Biostatistics (2019) **20**, 2, *pp.* 347–357
doi:10.1093/biostatistics/kxy006
Advance Access publication on February 16, 2018

The c-index is not proper for the evaluation of *t*-year predicted risks

PAUL BLANCHE

Section of Biostatistics, Department of Public Health, University of Copenhagen, Oester Farimagsgade 5, 1014 Copenhagen, Denmark

MICHAEL W. KATTAN

Department of Quantitative Health Sciences, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, OH 44195, USA

THOMAS A. GERDS*

Section of Biostatistics, Department of Public Health, University of Copenhagen, Oester Farimagsgade 5, 1014 Copenhagen, Denmark

tag@biostat.ku.dk

Summary

We show that the widely used concordance index for time to event outcome is not proper when interest is in predicting a *t*-year risk of an event, for example 10-year mortality. In the situation with a fixed prediction horizon, the concordance index can be higher for a misspecified model than for a correctly specified model. Impropriety happens because the concordance index assesses the order of the event times and not the order of the event status at the prediction horizon. The time-dependent area under the receiver operating characteristic curve does not have this problem and is proper in this context.

Keywords: Cox regression; Concordance index; Discrimination ability; Model comparison; Survival prediction.

1. Introduction

The concordance index, also known as c-index or c-statistic (Harrell *and others*, 1982, 1996; Pencina and D'Agostino, 2004) has been an enormously popular metric for evaluating the performance of a predictive or prognostic statistical model. The c-index has the desirable features of being relatively simple to calculate and explain to a clinical audience. Moreover, it is on a user-friendly scale that ranges from 0.5 (no discriminating ability) to 1.0 (perfect ability to discriminate between cases with different outcomes). For these reasons, a very high proportion of prognostic models in the literature will report their performance as measured by the c-index, and most readers and reviewers expect to see this metric reported. The typical case is a prognostic model that reports *t*-year predicted probabilities, say 5-year overall survival probabilities, that should be similar to 5-year observed outcomes. Examples where *t*-year risk predictions are evaluated with the c-index are Stephenson *and others* (2005) and Pencina *and others* (2009). The *t*-year area under the receiver operating characteristic (ROC) curve (AUC_t) is an alternative measure of the

*To whom correspondence should be addressed.

discriminative ability (Heagerty *and others*, 2000; Chambless and Diao, 2006). Examples where t-year predictions are evaluated with AUC_t are Cornec-Le Gall *and others* (2016) and Mortensen *and others* (2017).

To explain the main difference between the two major discrimination measures for survival analysis, we consider two random subjects (i, j) from a population for who we have t-year risk estimates, $\text{Risk}_t(i)$ and $\text{Risk}_t(j)$, respectively. Informally, the population parameters c-index $\mathcal{C}$ and AUC_t are given by (see Section 3 for a formal description, details and more references)

$$\mathcal{C} = \text{Prob}(\text{Risk}_t(i) > \text{Risk}_t(j) \,|\, i \text{ has event before } j)$$
$$\text{AUC}_t = \text{Prob}(\text{Risk}_t(i) > \text{Risk}_t(j) \,|\, i \text{ has event before } t \text{ and } j \text{ has event after } t).$$

The fact that two alternative discrimination measures are available is a dilemma for the practitioner who needs to choose between them (see Section 2). The current guidelines (e.g., Hlatky *and others*, 2009; Moons *and others*, 2015; Pencina and D'Agostino, 2015) clearly state that a discrimination measure should be reported, but they do not provide recommendations on how to choose the most appropriate discrimination measure. Another sign of the widespread confusion is that the term c-index sometimes refers to what we define as $\mathcal{C}$ and sometimes to what we define as AUC_t above.

However, the c-index has never been fully assessed as to whether it is a proper scoring rule. To satisfy this condition, the c-index must be able to always achieve a maximum value for the rival prediction model that truly outperforms any other model (i.e., the c-index needs to be highest for the best prediction model). The purpose of this study is to examine, mathematically, whether the c-index is a proper scoring rule in the t-year risk prediction setting.

Our derivations and conclusions are independent of the way c-index $\mathcal{C}$ and AUC_t are estimated, and we state them in terms of population level parameters. In particular, our derivations are not affected by whether data are censored.

We argue that the c-index $\mathcal{C}$ is not generally suitable to evaluate t-year predictions. The reason is that the outcome which corresponds to the predicted t-year risks is the binary event status at time t. However, the c-index compares the ranks of the predictions with the ranks of the actual event times, and not directly with the binary event status.

To make our argument rigorous, we state propriety (Gneiting and Raftery, 2007; Pepe *and others*, 2015) as a minimal requirement on a measure of discrimination. We then derive a sufficient condition on the underlying survival distribution under which the c-index satisfies the minimal requirement. But, we also provide an example in which c-index fails this requirement. It is important to recognize that the condition is about the unknown true distribution of the data and not satisfied in general. On the other hand, AUC_t satisfies this requirement without further conditions.

Based on these theoretical considerations, an outcome of our study is a new guideline for applied research. We recommend that for evaluation of t-year risks, analysts should use the time-dependent area under the ROC curve instead of the c-index.

2. MOTIVATION

The fact that there are several possibilities to calculate a concordance index for a survival model puts high responsibility on the applied researcher and shows a demand for guidelines. A particular problem is that the actual values of the different discrimination measures can be quite different in a given situation. We first illustrate this dilemma by using five different survival data sets with fixed prediction horizon, and then by varying the prediction horizon in a single data set.

2.1. *Illustration of the difference between* $\mathcal{C}$ *and* AUC_t

In each data set, we first fit a Cox regression model (see Appendix for details) and then predict the 5-year risk of the event of interest. Thus, the prediction horizon is the same ($t = 5$) in all examples but the

maximal end of follow-up time (τ) differs between the different data sets. We then calculate the following estimates of the discrimination measures (see Appendix for details):

1. IPCW estimate of AUC(t): $\widehat{\text{AUC}}_{\text{IPCW}}(t)$ (Uno *and others*, 2007)
2. IPCW estimate of c-index (Uno *and others*, 2011)
   - outcome not artificially censored: $\widehat{C}_{\text{IPCW}}(\tau)$
   - outcome artificially censored at 5 years: $\widehat{C}_{\text{IPCW}}(t)$
3. Harrell's estimate of the c-index (Harrell *and others*, 1996)
   - outcome not artificially censored: $\widehat{C}(\tau)$
   - outcome artificially censored at 5 years: $\widehat{C}(t)$.

The differences among the measures in Table 1 are substantial, underscoring the importance of choosing a measure before the analysis. In Table 1, the values of $\widehat{AUC}_{\text{IPCW}}(t)$ are all larger than those of $\widehat{C}_{\text{IPCW}}(t)$, but this can be the other way around, see e.g., Figure 1.

Table 1. *Estimates of discrimination measures for Cox regression models that predict t-year survival ($t = 5$ years) in five different data sets*

Data set	$\widehat{AUC}_{\text{IPCW}}(t)$	$\widehat{C}_{\text{IPCW}}(t)$	$\widehat{C}_{\text{IPCW}}(\tau)$	$\widehat{C}(t)$	$\widehat{C}(\tau)$	τ (years)
`pbc`	89.2	83.4	77.1	83.7	81.6	13.1
`follic`	64.0	61.3	63.7	61.3	62.4	31.1
`GBSG2`	75.4	68.2	67.8	69.3	69.2	7.3
`cost`	75.5	70.2	68.6	70.2	68.6	11.7
`recc`	81.3	81.2	67.2	83.1	82.2	17.0

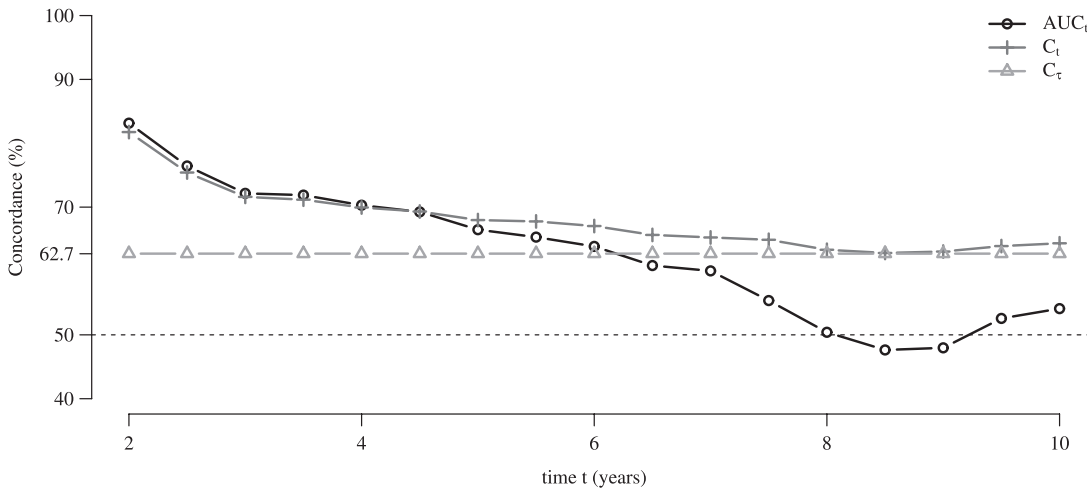

Fig. 1. The ability of the standardized blood clotting time to discriminate patients who will either die or be transplanted within the next t-years from those who will be event-free at time t. Discrimination ability is quantified by AUC$_t$, $\mathcal{C}_\tau$ (where $\tau = 13.1$ years), and by $\mathcal{C}_t$ obtained by artificially censoring the outcome at the prediction horizon t.

2.2. *Varying prediction time horizon*

We now vary the prediction time horizon in the `pbc` data set (see Appendix for details) to show how contradicting conclusions can occur depending on whether $\mathcal{C}_\tau$ or AUC_t is used. For this example, we consider the discriminative ability of a single predictor variable: the standardized blood clotting time and vary the prediction horizon between 2 and 10 years (Figure 1). We estimate AUC_t and $\mathcal{C}_\tau$ and also $\mathcal{C}_t$ in a version of the data where we artificially censor the outcome at the current prediction horizon t.

For illustration purposes we now interpret the estimates in Figure 1 without consideration of their uncertainty. In Section 4, we provide an example based on simulated data where we can compute the true values. For instance, the estimate of AUC_t at $t = 8$ years suggests that the standardized blood clotting time does not help to discriminate the 8-year outcome, as the value is very close to the benchmark 50% (coin flip). However, according to overall $\mathcal{C}_\tau$ and $\mathcal{C}_t$ the discrimination ability of the standardized blood clotting time discriminates better than the benchmark for all prediction horizons.

3. DISCRIMINATION MEASURES

Suppose we observe the continuous time to an event T and a d-dimensional vector of baseline covariates Z. For a fixed prediction horizon $t < \tau$ where τ is the maximum follow-up time, which is smaller than or equal to the study duration, the aim is to discriminate the cumulative risk of the event before time t based on information in Z. Denote by P the joint law of (T, Z) and by (T_1, Z_1) and (T_2, Z_2) two independent and identically distributed replicates of (T, Z). For any function $R : \mathbb{R}^d \to \mathbb{R}$, the population parameter, which is estimated by the c-index, is given in its truncated version (Heagerty and Zheng, 2005; Uno *and others*, 2011; Gerds *and others*, 2013) by

$$\mathcal{C}_\tau(R) = \text{P}(R(Z_1) > R(Z_2) | T_1 < T_2, T_1 \le \tau) + 0.5\,\text{P}(R(Z_1) = R(Z_2) | T_1 < T_2, T_1 \le \tau).$$

Because of the finite maximum follow-up time τ, only this truncated version of the c-index is identifiable without assuming not testable assumptions (Uno *and others*, 2011). Note that in much of the applied research, c-index is reported without explicit reference to τ. A typical way to use the c-index is when $R(Z)$ is a linear predictor, that is $R(Z) = Z'\beta$, or a monotone function of such a linear predictor (Harrell *and others*, 1996; Heagerty and Zheng, 2005; Uno *and others*, 2011). Note that we allow for ties in the predicted risks and add a value of 0.5 to the concordance index in this case. This is in accordance with the more general definition of a probabilistic index (Thas *and others*, 2012).

On the other hand, the so called cumulative-dynamic time-dependent area under the ROC curve (AUC) (Heagerty *and others*, 2000; Heagerty and Zheng, 2005; Chambless and Diao, 2006; Uno *and others*, 2007; Blanche *and others*, 2013) is defined for all $t \le \tau$ by

$$\text{AUC}_t(R) = \text{P}(R(Z_1) > R(Z_2) | T_1 \le t, T_2 > t) + 0.5\,\text{P}(R(Z_1) = R(Z_2) | T_1 \le t, T_2 > t).$$

Any estimate of $\text{P}(T \le t|Z)$ can be interpreted as the t-year predicted risk and is a Z-measurable random variable that can be assessed by both discrimination measures. The difference between these two discrimination measures is that the Harrell-type concordance index rank correlates predictions with the actual event times whereas the time-dependent AUC rank correlates predictions with the binary event status at time t.

3.1. *Minimal requirement for discrimination measures*

Consider a setting where the aim is to evaluate the discriminative ability of a t-year prediction rule based on a discrimination measure. We require that the true t-year risk:

$$F_t(Z) = \mathrm{P}(T \leq t|Z)$$

achieves the maximal discrimination value. Note that the set of all functions $R : \mathbb{R}^d \to \mathbb{R}$ evaluated at Z in the following definition includes any estimate of $F_t(Z)$.

DEFINITION 1 The discrimination measure $\mathcal{D}$ is proper for evaluating t-year risk predictions if for all $R : \mathbb{R}^d \to \mathbb{R}, \quad \mathcal{D}(R) \leq \mathcal{D}(F_t)$.

We show that $\mathcal{D} = \mathrm{AUC}_t$ is proper according to Definition 1 but that without untestable assumptions $\mathcal{D} = \mathcal{C}_\tau$ is not proper. The minimal requirement means that no prediction model using the information Z should have a higher discriminative value than the true data generating risk function.

3.2. *Harrell's c-index*

The result of the following theorem states a sufficient condition under which the minimal requirement (Definition 1) is satisfied for the c-index. We emphasize that the condition depends on both the function R, which usually relates to a working prediction model, and the true conditional survival function $S : t \mapsto 1 - \mathrm{P}(T \leq t|Z)$. Hence, the condition cannot be checked solely by choosing the working prediction model carefully, as illustrated in Section 4.

THEOREM 3.1 Let $\lambda(s|Z)$ denote the true conditional hazard at time s that is

$$\lambda(s|Z) = \frac{\mathrm{d}}{\mathrm{d}s}(-\log\{1 - \mathrm{P}(T \leq s|Z)\}).$$

If there exists $R^* : \mathbb{R}^d \to \mathbb{R}$ such that for all $s \in [0, \tau]$ and all z_1, z_2 the following ranking condition holds:

$$R^*(z_1) > R^*(z_2) \Leftrightarrow \lambda(s|Z_1 = z_1) > \lambda(s|Z_2 = z_2)$$

then R^* achieves the maximal c-index:

$$\forall R : \mathbb{R}^d \to \mathbb{R}, \quad \mathcal{C}_\tau(R) \leq \mathcal{C}_\tau(R^*).$$

In this case, the true t-year risk function $F_t : Z \mapsto \mathrm{P}(T \leq t|Z)$ also attains the maximal discrimination: $\mathcal{C}_\tau(R^*) = \mathcal{C}_\tau(F_t)$ for all $t \leq \tau$.

Obvious cases in which the sufficient condition is satisfied include those for which the true data generating model follows a Cox model: $\lambda(s|Z = z) = \lambda_0(s)\exp(z'\beta)$. Cases in which the sufficient condition is not satisfied can occur when the true data generating model is a Cox model with time dependent covariate effects: $\lambda(s|Z = z) = \lambda_0(s)\exp(z'\beta(s))$.

Proof: At time $s \in [0, t]$ the so-called incident-dynamic area under the curve (Heagerty and Zheng, 2005) is given by

$$\mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R) = \mathrm{P}(R(Z_1) > R(Z_2)|T_1 = s, T_2 > s) + 0.5\mathrm{P}(R(Z_1) = R(Z_2)|T_1 = s, T_2 > s).$$

The following result is due to Heagerty and Zheng (2005, Section 2.4). Denote by $S_T(s) = \mathrm{P}(T > s)$ and $f_T(s) = -\frac{\mathrm{d}}{\mathrm{d}s}S_T(s)$ the marginal survival and density function of T, respectively. For any $R : \mathbb{R}^d \to \mathbb{R}$

$$\mathcal{C}_\tau(R) = \int_0^\tau \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R)\, \omega_\tau(s)\mathrm{d}s, \tag{3.1}$$

where $\omega_\tau(s) = f_T(s)S_T(s)/\mathrm{P}(T_1 < T_2, T_1 \leq \tau)$.

By extending the arguments of McIntosh and Pepe (2002) to the time-dependent setting, see also Lemma 3.2 of Section 3.3 below, it follows that the incident-dynamic area under the curve is maximized by the true conditional hazard function $\lambda_s : Z \mapsto \lambda(s|Z)$:

$$\forall R : \mathbb{R}^d \to \mathbb{R}, \quad \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R) \leq \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(\lambda_s).$$

From the presumption of the theorem it follows that R^* also attains this maximal value:

$$\forall R : \mathbb{R}^d \to \mathbb{R}, \quad \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R) \leq \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R^*).$$

Since $\omega_\tau(s)$ is greater than zero and does not depend on R, and since $\mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R^*) > 0$ for all $s \in [0, \tau]$ we deduce from equation (3.1) that

$$\forall R : \mathbb{R}^d \to \mathbb{R} \quad \mathcal{C}_\tau(R) \leq \mathcal{C}_\tau(R^*).$$

The last statement of the theorem follows from the fact that under the ranking condition we have

$$\left\{\text{for all } s \in [0, \tau] \quad \lambda(s|Z = z_1) < \lambda(s|Z = z_2)\right\} \Leftrightarrow F_t(z_1) < F_t(z_2).$$

The direction $\Rightarrow$ is obvious from the relationship

$$F_t(z) = 1 - \exp\left(-\int_0^t \lambda(s|Z = z)\mathrm{d}s\right).$$

For $\Leftarrow$ assume that $F_t(z_1) < F_t(z_2)$. This can only happen if there exists $s \in [0, t]$ such that $\lambda(s|Z = z_1) < \lambda(s|Z = z_2)$ and the claim follows since the ranking condition of the theorem implies that for all $s, t \in [0, \tau]$ and all z_1, z_2

$$\lambda(s|Z = z_1) < \lambda(s|Z = z_2) \Leftrightarrow \lambda(t|Z = z_1) < \lambda(t|Z = z_2).$$

3.3. *Time-dependent AUC*

The t-year area under the ROC curve fulfills our minimal requirement (Definition 1). As mentioned in Zheng *and others* (2006) this follows directly from the results of McIntosh and Pepe (2002) and of Eguchi and Copas (2002). The following result is due to McIntosh and Pepe (2002) suitably adapted

to the time-dependent setting and given for the sake of completeness. It includes a statement about the incident-dynamic AUC which has been used in the proof of Theorem 3.1 above.

LEMMA 3.2 For all $s \geq 0$ and $t \in [0, \tau]$,

$$\forall R : \mathbb{R}^d \to \mathbb{R}, \quad \mathrm{AUC}_t(R) \leq \mathrm{AUC}_t(F_t) \quad \text{and} \quad \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R) \leq \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(\lambda_s),$$

where $F_t : Z \mapsto \mathrm{P}(T \leq t|Z)$ and $\lambda_s : Z \mapsto \lambda(s|Z)$.

Proof: In what follows, the case $t > s = 0$ gives the result $\mathrm{AUC}_t(R) \leq \mathrm{AUC}_t(F_t)$. The case $t = s \geq 0$ the result $\mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(R) \leq \mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(\lambda_s)$. For fixed t and s with $t \geq s \geq 0$ we condition on the event $\{T \geq s\}$ and consider the null hypothesis $\mathcal{H}_0 : T > t$ such that the alternative hypothesis is $\mathcal{H}_1 = \overline{\mathcal{H}}_0 : s \leq T \leq t$.

For any $\alpha \in]0, 1[$, the likelihood ratio test at level α is defined as follows. We reject $\mathcal{H}_0$ if $\mathrm{LR}_{s,t}(Z) > c_\alpha$, where $\mathrm{LR}_{s,t}(Z) = f_{Z|s\leq T\leq t}(Z) \big/ f_{Z|T>t}(Z)$, with $f_{Z|s\leq T\leq t}$ and $f_{Z|T>t}$ which denote the conditional densities of Z given $s \leq T \leq t$ and given $T > t$, and with c_α such that $\mathrm{P}\big(\mathrm{LR}_{s,t}(Z) > c_\alpha|T > t\big) = \alpha$. The Neyman–Pearson lemma states that among all Z-measurable test statistics the likelihood ratio test $\mathrm{LR}_{s,t}(Z)$ has the highest power. The power of the likelihood ratio test is $\mathrm{P}\big(\mathrm{LR}_{s,t}(Z) > c_\alpha|s \leq T \leq t\big)$. By definition the ROC curve of the random variable $\mathrm{LR}_{s,t}(Z)$ is the graph of the false positive rate $\mathrm{FP}_{s,t}(\alpha) = \mathrm{P}\big(\mathrm{LR}_{s,t}(Z) > c_\alpha|T > t\big)$ versus the true positive rate $\mathrm{TP}_{s,t}(\alpha) = \mathrm{P}\big(\mathrm{LR}_{s,t}(Z) > c_\alpha|s \leq T \leq t\big)$ obtained by varying the threshold α in the interval $[0, 1]$. Therefore $\mathrm{LR}_{s,t}(Z)$ maximizes the height of the ROC curve for all α, i.e., all along the curve, it also maximizes the area under the ROC curve $\int_0^1 \mathrm{TP}_{s,t}\big(\mathrm{FP}_{s,t}^{-1}(\alpha)\big)\mathrm{d}\alpha$. In the case $t > s = 0$, $\mathrm{AUC}_t(\mathrm{LR}_{s,t}) = \int_0^1 \mathrm{TP}_{s,t}\big(\mathrm{FP}_{s,t}^{-1}(\alpha)\big)\mathrm{d}\alpha$ and it remains to show that there is a one-to-one relationship between $F_t(Z)$ and $\mathrm{LR}_{s,t}(Z)$. Similarly, in the case $t = s \geq 0$, $\mathrm{AUC}_s^{\mathbb{I}/\mathbb{D}}(\mathrm{LR}_{s,t}) = \int_0^1 \mathrm{TP}_{s,t}\big(\mathrm{FP}_{s,t}^{-1}(\alpha)\big)\mathrm{d}\alpha$ and it remains to show a one-to-one relationship between $\lambda_s(Z)$ and $\mathrm{LR}_{s,t}(Z)$.

The two one-to-one relationships follow from the Bayes theorem:

$$\frac{O_{s,t}\,\mathrm{LR}_{s,t}(Z)}{O_{s,t}\,\mathrm{LR}_{s,t}(Z) + 1} = \begin{cases} F_t(Z) & \text{when} \quad t > s = 0 \\ \lambda_s(Z) & \text{when} \quad t = s \geq 0 \end{cases}$$

where $\lambda(s) = \frac{\mathrm{d}}{\mathrm{d}s}\left(-\log\{1 - \mathrm{P}(T \leq s)\}\right)$ denotes the marginal hazard rate at time s and where we define $O_{s,t} = \mathrm{P}(s \leq T \leq t) \big/ \big\{1 - \mathrm{P}(s \leq T \leq t)\big\}$ if $t > s = 0$ and $O_{s,t} = \lambda(s) \big/ \big\{1 - \lambda(s)\big\}$ if $t = s \geq 0$.

4. ILLUSTRATIVE EXAMPLE

Figure 2 shows a constructed example where the true data generating function F_t does not maximize the c-index. That is, there exists a function R such that $\mathcal{C}_\tau(R) > \mathcal{C}_\tau(F_t)$. In this example, Z is 1D and follows a normal distribution with mean 0 and standard deviation 3. We also fix $t = 1$ and $\tau = 1.2$. The functions shown in Figure 2 are

$$s \mapsto \lambda(s|Z) = \exp(Z)s^{\exp(Z)-1}$$
$$s \mapsto F_s(Z) = 1 - \exp(-s^{\exp(Z)})$$

for five selected values that Z can take. The five values correspond to the 0.25, 0.4, 0.5, 0.6, and 0.75 quantiles of the distribution of Z. Thus, T given Z follows a Weibull distribution with shape parameter

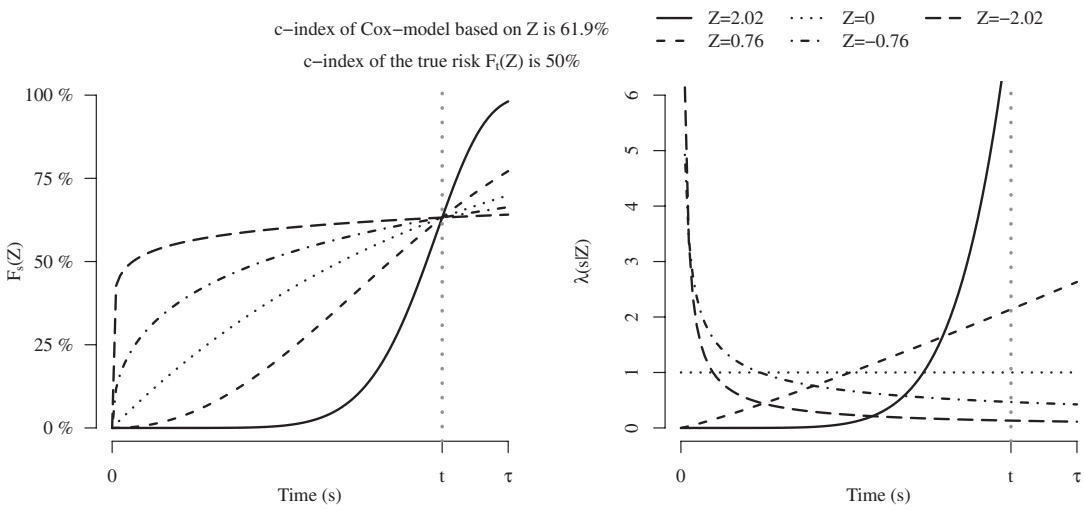

Fig. 2. Illustration of a setting in which the c-index is not appropriate to evaluate t-year predicted risks.

$\exp(Z)$ and scale parameter 1. It is shown that the marker Z discriminates the underlying risks at all time points except for t, where here $t = 1$. Obviously the true risk at $t = 1$ does not depend on Z, i.e., $F_1(Z) = 1-\exp(-1)$, and hence $\mathcal{C}_\tau(F_t) = 50\%$. However, the c-index for the identity function $R_I : Z \mapsto Z$ is computed by Monte-Carlo simulation as $\mathcal{C}_\tau(R_I) = 61.9\%$, which is higher than $\mathcal{C}_\tau(F_t) = 50\%$, and hence the c-index is not proper. Note that also the c-index estimated in the artificially censored data at time t is not proper: $\mathcal{C}_t(R_I) = 67.9\%$.

The reason this can happen is that the condition of Theorem 3.1 is violated in this case. To see this, choose for example $Z_1 = 0.76$ and $Z_2 = 0$. Since the curve $s \mapsto \lambda(s|Z = 0.76)$ crosses the horizontal line $y = \lambda(t|Z = 0) = 1$ inside the interval $[0, \tau]$ it is not possible to construct a map $R^* : \mathbb{R}^d \to \mathbb{R}$ such that for all $s \in [0, \tau]$

$$R^*(0.76) > R^*(0) \Leftrightarrow \lambda(s|Z_1 = 0.76) > \lambda(s|Z_2 = 0).$$

In the remainder of this section, we illustrate that this problem cannot be avoided by carefully checking that the prediction model under consideration satisfies the ranking condition of Theorem 3.1. More specifically we now show that when a misspecified prediction model satisfies the ranking condition but the true distribution does not, then it is possible that the misspecified model achieves a misleadingly high c-index. In the setting of Figure 2 we consider the simple Cox regression model: $\lambda(t|Z) = \lambda_0(t)\exp(Z'\beta)$. We assume that this model has been fitted using a sample of data generated in this setting and we also assume that the estimated value $\widehat{\beta}$ is positive. Because of the one-to-one relationship between the predicted risk at any time t, that is $\widehat{F}_t(Z) = 1 - \widehat{S}_0(t)^{\exp(Z\widehat{\beta})}$, and the covariate Z, this Cox model also has a c-index $\mathcal{C}_\tau(\widehat{F}_t) = 61.9\%$. Hence an obviously misspecified model would score higher than the true risk.

5. CONCLUSIONS AND DISCUSSION

We have shown that there is a problem with Harrell's c-index in the context of t-year risk predictions. Our theorem shows that the c-index is maximized by the true risk (our minimal requirement) if the underlying survival distribution satisfies a specific condition. However, in any application this condition

may or may not be satisfied, and this cannot be fully known or tested. If the condition is not satisfied, then it is possible that the true data generating risk model does not achieve the maximal c-index. Figure 2 shows an example of such a situation. In this situation, the c-index can declare that a model based on the marker has decent discriminative ability whereas the true t-year risks are the same for all marker values.

Simple risk prediction models can be derived from a Cox regression analysis, and it is commonly agreed that such models are useful. However, from a clinical point of view, it is often more realistic to suspect time-dependent effects especially when follow-up periods are long (Martinussen and Scheike, 2006). Because our minimal requirement is not generally fulfilled in this very common setting, our findings may raise serious concerns about the use of the c-index to evaluate t-year risk predictions even in the simple case where Cox prediction model is used. Testing the proportional hazards assumptions is not straightforward and goodness of fit tests often have limited power. Therefore it will generally be unclear if it is ok to use c-index, but we recommend to use AUC because then one does not have to worry about the proportional hazards assumption.

It is worth noting that in our example the t-year area under the ROC curve has value 50% for both the t-year predicted risks by the Cox model and for the true t-year risks. More generally, since the time-dependent area under ROC curve does not need any condition on the underlying survival distribution to fulfill our minimal requirement, we recommend to use it instead of the c-index when the aim is to evaluate t-year risk predictions.

Our criticism of the c-index holds only for situations in which the aim is to predict the risk of an event for a given time horizon. The c-index may still be valuable when the aim is to evaluate a correlation between the continuous event time and a prediction of (the order of) the event times.

In conclusion, we have shown that the c-index is not a proper scoring rule to evaluate t-year predicted risks. We have provided examples where the model with the more accurate t-year predicted probabilities does not have the higher c-index. We have mathematically derived a condition in which this cannot happen. However, since the condition is not satisfied in general, we can no longer recommend use of the c-index when evaluating the performance of a model that predicts t-year predicted probabilities. We suggest that analysts use the time-dependent area under the ROC curve instead of the c-index.

Acknowledgments

Conflict of Interest: None declared.

Appendix

For the sole purpose of illustration, Section 2 uses the data sets detailed in Table 2. For Section 2.1, in each data set Cox regression models were estimated adjusted for the risk factors shown in Table 2. Based on the Cox regression models, we predict the 5-year outcome risk and the calculate the inverse probability of censoring weighted (IPCW) estimate of $\mathcal{C}_\tau$ (Uno *and others*, 2011; Gerds *and others*, 2013) and AUC_t (Uno *and others*, 2007; Blanche *and others*, 2013). For simplicity, we use the marginal Kaplan–Meier method to estimate the censoring weights and circumvent issues with competing risks by studying the combined endpoints. For Section 2.2, we fit a simple Cox regression model adjusted only for standardized blood clotting time and predict the outcome risk for prediction horizons between 2 and 10 years with a 6 months interval. The R-code and data of all results are available at https://github.com/tagteam/webappendix-cindex-not-proper.

Table 2. *Details of data sets and Cox regression models. The risk factors entered additively into the linear predictor. For the* `pbc` *analysis the variables bilirubin, standardized blood clotting time and albumin were log-transformed*

Data set	References	Outcome	Risk factors
`pbc`	Therneau and Grambsch (2000)	Transplant-free survival	Edema, age, bilirubin, standardized blood clotting time, albumin
`follic`	Pintilie (2006)	Relapse-free survival	Age, hemoglobin, clinical stage, chemotherapy
`GBSG2`	Schumacher *and others* (1994)	Recurrence-free survival	Hormone therapy, age, menopause status, tumor size, tumor grade, number of positive lymph nodes, progesterone receptor, estrogen receptor
`cost`	Jørgensen *and others* (1996)	All-cause mortality	Age, sex, hypertension, ischemic heart disease, previous stroke, cholesterol, atrial fibrillation, Scandinavian Stroke Score
`recc`	Lee *and others* (2016)	Recurrence-free survival	Post-operative nomogram

REFERENCES

BLANCHE, P., DARTIGUES, J.-F. AND JACQMIN-GADDA, H. (2013). Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* **55**, 687–704.

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

CORNEC-LE GALL, E., AUDRÉZET, M.-P., ROUSSEAU, A., HOURMANT, M., RENAUDINEAU, E., CHARASSE, C., MORIN, M.-P., MOAL, M.-C., DANTAL, J., WEHBE, B. *and others* (2016). The PROPKD score: a new algorithm to predict renal survival in autosomal dominant polycystic kidney disease. *Journal of the American Society of Nephrology* **27**, 942–951.

EGUCHI, S. AND COPAS, J. (2002). A class of logistic-type discriminant functions. *Biometrika* **89**, 1–22.

GERDS, T., KATTAN, M., SCHUMACHER, M. AND YU, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32**, 2173–2184.

GNEITING, T. AND RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

HARRELL, F. E., CALIFF, R. M., PRYOR, D. B., LEE, K. L. AND ROSATI, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* **247**, 2543–2546.

HARRELL, F. E., LEE, K. L. AND MARK, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

HEAGERTY, P. J., LUMLEY, T. AND PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

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

HLATKY, M. A., GREENLAND, P., ARNETT, D. K., BALLANTYNE, C. M., CRIQUI, M. H., ELKIND, M. S. V., GO, A. S., HARRELL, F. E., HONG, Y., HOWARD, B. V. *and others* (2009). Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* **119**, 2408–2416.

JØRGENSEN, H. S., NAKAYAMA, H., REITH, J., RAASCHOU, H. O. AND OLSEN, T. S. (1996). Acute stroke with atrial fibrillation. The Copenhagen Stroke Study. *Stroke* **27**, 1765–1769.

LEE, B. H., FEIFER, A., FEUERSTEIN, M. A., BENFANTE, N. E., KOU, L., YU, C., KATTAN, M. W. AND RUSSO, P. (2016). Validation of a postoperative nomogram predicting recurrence in patients with conventional clear cell renal cell carcinoma. *European Urology Focus*, doi:10.1016/j.euf.2016.07.006.

MARTINUSSEN, T. AND SCHEIKE, T. H. (2006). *Dynamic Regression Models for Survival Data*. New York, USA: Springer.

MCINTOSH, M. W. AND PEPE, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657–664.

MOONS, K. G., ALTMAN, D. G., REITSMA, J. B., IOANNIDIS, J. P., MACASKILL, P., STEYERBERG, E. W., VICKERS, A. J., RANSOHOFF, D. F. AND COLLINS, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine* **162**, W1–W73.

MORTENSEN, R. N., GERDS, T. A., JEPPESEN, J. L. AND TORP-PEDERSEN, C. (2017). Office blood pressure or ambulatory blood pressure for the prediction of cardiovascular events. *European Heart Journal* **38**, 3296–3304.

PENCINA, M. J. AND D'AGOSTINO, R. B. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* **23**, 2109–2123.

PENCINA, M. J. AND D'AGOSTINO, R. B. (2015). Evaluating discrimination of risk prediction models: the c statistic. *JAMA* **314**, 1063–1064.

PENCINA, M. J., D'AGOSTINO, R. B., LARSON, M. G., MASSARO, J. M. AND VASAN, R. S. (2009). Predicting the 30-year risk of cardiovascular disease. *Circulation* **119**, 3078–3084.

PEPE, M. S., FAN, J., FENG, Z., GERDS, T. AND HILDEN, J. (2015). The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets. *Statistics in Biosciences* **7**, 282–295.

PINTILIE, M. (2006). *Competing Risks: A Practical Perspective*, Volume 58. Chichester, England: John Wiley & Sons.

SCHUMACHER, M., BASTERT, G., BOJAR, H., HUEBNER, K., OLSCHEWSKI, M., SAUERBREI, W., SCHMOOR, C., BEYERLE, C., NEUMANN, R. AND RAUSCHECKER, H. (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology* **12**, 2086–2093.

STEPHENSON, A. J., SCARDINO, P. T., EASTHAM, J. A., BIANCO JR, F. J., DOTAN, Z. A., DIBLASIO, C. J., REUTHER, A., KLEIN, E. A. AND KATTAN, M. W. (2005). Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of Clinical Oncology* **23**, 7005–7012.

THAS, O., NEVE, J. D., CLEMENT, L. AND OTTOY, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) **74**, 623–671.

THERNEAU, T. M. AND GRAMBSCH, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

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.

UNO, H., CAI, T., TIAN, L. AND WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102**, 527–537.

ZHENG, Y., CAI, T. AND FENG, Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62**, 279–287.

[*Received May 18, 2017; revised January 17, 2018; accepted for publication January 17, 2018*]