The concordance statistic and the Cox model

Terry M. Therneau
David A. Watson

Technical Report #85
December 8, 2017

Department of Health Sciences Research
Mayo Clinic
Rochester, Minnesota

Abstract

The concordance statistic is used to measure the amount of agreement between two variables, often a risk score and time until an event in survival analysis. Surprisingly, the concordance statistic is a score statistic from a Cox model with a time varying coefficient. This relationship connects the literature on the concordance statistic and Cox model, specifically non-parametric techniques for survival analysis with time weighted Cox models. We also discuss the sensitivity of the concordance statistic with respect to the censoring distribution and introduce robust variance estimators for both the concordance statistic as well as comparisons between two correlated concordance statistics.

1 Introduction

The concordance between two variables X and Y is an old idea in statistics [4]. Two observed pairs (x_i, y_i) and (x_j, y_j) are said to be concordant if x and y have the same ordering, that is, either $\{x_i < x_j, y_i < y_j\}$ or $\{x_i > x_j, y_i > y_j\}$. A concordance statistic is the proportion of all possible pairs in a sample that are concordant:

$$\frac{\sum_{i \neq j} \mathbb{1}\{x_i < x_j, y_i < y_j\} + \mathbb{1}\{x_i > x_j, y_i > y_j\}}{(n-1)^2}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function and n is the sample size. Variations on the statistic handle ties in different ways, sometimes counting them as half a concordant pair and sometimes ignoring them.

In survival analyses, some event times are unobserved or right censored, complicating the notion of a concordant pair because not all pairs can be ordered. We consider a set of n subjects followed forward from time $t = 0$ until an event of interest occurs (e.g., death). The response for each subject is represented as a pair (t, δ) where survival time t is either the event time, if it was observed, or the last known time at which the event had not yet occurred and $\delta = 1$ if the event was observed and 0 otherwise (i.e., censored). Due to censoring, some pairs of survival times will be incomparable, for example a censored observation at 10 years and an event time at 12 years.

In such situations, [5] modified the concordance statistic to only consider comparable pairs. As a convenience, define the function $K(i, j)$ to be 1 if event time of subject i is known to be smaller than event time of subject j and 0 otherwise, that is

$$\begin{aligned} K(i,j) &= \mathbb{1}\{t_i < t_j, \delta_i = 1\} + \mathbb{1}\{t_i = t_j, \delta_i = 1, \delta_j = 0\} \\ &= \mathbb{1}\{\delta_i = 1\} \left[\mathbb{1}\{t_i < t_j\} + \mathbb{1}\{t_i = t_j, \delta_j = 0\}\right] \end{aligned}$$

The second indicator function addresses ties and follows the convention used in other survival methods that if subject j is censored at the same time that subject i has an event, then subject j has a longer event time than subject i. Other definitions do not consider such ties ordered [15].

[5] defined the concordance as the fraction of all the ordered time pairs (i.e., all i and j such that $K(i,j) = 1$ or $K(j,i) = 1$) in which the risk score, x, correctly predicts the order. Let τ be an upper time limit for comparison, for instance in advanced cancer any survival beyond five years might represent a cure and thus the relative ordering of any times beyond that point are uninteresting. The concordance statistic C is

$$C = \frac{\sum_{i \neq j} \mathbb{1}\{t_i < \tau\} K(i,j) \left[\mathbb{1}\{x_i > x_j\} + \mathbb{1}\{x_i = x_j\}/2\right]}{\sum_{i \neq j} \mathbb{1}\{t_i < \tau\} K(i,j)}. \tag{1}$$

Pairs where the risk score is tied count as half of an agreement, and tied event times are not counted as comparable and appear in neither the numerator nor denominator (these conventions are analogous to the AUC in logistic regression). For a Cox model, higher risk scores predict shorter event times, so C inverts the standard definition of concordance. Values of C range from 0 to 1 indicating a perfectly discordant to concordant risk score, and a value of 1/2 indicates the risk score is independent of the event times. The concordance statistic has become increasingly popular as a summary statistic for survival analyses, with recent work by [6], [14], [7], [24], and [15].

We show in the next section that the concordance statistic can be written as a sum over the distinct event times, where only the individuals still at risk contribute to the statistic. Rewriting the statistic in this way leads to many connections to the literature on survival analysis. Foremost, the concordance statistic is a score statistic from a Cox model with time varying covariates (Section 3.1). This realization relates the concordance statistic to previous work on rank transformations and time-weighted Cox models (Section 3.2). Moreover, in the two sample case (i.e., $x = 0, 1$), it leads to a connection between concordance and the Gehan-Wilcoxon test (Section 4.1), and this insight suggests a new class of concordance statistics with time-dependent weights (Section 4.2). Our reformulation suggests variance estimators for the concordance statistic (Section 4.3) and the difference between two correlated concordance statistics (Section 4.4).

2 Efficient computation

The obvious way to compute the concordance is to consider and rank all pairs of observations, which is a $O(n^2)$ computation and can be very slow for large data sets. Without loss of generality assume that the data has been sorted in time order such that $t_i \leq t_j$ for all $i < j$, with events preceding censorings in the case of tied times. Then $K(i,j) = 0$ for all $i \geq j$ and we can then rewrite the concordance as

$$2C - 1 = \frac{\sum_{i<j} \mathbb{1}\{t_i < \tau\} K(i,j) [\mathbb{1}\{x_i > x_j\} - \mathbb{1}\{x_j > x_i\}]}{\sum_{i<j} \mathbb{1}\{t_i < \tau\} K(i,j)} \equiv \frac{U}{D}.$$

Focusing on the numerator, we have

$$
\begin{aligned}
U &= \sum_{i:t_i<\tau} \delta_i \sum_{j>i} K(i,j)\,\mathrm{sign}(x_i - x_j) && (2)\\
&= \sum_{i:t_i<\tau} \delta_i \sum_{j\in\mathcal{R}(t_i)} \mathrm{sign}(x_i - x_j) && (3)\\
&= \sum_{i:t_i<\tau} \delta_i 2[r_i(t_i) - \bar{r}(t_i)] && (4)
\end{aligned}
$$

where

$$
\mathrm{sign}(z) = \begin{cases} -1, & z < 0 \\ 0, & z = 0 \\ 1, & z > 0. \end{cases}
$$

Equation (2) rewrites the sum as one term per event time, a common form for survival statistics. Equation (3) carries this further and writes this term as a sum over the risk set where $\mathcal{R}(t) = \{j : t_j \geq t\}$ which is the set of all observations at risk at time t. In Equation (4), $r_i(t)$ is the rank of x_i among all risk scores for those still at risk at time t and $\bar{r}(t)$ is the average of these ranks. The proof for these steps with further extensions to case weights is given in Appendix A of the supplementary materials.

Computation of Equation (4) requires ranking each observation within the risk set at each event time. This task is analogous to the problem of maintaining an ordered list as elements are added and removed from it and has a rich history in the computing literature on binary search trees. Using these methods, [12] computed C with computational time of $O(n \log_2 n)$. Whereas [12] applied this approach directly to Equation (1), the advantage of our reformulation is that it facilitates comparison to other statistical methods.

3 Cox model

3.1 Concordance as score statistic

The most surprising connection comes from realizing that Equation (4) is precisely the score statistic for a Cox model [1] with time-dependent covariate $r(t)$, namely the rank of the risk score, x, within the current risk set at time t. The (unstandardized) score statistic for the Cox model with a time-varying covariate $z(t)$ is

$$
\sum_i \delta_i \left[z_i(t_i) - \frac{\sum_{j\in\mathcal{R}(t_i)} z_j(t_i)\exp[z_j(t_i)\beta]}{\sum_{j\in\mathcal{R}(t_i)} \exp[z_j(t_i)\beta]} \right] \Bigg|_{\beta=0} = \sum_i \delta_i [z_i(t_i) - \bar{z}(t_i)],
$$

and letting $z(t) = 2r(t)$ exactly recovers Equation (4). When there are tied death times, the appendix verifies that Equation (4) matches the Breslow approximation for ties.

Equivalently, we can let $r_i^*(t) = r_i(t)/n(t)$ be the scaled ranks where $n(t)$ is the number of subjects still at risk at time t. Equation (4) can be rewritten as a time-weighted sum

$$U = \sum_{i:t_i<\tau} \delta_i 2n(t_i) \left[r_i^*(t) - \bar{r}^*(t) \right]. \tag{5}$$

Thus, we can also interpret U as a the score statistic of a time-weighted Cox model using the scaled ranks of the risk score, $r^*(t)$, as a time-dependent covariate and event time weights of $n(t)$. Both of these interpretations lead to interesting connections to previous work and suggest some interesting extensions.

3.2 Rank transformations and time-weighted Cox models

We relate the concordance statistic to previous work on scaled rank transformations and time-weighted Cox models, two ideas that have been considered separately in the literature on survival analysis.

[13] proposed a modification to the single covariate Cox model as a way to protect against outliers in x by reranking the covariate at each individual event time. The suggested statistic uses the sum of the logit transformed scaled ranks (which have expectation 0 under the null hypothesis): $\sum_i \delta_i \text{logit}(r_i^*(t_i))$. In practice, the final logit transform has little impact on the statistical properties of the O'Brien test, similar to the what is seen between a rank-sum and normal scores test [8]. Thus the primary difference between the concordance and O'Brien's test statistic is that the former weights each event by $n(t)$ and the latter weights events evenly.

[11], [21], [18], and [20] discuss time-weighted Cox regression, in particular for assessing and handling non-proportional hazards. If we apply this method with the time-dependent scaled ranks $r^*(t)$ as the covariate and using event time weight of $n(t)$, their score statistic is exactly U of Equation (5). The scaled ranks (compared to unscaled) behave more like a typical covariate in a Cox model. For instance, the resulting score statistic has $\bar{r}^*(t) = 1/2$, which is similar to the usual Cox models where $\bar{z}(t)$ normally has only small variation over time. Although previous work on time-weighted Cox models focused primarily on time fixed covariates, often baseline scaled ranks, $r^*(0)$, are very similar to time-dependent scaled ranks, $r^*(t)$, which we show with the following example.

Example 3.1. The lung cancer dataset included in the `survival` package in `R` consists of 228 patients enrolled in chemotherapy trials for advanced disease [23]. We fit a simple Cox model for the time until death with predictors age, sex, and patient's assessment of their own Karnofsky score. Figure 1 shows that the time-dependent scaled ranks of the risk score contribute approximately the same as the baseline (or time fixed) ranks to the respective score statistics.

In light of this previous work, we can view the concordance statistic as a combination of a time-weighted Cox model on the rank transformed risk score.

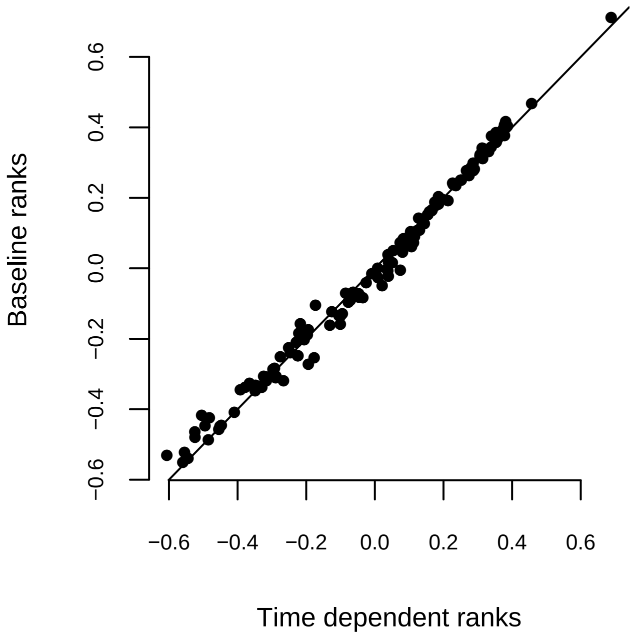


Figure 1: Contribution to score statistic for time-dependent scaled ranks, $r_i^*(t_i) - \bar{r}^*(t_i)$, versus baseline scaled ranks, $r_i^*(0) - \bar{r}^*(0)$. Solid line is 45 degree line.

4 Additional connections

4.1 Two sample rank tests

It is well known that for a binary covariate the score test of the Cox model is equivalent to a log rank test. Moreover, allowing for a time-dependent weights leads to a whole class of rank tests that compare the number of the observed events to expected events under the null hypothesis at each event time, that is a test statistic

$$V_W = \sum_k W_k \left[d_{k1} - \frac{n_{k1} d_k}{n_k} \right] \tag{6}$$

where the sum is over all distinct event times, d_{1k} and d_k are the number of events at the kth event time in group 1 and overall respectively, n_{1k} and n_k are the number of individuals at risk at the kth event time in group 1 and overall respectively, and W_k is some time-dependent weighting function. Under the null hypothesis that the two groups have identical survival functions, standardized V_W is asymptotically a standard normal random variable. Letting $W_k = 1$, as mentioned, produces the log rank test. A weight of $W_k = n_k$ produces the Gehan-Wilcoxon test and this test statistic is in fact the same as U for a binary risk score (for details see Appendix B of the supplementary materials); that is, the Gehan-Wilcoxon test statistic is the concordance statistic.

[16] and [17] showed theoretical and substantive examples respectively of how the censoring distribution can adversely affect the Gehan-Wilcoxon statistic. Heuristically, letting $S(t)$ and $G(t)$ be the underlying survival functions of the time until event and censoring respectively, if event times and censoring times are independent then $E[n(t)] \propto S(t)G(t)$. Thus $n(t)$ depends on the censoring distribution, which may differ across studies or between groups. In particular, heavy censoring puts much of the weight on early events. [16] suggested a weight of $W(t) = \hat{S}(t)$, an estimate of the survival function of the event times. Several variants have been explored; for an overview, see Section 7.3 of [10].

4.2 Time-weighted concordance

The same criticism of the Gehan-Wilcoxon test should also apply to the concordance statistic, which implicitly weights the scaled ranks by $n(t)$. The general form of the rank test suggests a new class of concordance statistics with a choice of time-dependent weights. In particular, let

$$U_W = \sum_{i:t_i<\tau} \delta_i 2W(t_i)[r_i^*(t_i) - \bar{r}^*(t_t)], \qquad D_W = \sum_{i:t_i<\tau} W(t_i) \frac{n(t_i) - d(t_i)}{n(t_i)},$$

where $d(t)$ is the number of events at time t. For D_W, a similar argument to Equations (2)–(4) shows subject i contributes $\delta_i[n(t_i) - d(t_i)]$ to the sum of D, and in order to generalize for time-dependent weights, each term is multiplied and divided by $W(t_i)$ and $n(t_i)$ respectively. The time-weighted concordance

$E[W] \propto$	W	$V_W/se(V_W)$	C_W
SG	n	0.3	0.517
	$\hat{S}\hat{G}$	0.3	0.516
S	$n/\hat{G}$	2.4	0.554
	$\hat{S}$	2.4	0.555
S/G	$n/\hat{G}^2$	4.4	0.538
	$\hat{S}/\hat{G}$	4.4	0.541

Table 1: Two sample test statistics and concordance for different weights for the data in Example 4.1

measure C_W with weight $W(t)$ is

$$C_W = \frac{U_W}{2D_W} + \frac{1}{2}.$$

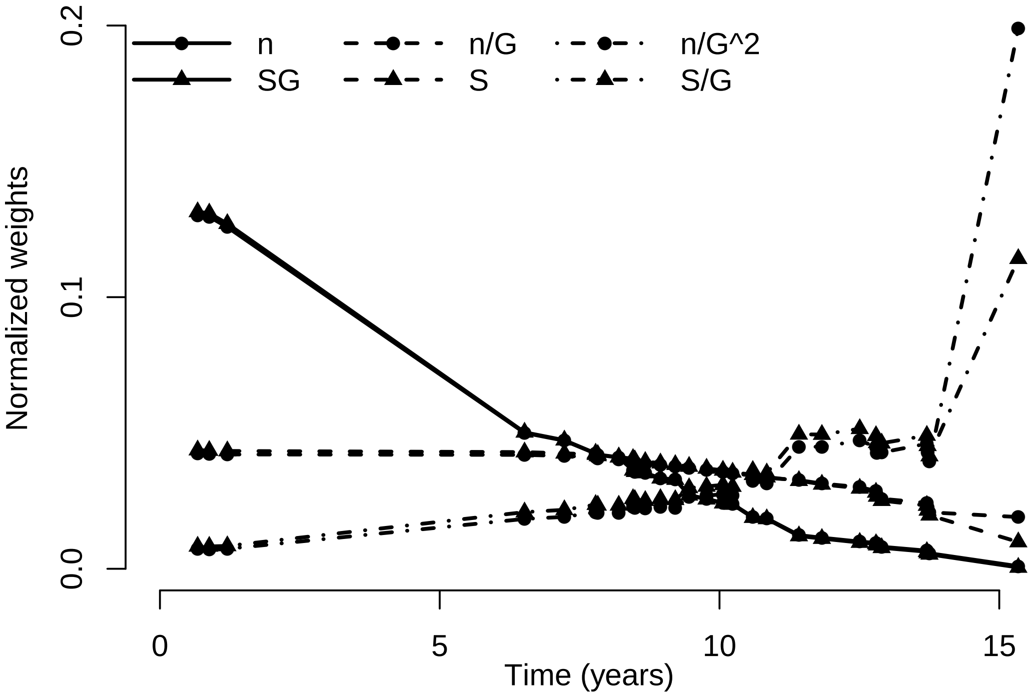
[24] first proposed a time-weighted concordance statistic in order to make the resulting measure robust to different censoring patterns; in particular, a weight of $W(t) = n(t)\hat{G}(t)^{-2}$ is recommended where $\hat{G}(t)$ is an estimate of the survival function of the censoring distribution. The weighted concordance statistic of [24] is written in a form similar to the definition in Equation (1) and does not explicitly include a factor of $n(t)$ in the weight, but we see this factor implicitly applies from Equation (5). Again as a heuristic, $E[n(t)\hat{G}(t)^{-2}] \propto S(t)G(t)^{-1}$. A weight of $W(t) = \hat{S}(t)\hat{G}(t)^{-1}$ is recommended by [20], but for the time-weighted Cox model.

In the presence of no censoring, $E[n(t)] \propto S(t)$, which suggests a weight proportional to (an estimate of) the survival function of the event times. Indeed, this is the recommendation of [16] for their rank test. We explore the weighted concordance with weights that are proportional to $S(t)G(t)$, $S(t)$, and $S(t)G(t)^{-1}$ with a substantive example and simulations.

Example 4.1. [17] reported data on time until developing cancer for 281 dogs that either did or did not receive a marrow transplant and total body irradiation. This data suffers from extreme censoring and provides a substantive example of a disparity between the Gehan and Peto generalizations of the Wilcoxon rank-sum test. The test and concordance statistics are provided in Table 1 and the weights for each event time are plotted in Figure 2. Weights proportional to SG, which the concordance statistic implicitly use, put much of the weight at early time points. In contrast, the weights proportional to S/G heavily weight the later times and become slightly erratic near the end due to extreme censoring. Interestingly, these weights provide the largest (and most significant) test statistic, but this significance does not translate into the largest concordance statistic.

Example 4.2. We consider a simulation to demonstrate potential disparities in the weighted concordance. The risk score, which is distributed as a standard

Figure 2: Weights at different event times for the data in Example 4.1. Weights are normalized to sum to 1.

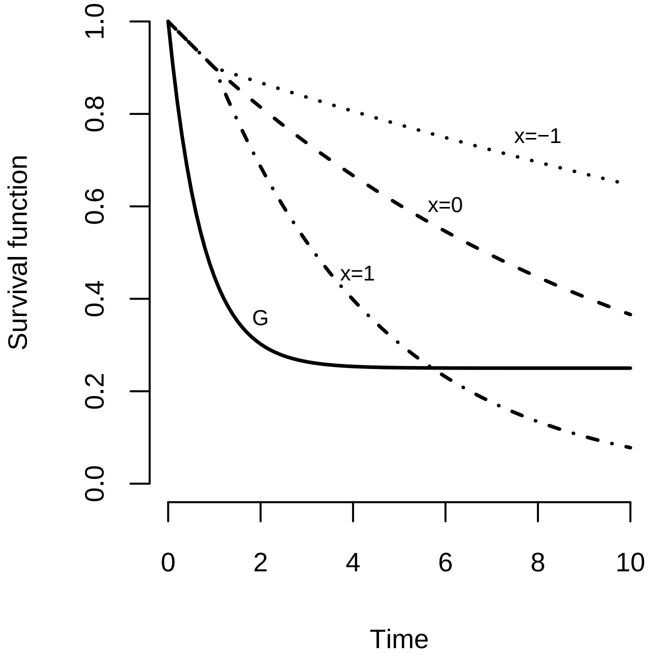


normal random variable, only differentiates longer event times, specifically after one year. A quarter of the sample of size $n = 200$ is followed until an event is observed, but the remaining three quarters are heavily censored early. The survival functions of the event and censoring times are plotted in dashed and solid lines respectively in Figure 3. The uncensored concordance is 0.77 and the simulated averages of the weighted concordances for weights proportional to SG, S, and S/G are 0.67, 0.72, and 0.77. Clearly, the standard concordance statistic, with a weight proportional to SG, is biased for this model where as the weights proportional to S/G work well.

and the weights proportional to S/G perform the best with slight positive bias. Under the conditionally independent censoring mechanism, all weights underestimate the uncensored concordance, though again weights proportional to S/G were the least biased.

The theory for the weight $n/\hat{G}^2$ recommended in [24] assumes that the censoring distribution is independent of the event distribution and risk score. If this

Figure 3: Survival curves for event and censoring times for simulation in Example 4.2. Survival curves for event times are dashed and dotted lines for low ($x = -1$), medium ($x = 0$), and high risk ($x = 1$) risk scores. The solid line is the censoring distribution.



assumption is violated, [3] suggests weighting with an estimate of the censoring distribution that conditions on the risk score.

The first factor of $1/G$ accounts for the fact that there would have been more terms in the sum without censoring, and the second for the fact that $n(t)$ would be larger within each term. Replacing $n(t)$ with $G(t)S(t)$ leads to score statistic of Schemper, showing a close relationship between the two approaches.

4.3 Variance estimation

A natural estimate of the variance of U, and thus for concordance C as well, is to use the variance of the Cox score statistic, that is the observed information evaluated at $\beta = 0$. For the Cox model, this estimate is simply the sum of the variances of the ranks, $r(t)$, over the risk set at each event time. The resulting variance estimator is

$$\hat{V}_{\text{sc}} = \frac{1}{D^2} \sum_i \delta_i \left[\sum_{j \in \mathcal{R}(t_i)} \frac{r_j(t_i)^2}{n(t_i)} - \bar{r}(t_i)^2 \right].$$

Prop. to	Weight	Independent censoring	Cond. indep. censoring
SG	n	0.55	0.57
	$\hat{S}\hat{G}$	0.55	0.57
S	$n/\hat{G}$	0.58	0.58
	$\hat{S}$	0.58	0.58
S/G	$n/\hat{G}^2$	0.63	0.59
	$\hat{S}/\hat{G}$	0.63	0.59

Table 2: Simulated expectations of weighted concordance statistics under an independent and conditionally independent censoring distribution. Expected value of the concordance statistic without censoring is 0.62. All Monte Carlo standard errors are less than 10^{-3}.

When the risk score is independent of the event time (i.e., $\beta = 0$ or a true concordance of 1/2), this variance estimate is correct, so this estimate is appropriate for testing whether the concordance is 1/2. However, if the risk score is associated with the event times, this estimator tends to be an overestimate of the variance.

When the risk score is associated with the event we would prefer an estimator that is robust to alternatives to the null hypothesis of no association. The infinitesimal jackknife of [9] offers a robust variance estimator for the Cox model e.g., see Chapter 7.2 of [22], and thus could work well for the concordance statistic. The results of Section 2 facilitate efficient computation of the infinitesimal jackknife estimator (formulas are given in Appendix A of the supplementary materials).

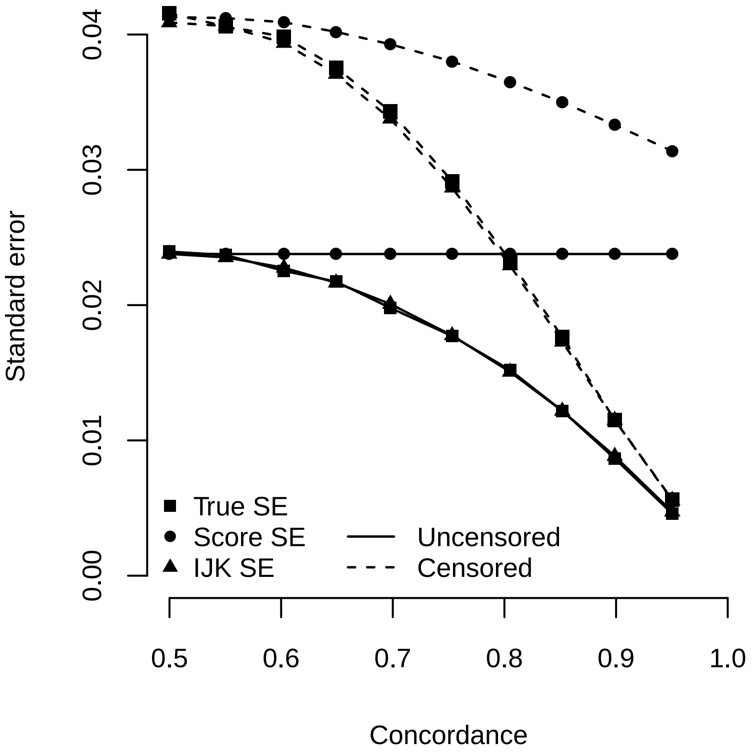
Example 4.3. We show the performance of these variance estimators via simulation of a sample of size $n = 200$. The risk score is drawn from a standard normal distribution. Event times are drawn from an exponential distribution with mean $e^{-\gamma x}$ for $\gamma \geq 0$. Values of γ are chosen to correspond to an (uncensored) concordance of approximately 0.5, 0.55, ..., 0.95. We consider both no censoring and exponentially distributed censoring times with mean 2. Estimated and true standard errors are plotted against the concordance in Figure 4. The variance of the score statistic is correct for a concordance of 0.5, but overestimates the variance for larger values. The infinitesimal jackknife has little bias.

Other estimates of the variance have been proposed in [14] and [12]. Future work should assess these variance estimators, including frequency properties and computational complexity.

4.4 Comparing correlated concordances

A reasonable comparison of two risk scores is the difference of the resulting concordance statistics. The results from the previous section on the infinitesimal

Figure 4: Simulated results of Example 4.3: true standard errors (squares) and estimated standard errors via the variance of the score statistic (circles) and infinitesimal jackknife (triangles) for uncensored (solid line) and censored data (dashed line). All Monte Carlo standard errors are less than 0.005.



jackknife are easily extended to estimating the variance of the difference of two (potentially correlated) concordance statistics (see Appendix A of supplementary materials or details). The resulting variance estimator is comparable to the results of [2] for the AUC based on U-statistics theory because the resulting variance estimators from the infinitesimal jackknife and asymptotic U-statistic theory are identical [19].

Example 4.4. We revisit the dataset from Example 3.1 to compare the concordance statistic from risk scores from two Cox models, one with predictors age, sex, and patient's assessment of their own Karnofsky score and the other with just predictors age and sex. The concordance statistics for the larger and smaller models are 0.64 and 0.60 respectively. The standard error for the estimated difference of 0.04 is 0.02, so the difference is statistically significant at the 0.05 level. Thus, the improvement in the concordance when including the Karnofsky score in the Cox model is statistically significant.

5 Conclusion

It is a surprising fact that the concordance statistic is the score statistic from a Cox model. This realization helps in understanding previous work showing the dependence of the concordance statistic on the censoring distribution [24] and provides easy to compute variance estimators for the concordance statistic. There are potentially many other connections to explore between the concordance statistic and the Cox model.

A Case weights

Define analogous definitions of U and W using case weights, w_i (not to be confused with time-dependent weights, $W(t)$), as

$$U = \sum_i w_i \delta_i \sum_{j:K(i,j)=1} w_j \text{sign}(x_i - x_j)$$
$$D = \sum_i w_i \delta_i \sum_{j:K(i,j)=1} w_j$$

If the weights w are positive integers, this is exactly the result obtained by making a new data set that contains w_i copies of individual i.

The step from (2) to (3) is obvious if there are no ties, thus we focus only on the situation in which ties are present. Consider a and b such that $a < b$ and $t_a = t_b$. If both a and b have events at the observed time, $\delta_a = \delta_b = 1$, then $K(a,b) = 0$ and this pair does not contribute to the sum of (2). In (3), when $i = a$, a term of $w_a w_b \text{sign}(x_a - x_b)$ contributes to the sum, but it is cancelled by the term $w_a w_b \text{sign}(x_b - x_a)$, which contributes to the sum when $i = b$. If individual b is censored, then $\delta_b = 0$ and the term $w_a w_b \text{sign}(x_a - x_b)$ contributes to the sum in (2). This term also contributes to the sum in (3) when $i = a$, and nothing contributes to the sum when $i = b$ because the δ_i term zeros the summation over the risk set. Including all tied survival times in the other's risk set, that is letting $\mathcal{R}(t_i) = \{j : t_j \geq t_i\}$ corresponds to the Breslow approximation for ties in a Cox model computation.

To see the step from (3) to (4), we define the rank of x_i within a set of m weighted observations as the sum of the weights for all observations with x_j less than x_i, plus half the sum of weights for ties, that is

$$\begin{aligned} r_i &= \sum_{j=1}^{m} \left[w_j I(x_i > x_j) + \frac{w_j}{2} I(x_i = x_j) \right] \\ &= \sum_{j=1}^{m} w_j \left[\text{sign}(x_i - x_j) + 1 \right] / 2 \\ &= \bar{r} + \frac{1}{2} \sum_{j=1}^{m} w_j \text{sign}(x_i - x_j) \end{aligned}$$

where $\bar{r} = \sum w_i r_i / \sum w_i$, the weighted mean of the ranks. To see the last line, the numerator of $\bar{r}$ is

$$\sum_{i<j} w_i w_j + \sum_i w_i^2/2 = \sum_{i \neq j} w_i w_j/2 + \sum_i w_i^2/2 = \frac{1}{2} \sum_i \sum_j w_i w_j = \frac{1}{2} \left[\sum_i w_i \right]^2$$

and thus $\bar{r} = \sum w_i/2$. This justifies the transition from (3) to (4), or more generally, $U = \sum_i w_i \delta_i 2[r_i(t_i) - \bar{r}(t_i)]$. When all the weights are 1 the ranks defined above are 0.5, 1.5, 2.5, ..., rather than the more usual 1, 2, 3, ..., but this convention makes no difference for Equation (4).

Introducing case weights, in addition to their obvious use, facilitates the derivation of the variance estimator via the infinitesimal jackknife [9]. The variance estimator is

$$\hat{V}_{\text{ijk}} = \sum_i w_i \left(\frac{\partial C}{\partial w_i}\right)^2 = \sum_i w_i \left(\frac{DU_i' - UD_i'}{2D^2}\right)^2$$

where

$$\begin{aligned} U_i' \equiv \frac{\partial U}{\partial w_i} &= \sum_{j:K(i,j)=1} w_j \delta_i \text{sign}(x_i - x_j) - \sum_{j:K(j,i)=1} w_j \delta_j \text{sign}(x_i - x_j) \\ D_i' \equiv \frac{\partial D}{\partial w_i} &= \sum_{j:K(i,j)=1} w_j \delta_i - \sum_{j:K(j,i)=1} w_j \delta_j . \end{aligned}$$

The above abuse of notation means evaluate the partial derivative of the concordance at the specified weights. The second sum on the right hand side of U_i' compares x_i to all x_j for which i is at risk at time t_j (and not a tied event). Using similar arguments as above, U_i' can be rewritten in terms of ranks over the risk set and event set (i.e., the observed events up to a specified time). Computing $\hat{V}_{\text{ijk}}$ can be implemented in $O(n \log_2 n)$ computational time using binary search trees.

Now suppose we want to compare two concordance statistics from two risk scores fit to the same dataset, that is $C_v = (U_v/D + 1)/2$ where $v = 1, 2$ denotes the risk score and noting D is the same for both risk scores. The infinitesimal jackknife variance estimator of the difference $C_1 - C_2$ is

$$\hat{V}_{\text{diff}} = \sum_i w_i \left[\frac{\partial (C_1 - C_2)}{\partial w_i}\right]^2 = \sum_i w_i \left[\frac{D(U_{1i}' - U_{2i}') - (U_1 - U_2)D_i'}{2D^2}\right]^2 .$$

B Binary risk score

We show for a binary risk score, that is $x = 1$ for high risk subjects and $x = 0$ otherwise, U of Equation (6) equals Z of Equation (5) when $W_k = n_k$, that is the Gehan-Wilcoxon test statistic. For this section, we assume all survival times are unique. For individual i corresponding to the kth event, it suffices to show

$$\begin{aligned} 2n(t_i)[r_i^*(t_i) - \bar{r}^*(t_i)] &= n_k[d_{1k} - n_{k1}/n_k] \\ &= n(t_i)[x_i - n_1(t_i)/n(t_i)] \end{aligned} \tag{7}$$

where the second line follows from exchanging the event time index of k for the individual index of i and $n_x(t)$ is the number of individuals at risk in group $x = 0, 1$ at time t. If $x_i = 1$, then the left hand side of Equation (7) is $2[n_0(t_i) + n_1(t_i)/2 - n(t_i)/2] = n_0(t_i)$. Similarly, if $x_i = 0$, then the left hand side of Equation (7) is $2[n_0(t)/2 - n(t)/2] = -n_1(t_i)$. Thus, the Gehan-Wilcoxon statistic equals U.

For variance estimators, similar argument show each term in the sum of $\hat{V}_{\mathrm{sc}}$ is simply $\delta_i n_0(t_i) n_1(t_i)/4$ and, letting $e_1(t)$ be the number of events in group 1 occurring before time t, $U_i' = e_1(t_i) - x_i e(t_i) + \delta_i[x_i n(t_i) - n_1(t_i)]$, which can be plugged into the formula for $\hat{V}_{\mathrm{ijk}}$.

B.1 The details for D_W

Following similar logic as Equations (2)–(4), we have

$$D = \sum_{i:t_i<\tau} \delta_i \sum_{j>i} K(i,j) = \sum_{i:t_i<\tau} \delta_i [n(t_i) - d(t_i)],$$

where $d(t)$ is the number if individuals who are have an event at time t. To generalize the concordance statistic for time weights, we multiply each term by $W(t_i)$ and divide by $n(t_i)$. Performing the same operation leads to D_W.

C Derivatives

$$\begin{aligned} C - D &= \sum_i w_i \delta_i \sum_{t_j>t_i} w_j \mathrm{sign}(\mathrm{r_i} - \mathrm{r_j}) \\ C + D + T &= \sum_i w_i \delta_i \sum_{t_j>t_i} w_j \end{aligned}$$

The first derivatives of these quantities with respect to an arbitrary subject k are

$$\frac{\partial C - D}{\partial w_k} = \sum_{t_j>t_k} w_j \delta_k \mathrm{sign}(\mathrm{r_k} - \mathrm{r_j}) - \sum_{\mathrm{t_j<t_k}} \mathrm{w_j} \delta_{\mathrm{j}} \mathrm{sign}(\mathrm{r_k} - \mathrm{r_j}) \tag{8}$$

$$\frac{\partial C + D + T}{\partial w_k} = \sum_{t_j>t_k} w_j \delta_k + \sum_{t_j<t_k} w_j \delta_j \tag{9}$$

References

[1] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34(2):187–220, 1972.

[2] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[3] Thomas A Gerds, Michael W Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2013.

[4] Leo A Goodman and William H Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764, 1954.

[5] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

[6] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[7] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[8] JL Hodges Jr and Erich L Lehmann. Comparison of the normal scores and Wilcoxon tests. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 307–317, 1961.

[9] Louis A Jaeckel. The infinitesimal jackknife. Technical Memo MM 72-1215-11, Bell Laboratories, 1972.

[10] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer-Verlag, New York, 2003.

[11] DY Lin. Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association*, 86(415):725–728, 1991.

[12] R Newson. Confidence intervals for rank statistics: Somers' D and extensions. *Stata Journal*, 6(3):309–334, 2006.

[13] Peter C O'Brien. A nonparmetric test for association with censored data. *Biometrics*, 34(2):243–250, 1978.

[14] Michael J Pencina and Ralph B D'Agostino. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004.

[15] Michael J Pencina, Ralph B D'Agostino, and Linye Song. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Statistics in Medicine*, 31(15):1543–1553, 2012.

[16] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society Series A*, 135(2):185–206, 1972.

[17] Ross L Prentice and P Marek. A qualitative discrepancy between censored data rank tests. *Biometrics*, 35(4):861–867, 1979.

[18] Peter Sasieni. Maximum weighted partial likelihood estimators for the Cox model. *Journal of the American Statistical Association*, 88(421):144–152, 1993.

[19] Edna Schechtman. On estimating the asymptotic variance of a function of U statistics. *The American Statistician*, 45(2):103–106, 1991.

[20] M. Schemper, S. Wakounig, and G. Heinze. The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, 28(19):2473–2489, 2009.

[21] Michael Schemper. Cox analysis of survival data with non-proportional hazard functions. *The Statistician*, 41(4):455–465, 1992.

[22] T. M. Therneau and P. M. Grambsch. *Modeling survival data: extending the Cox model*. Springer-Verlag, New York, 2000.

[23] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. Version 2.38.

[24] H. Uno, T. Cai, M. J. Pencina, R. B D'Agnostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.