Advances in Methodology and Statistics, 2024, *21*(1), 1–6. https://doi.org/10.51936/qxeu3315

A cautionary note on two measures of explained variation in survival analysis

Robin Henderson[a], Damjan Manevski[b], Tina Košuta[b], Janez Stare[b,*]

[a]*Newcastle University, School of Mathematics, Statistics and Physics, Newcastle upon Tyne, United Kingdom*
[b]*University of Ljubljana, Faculty of Medicine, Ljubljana, Slovenia*

Abstract

In the paper we cast serious doubt on the usefulness of the index of concordance (C-index) and the coefficient of determination for survival models. The index of concordance is known for not being useful when selecting the best of several competing models because of its nonsensitivity. We show that it can even go down when a significant covariate is added to a correctly specified model, including cases when there is no censoring and no tied event times or covariate values, meaning that the usual suspects are not responsible. The coefficient of determination was originally suggested because its formula gives the usual R^2 when used in linear regression. But, the analogy with the linear model is gone when we use survival analysis models, since it crucially depends on the null model which changes with the change in the fitted model. We illustrate the expected behaviour of such measures and compare with an alternative Schemper–Henderson measure.

Keywords: survival analysis, explained variation, C-index, Schemper–Henderson measure

1. Introduction

Many measures of explained variation and predictive accuracy have been proposed for use in survival data modelling (Austin et al., 2017; Choodari-Oskooei et al., 2012a, 2012b). The most-cited are variants of the C-index of concordance, orginally proposed in 1982 by Harrell et al. (1982), and of a coefficient of variation, popularised by Nagelkerke (1991). The former has more than 3200 Google Scholar citations as of autumn 2023 and the latter more than 6700. Both papers continue to be highly cited, with about 1400 and 1700 Google Scholar citations respectively since 2020. Although the C-index and coefficient of determination are defined generally, our focus is on their use in survival analysis.

Recently Hartman et al. (2023) described some limitations of the C-index for survival outcomes. We agree with these, and in this short note we draw attention to what we consider to be a further and major drawback. We also take the opportunity to describe a concern we have with the use of the coefficient of determination.

*Corresponding author
Email address: janez.stare@mf.uni-lj.si (JS)
ORCID iD: 0000-0002-2564-8781 (JS)

2. The C-index can go down when it shouldn't

Table 1 shows some of the results of two Cox proportional hazards models fit to the well known Primary Biliary Cirrhosis (PBC) dataset often used for illustrating survival data methods (Fleming & Harrington, 1991). We used the version available in the `survival` package in R and removed, as is common, 25 patients who had a transplant during follow-up, to leave $n = 393$ patients and 59 % censoring. In the first model we include a single covariate, *bilirubin*, and in the second model we added *albumin* as another covariate. Results with transplants considered as either events or as censored cases were very similar to those in Table 1. The C-index is produced by the `coxph()` routine and takes into account censoring and tied data. We include also, as a comparator, an explained variation measure proposed by Schemper and Henderson (2000), which we label SH.

Table 1. Summary of Primary Biliary Cirrhosis dataset analysis.

	Model 1			Model 2		
Covariate	Coef.	Std. Err.	Z	Coef.	Std. Err.	Z
Bilirubin	0.145	0.012	12.5	0.131	0.012	10.7
Albumin				−1.318	0.194	−6.8
C-index	0.789			0.771		
SH	0.163			0.245		

Legend: SH = Schemper–Henderson measure.
Note: Likelihood ratio test for Model 2 v Model 1: 40.6 on df = 1 ($p < 2 \times 10^{-10}$)

We draw attention to the fall in C-index when Model 2 is fitted. Neither the C-index nor the SH measure is guaranteed to increase when a model is extended as they are not directly linked to the partial likelihood, which of course cannot decrease. If an unimportant covariate is added we would expect the C-index and SH to stay around the same, or perhaps fall a little. But what caught our attention in these fits was the decrease in C-index despite the addition of a highly statistically significant new covariate. The SH measure increases as we would hope when we move from Model 1 to Model 2.

The C-index can be affected by tied covariates, tied event times, censoring and model misspecification. To rule these out we generated 1000 simulated data sets, loosely based on the PBC data but without ties or censoring, with Weibull event times and covariates similar to bilirubin and albumin. We took a sample size $n = 250$, chosen to be intermediate between the number of observed events in the PBC data (161) and the total sample size (393). Weibull is of course a special case of a Cox proportional hazards model, meaning a Cox fit is correctly specified.

Figure 1 summarises the results. In the left panel (a) we compare the C-indices produced when Cox models with one and two covariates were fitted to the data. In the right panel (b) we do the same for the SH measure. The lines of equality are shown and different symbols are used for simulations where a likelihood ratio test does or does not reject the simpler one-covariate model.

The C-index fell when the second covariate was added in 14 % of simulations. It fell or increased by just a small amount (0.01 or less) in 30 % of simulations. Some 93 % of simulations produced statistically significant likelihood ratio tests when the two models were compared. Within this group, 13 % of simulations still had a reduction in C-index and 28 % had a reduction or only a very small increase. The SH measure never fell when the

second covariate was added.

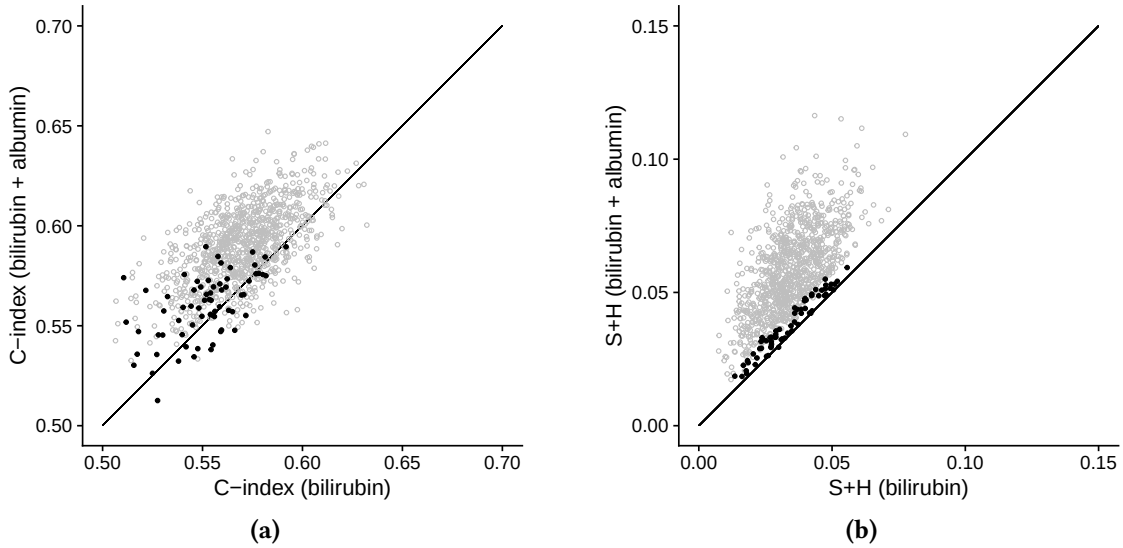

Figure 1. Simulation results. C-index (a) and SH scores (b) with one and two covariates based on PBC data. Gray dots mark simulations where a likelihood ratio test indicates significant model improvement at the 5 % level. Black dots indicate simulations with no significant model improvement.

To explore further we simulated uncensored exponential survival data with bivariate Normal covariates with standard $\mathcal{N}(0,1)$ marginals and correlation ρ. The true model had hazard $\exp(\beta_1 x_1 + \beta_2 x_2)$ and again we fitted a Cox model first with just x_1 and then with both x_1 and x_2 included. We used a likelihood ratio test to assess whether including x_2 led to an improved fit, testing at the 5 % level, and we calculated the C-index and SH for both model fits. To investigate the effect of covariate skewness we repeated the simulations with as covariates the squared values of x_1 and x_2, standardised to zero mean and unit variance. These transformed covariates were used for both simulating data and subsequent model fitting, so there was no misspecification.

Table 2 summarises some of our results. The values shown are based on 1000 repetitions of samples of size $n = 250$, with $\beta_1 = 0.25$ and $\beta_2 = 0.15$. The table shows, as percentages, the proportion $P(\text{sig})$ of simulations where adding x_2 was deemed statistically significant, and then for each of C-index and SH the overall proportion of repetitions where the measure fell when x_2 was added, $P(\text{C}\downarrow)$ and $P(\text{SH}\downarrow)$ respectively, and the proportions of repetitions with significant x_2 where the measure fell, $P(\text{C}\downarrow|\,\text{sig})$ and $P(\text{SH}\downarrow|\,\text{sig})$.

In all of the scenarios there are occasions where the C-index falls when a statistically significant covariate is added. The proportions are sometimes low, but they are not zero. The issue seems to be exacerbated by having skew covariates.

3. The coefficient of determination doesn't compare models well

Nagelkerke (1991) discusses a *coefficient of determination* that can be used in general statistical modelling. It is

$$R^2 = 1 - \exp\left[-\frac{2}{n}\{\ell(\hat{\beta}) - \ell(0)\}\right],$$

where n is the sample size and $\ell(\hat{\beta})$ and $\ell(0)$ are the maximised log-likelihoods with and without covariates. In the survival analysis context n might be replaced with the number of uncensored observations.

Table 2. Simulation results with Normal or transformed (squared) Normal covariates.

			C-index		SH	
ρ	is_tr	P(sig)	P(C↓)	P(C↓|sig)	P(SH↓)	P(SH↓|sig)
0.0	No	63.1	8.6	2.2	1.6	0.6
−0.5	No	52.0	11.9	1.9	3.8	0.0
0.5	No	54.0	11.4	3.0	2.6	0.0
0.0	Yes	61.1	10.0	4.4	1.1	0.0
−0.5	Yes	60.2	17.0	7.8	1.3	0.0
0.5	Yes	59.1	16.4	7.1	1.8	0.0

Notes: For every simulation scenario (row), we report the following quantities in the columns: the correlation value ρ, whether transformed (squared) Normal covariates are used (is_tr ∈ {No, Yes), the percentages of repetitions with statistically significant x_2 after allowing for x_1, and the percentages where the measure fell when x_2 was added, both overall and conditional on significant x_2 for the C-index (fourth and fifth column) and SH (sixth and seventh column).

To illustrate our concern, we simulated a single sample of $n = 1000$ uncensored exponential survival data with a single binary covariate, and we fitted three nested models. These, and their survival functions, are

$$\text{Exponential: } S(t) = \exp\{-e^{\beta x}t\}$$
$$\text{Weibull: } S(t) = \exp\{-e^{\beta x}t^{\sigma}\}$$
$$\text{Cox: } S(t) = \exp\{-e^{\beta x}A_0(t)\}$$

where $\sigma > 0$ and with $A_0(t)$ unspecified. The resulting coefficients of determination and SH measures are presented in Table 3.

Table 3. Calculated coefficient of determination (R^2) and Schemper–Henderson (SH) measure for the three simulated scenarios.

	Model		
	Exponential	Weibull	Cox
R^2	0.572	0.461	0.427
SH	0.265	0.265	0.265

Thus the coefficients decrease as we fit a more flexible model, which is clearly counterintuitive. Figure 2 provides the explanation. Conditional upon the covariate all three models are correctly specified and all three fit the data very well. But if covariates are excluded the true marginal survival distribution is *not* exponential and a fit of that model is poor. The Weibull model is more flexible and gives an improved fit to the marginal, and the flexible Cox model fits the marginal well. The more flexible the model the higher the $\ell(0)$ and so for the same $\ell(\hat{\beta})$ the lower the R^2. This coefficient of determination should not be used to compare different models.

Since all three models are correctly specified, a sensible measure of explained variation should give very similar values for them. As shown above, the SH measure gives equal values to the third decimal.

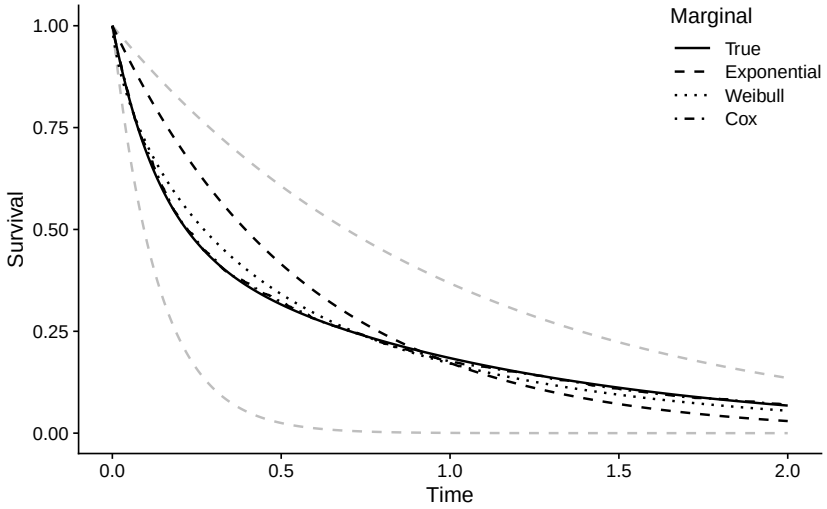

Figure 2. Fitted survival curves for coefficient of determination example. The grey dashed lines are conditional upon covariates. The three fits are almost indistinguishable from the true conditional survival curves.

4. Summary

We have added another concern to those already known about the usage of the C-index. We do not see any reason for using it.

The definition of the coefficient of determination is based on the fact that such a definition is equal to the usual R^2 in linear regression. But in linear regression the null model stays the same for different models, while in nonlinear models, including those in survival analysis, it does not. This makes this measure useless for comparing models and its interpretation essentially worthless.

Funding

This work was partially supported by the Slovenian Research and Innovation Agency under the core research programme Methodology for Data Analysis in Medical Sciences (P3-0154; DM and TK) and the research project Modelling Disease-Specific Mortality Using an Extended Multi-State Model (Z3-50124; DM).

References

Austin, P., Pencina, M., & Steyerberg, E. (2017). Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Statistical Methods in Medical Research*, *26*(3), 1053–1077. https://doi.org/10.1177/0962280214567141

Choodari-Oskooei, B., Royston, P., & Parmar, M. (2012a). A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine*, *31*(23), 2627–2643. https://doi.org/10.1002/sim.4242

Choodari-Oskooei, B., Royston, P., & Parmar, M. (2012b). A simulation study of predictive ability measures in a survival model II: Explained randomness and predictive accuracy. *Statistics in Medicine*, *31*(23), 2644–2659. https://doi.org/10.1002/sim.5460

Fleming, T., & Harrington, D. (1991). *Counting processes and survival analysis*. Wiley.

Harrell, F., Califf, R., Pryor, D., Lee, K., & Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, *247*(18), 2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

Hartman, N., Kim, S., He, K., & Kalbfleisch, J. (2023). Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, *42*(13), 2179–2190. https://doi.org/10.1002/sim.9717

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Schemper, M., & Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics*, *56*(1), 249–255. https://doi.org/10.1111/j.0006-341X.2000.00249.x