Comparison of two discrimination indexes in the categorisation of continuous predictors in time-to-event studies

Irantzu Barrio, María Xosé Rodríguez-Álvarez, Luis Meira-Machado, Cristobal Esteban and Inmaculada Arostegui

June 2017

This document contains supplementary material to the paper "Comparison of two discrimination indexes in the categorisation of continuous predictors in time-to-event studies". Web Appendix A shows the algorithms proposed to correct the optimism of the concordance probability estimators together with the proposal to select the optimal number of cutpoints in a Cox proportional hazards regression model. Web Appendix B shows the results obtained when we looked for a unique cutpoint in a univariate Cox proportional hazards regression model and finally, we briefly explain the software implementation of the methodology proposed in the main manuscript in Web Appendix C.

Web Appendix A

Optimism correction of the concordance probability estimators

The CPE was proposed as an unbiased alternative to Harrel's c-index by Gönen and Heller (2005) when the aim was to estimate the concordance probability and discriminatory power in a Cox proportional hazards regression model. Nevertheless, a discriminative ability measure estimator may be biased upward when the same data set is used to fit the model and estimate the model's discriminative ability. Hence, we proposed to correct the bias of both indexes since both were estimated using the same data that was previously used to estimate the vector of optimal cutpoints. Barrio et al. (2016) proposed a bootstrap bias correction approach for the correction of the AUC in a logistic regression setting. In this work, we have extended this proposal to the concordance probability estimators obtained from a Cox proportional hazards regression model. This algorithm can be summarised as follows:

Let us denote $\mathfrak{c}$ the concordance probability estimator, which can be either the c-index or the CPE.

Step 1. Categorise the predictor variable on the basis of the original sample $\{(x_i, \boldsymbol{z}_i, y_i, \delta_i)\}_{i=1}^{N}$ and compute the corresponding concordance probability (see equations (4) and (5) in the main manuscript). Let us denote this *apparent* concordance probability estimator as $\mathfrak{c}_{app}$.

Step 2. For $b = 1, \ldots, B$, generate the bootstrap resample $\{(x_{ib}^*, \boldsymbol{z}_{ib}^*, y_{ib}^*, \delta_{ib}^*)\}_{i=1}^{N}$ by drawing a random sample of size N with replacement from the original sample, and categorise the bootstrapped predictor $\{x_{ib}^*\}_{i=1}^{N}$ on the basis of the optimal cutpoints obtained in Step 1.

Step 3. Fit the Cox proportional hazards regression model to the bootstrap resample with the categorised version of the predictor. Let us denote as $\widehat{\boldsymbol{\beta}}^b$ the vector of the estimated regression coefficients based on this bootstrap resample. Compute the corresponding concordance probability, $\mathfrak{c}_{boot}^b$ for $b = 1, \ldots, B$.

Step 4. Obtain the linear predictor for the original sample based on the fitted Cox proportional hazards regression model obtained in Step 3, i.e,

$$\sum_{r=1}^{p} \widehat{\beta}_r^b z_{ri} + \sum_{q=p+1}^{p+k} \widehat{\beta}_q^b 1_{\{x_{cat_k i}=q\}}$$

and compute the concordance probability. Let's denote this estimator as $\mathfrak{c}_o^b$ for $b = 1, \ldots, B$.

Once the above process has been completed, the optimism O of the original concordance probability estimator is calculated as follows:

$$O = \frac{1}{B}\sum_{b=1}^{B} |\mathfrak{c}_{boot}^{b} - \mathfrak{c}_{o}^{b}|$$

and the bias-corrected concordance probability estimator is then computed as $\mathfrak{c}_{app} - O$.

Selection of the optimal number of cutpoints

Barrio et al. (2016) proposed a bootstrap confidence interval for the difference between the bias-corrected AUCs to select the optimal number of cutpoints in the categorisation of a continuous predictor variable in a logistic regression model. In this work, we have extended this proposal to obtain the optimal number of cutpoints in the categorisation of a continuous predictor variable in a Cox proportional hazards regression model. The aim is to compute a bootstrap confidence interval (CI) for the difference between the bias-corrected concordance probability of the two categorisation proposals in the Cox proportional hazards regression model in order to determine if an extra category is needed. This methodology is proposed when the maximisation index considered is either the c-index (Harrell et al., 1982) or the CPE (Gönen and Heller, 2005).

The procedure to compute the CI for the difference of the bias-corrected concordance probability estimator can be summarised as follows. For ease of notation, let us denote $\mathfrak{c}$ as the concordance probability estimator, which in our specific framework may be either the c-index or the CPE.

Step 1. For $v = 1, \ldots, V$, generate the bootstrap resample $\{(x_{iv}^{*}, \boldsymbol{z}_{iv}^{*}, y_{iv}^{*}, \delta_{iv}^{*})\}_{i=1}^{N}$ by drawing a random sample of size N with replacement from the original sample.

Step 2. Compute the bias-corrected concordance probability for the categorised variable for $k = l$ and $k = l+1$ and denote it as $\mathfrak{c}_{l,v}^{*}$ and $\mathfrak{c}_{l+1,v}^{*}$ respectively. The bias-corrected concordance probability is computed as explained above, now using for Step 1 the optimal cutpoints obtained for $k = l$ and $k = l+1$ on the basis of the original sample.

Step 3. Compute the difference between the bias-corrected concordance probabilities obtained for $k = l+1$ and $k = l$

$$\mathfrak{c}_{Diff,v}^{*} = \mathfrak{c}_{l+1,v}^{*} - \mathfrak{c}_{l,v}^{*}.$$

Once the above process has been completed, the $(1-\alpha)$ % limits for the CI for the difference are given by

$$\left(\mathfrak{c}_{Diff}^{\alpha/2}, \mathfrak{c}_{Diff}^{1-\alpha/2}\right)$$

where $\mathfrak{c}_{Diff}^{p}$ represents the p-percentile of the estimated $\mathfrak{c}_{Diff,v}^{*}$ $(v = 1, \ldots, V)$.

We propose to determine whether an extra optimal cutpoint is needed if the CI does not contain the value zero.

We conducted a simulation study to analyse the empirical performance of the bias corrected bootstrap CI when the c-index or the CPE concordance probability estimators were used. The study was performed in the same conditions as Scenario II of the main manuscript. Hence, we considered $k = 2$ as the theoretical number of cutpoints. We looked for $k = 1$, $k = 2$ and $k = 3$ number of cutpoints for censoring rates of 20% and 70%. We selected the optimal number of cutpoints using the bootstrap CI for the difference of the bias corrected estimated concordance probabilities when compared $k = 1$ vs $k = 2$ and $k = 2$ vs $k = 3$, and computed the percentage of runs in which the number of cutpoints selected were 2. Simulations were performed for a sample size of $N = 500$ and $R = 100$ replicates of simulated data. To perform the bootstrap CI $V = 100$ number of bootstrap resamples were used.

The results suggest that, when using the c-index, the optimal number of cutpoints can be selected based on the bootstrap CI for the difference of the bias corrected estimated concordance probability (Table B1). However, the results suggest that the CPE tends to select the largest number of cutpoints. Further work is therefore needed to provide accurate methods for the selection of optimal cutpoints using the CPE.

Table B1: *Percentage of replicates in which the selected number of cutpoints is k = 2, based on the 95% bootstrap CI for the difference of the bias corrected estimated concordance probability.*

Compared number of cutpoints	**Concordance probability estimator**	**Censoring Rate**	
		20%	**70%**
$k = 2$ **vs** $k = 1$	c-index	100%	81%
	CPE	100%	98%
$k = 3$ **vs** $k = 2$	c-index	100%	96%
	CPE	77%	47%

Web Appendix B: Selection of an optimal cutpoint in a univariate Cox proportional hazards regression model

The simulation study was performed in the same conditions as the ones detailed in section 3.1 of the main manuscript but considering in this case $\alpha = 0$. In particular, the simulations for the univariate Cox proportional hazards regression model were performed considering the parameters described in Table B2.

Table B2: *Description of the different scenarios considered for the simulation study in the univariate Cox proportional hazards regression model. γ and λ are the shape and scale parameters of the Weibull distribution and the censorship $C \sim U(0,\tau)$.*

Scenario	**Theoretical cutpoints**	**Parameters**	**Censorship (τ)**		
			20%	**50%**	**70%**
Ia-Univariate	0	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 2.5, \alpha = 0$	25	4.25	1.45
Ib-Univariate	1.5	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 2.5, \alpha = 0$	39	10.25	3.6
Ic-Univariate	-1.5	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 2.5, \alpha = 0$	10.5	2	0.85

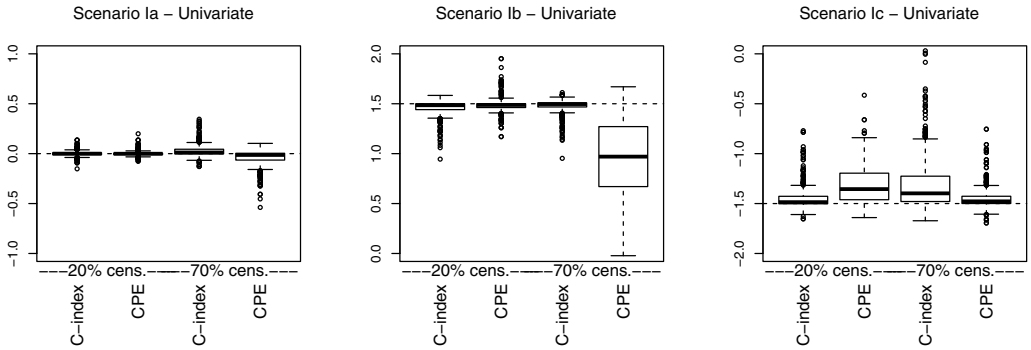

Figure B1: *Boxplot of the estimated optimal cutpoints based on 500 simulated data sets, $N = 500$ sample size and one theoretical cutpoint in a univariate Cox proportional hazards regression model. Results are shown for censoring rates of 20% and 70% and c-index and CPE concordance probability estimators. From left to right: (a) theoretical cutpoint, $c = 0$; (b) theoretical cutpoint, $c = 1.5$; and (c) theoretical cutpoint, $c = -1.5$.*

Figure B1 depicts the boxplot of the estimated optimal cutpoints over 500 simulated data sets, for the c-index and CPE estimators and a sample size of $N = 500$ for Scenarios Ia - Univariate, Ib - Univariate and Ic - Univariate, where a single optimal cutpoint is searched for. Simulation results suggest that when the theoretical optimal cutpoint is centred, i.e., $c_1 = 0$, the proposed method performs satisfactorily regardless of the concordance probability estimator used and the censorship rate. However, when the theoretical cutpoint is offset, the method is not able to find it, particularly when the censoring rate is high. At this point we must clarify the fact that depending on whether the cutpoint is shifted to the area of high risk ($c_1 = 1.5$) or low risk ($c_1 = -1.5$), differences between using the CPE or the c-index are considerable. For instance, in Scenario Ib - Univariate were the optimal cutpoint is 1.5, the cutpoints obtained with the c-index for a 70% censoring rate have a low bias whereas when using the CPE the method is not able to find the theoretical cutpoint. However, in Scenario Ic - Univariate were the optimal cutpoint is -1.5, the CPE performs better than the c-index. Differences can

Table B3: *Simulations results when one theoretical optimal cutpoints was chosen in a univariate Cox proportional hazards regression model for censoring rates of 20%, 50% and 70%. Mean, standard deviation, median, bias and mean MSE for the estimated cutpoints are reported when CPE or c-index concordance probability estimators are used as the maximisation criteria.*

Sample size	Cens.	theoretical cutpoint	Cutpoint Estimation							
			CPE				c-index			
			Mean (sd)	Median	Bias	MSE	Mean (sd)	Median	Bias	MSE
Scenario Ia - Univariate										
$N = 500$	20%	0	0.000 (0.026)	−0.001	0.000	0.001	0.000 (0.028)	0.000	0.000	0.001
	50%	0	−0.014 (0.047)	−0.005	−0.014	0.002	0.008 (0.033)	0.003	0.008	0.001
	70%	0	−0.043 (0.085)	−0.013	−0.043	0.009	0.028 (0.062)	0.012	0.028	0.005
$N = 1000$	20%	0	−0.003 (0.012)	−0.001	−0.003	0.000	0.000 (0.016)	0.000	0.000	0.000
	50%	0	−0.012 (0.023)	−0.005	−0.012	0.001	0.003 (0.019)	0.001	0.003	0.000
	70%	0	−0.033 (0.060)	−0.015	−0.033	0.005	0.013 (0.036)	0.004	0.013	0.001
Scenario Ib - Univariate										
$N = 500$	20%	1.5	1.485 (0.072)	1.488	−0.015	0.005	1.457 (0.076)	1.485	−0.043	0.008
	50%	1.5	1.420 (0.124)	1.462	−0.080	0.022	1.463 (0.077)	1.490	−0.037	0.007
	70%	1.5	0.952 (0.383)	0.970	−0.548	0.447	1.477 (0.071)	1.497	−0.023	0.006
$N = 1000$	20%	1.5	1.512 (0.086)	1.496	0.012	0.007	1.476 (0.040)	1.493	−0.024	0.002
	50%	1.5	1.468 (0.108)	1.482	−0.032	0.013	1.479 (0.045)	1.495	−0.021	0.002
	70%	1.5	0.948 (0.319)	0.956	−0.552	0.407	1.483 (0.045)	1.498	−0.017	0.002
Scenario Ic - Univariate										
$N = 500$	20%	−1.5	−1.308 (0.187)	−1.355	0.192	0.072	−1.443 (0.119)	−1.486	0.057	0.017
	50%	−1.5	−1.454 (0.088)	−1.486	0.046	0.010	−1.384 (0.195)	−1.452	0.116	0.051
	70%	−1.5	−1.444 (0.116)	−1.478	0.056	0.017	−1.296 (0.281)	−1.398	0.204	0.121
$N = 1000$	20%	−1.5	−1.327 (0.139)	−1.343	0.173	0.049	−1.468 (0.069)	−1.493	0.032	0.006
	50%	−1.5	−1.470 (0.059)	−1.491	0.030	0.004	−1.429 (0.109)	−1.476	0.071	0.017
	70%	−1.5	−1.475 (0.052)	−1.489	0.025	0.003	−1.366 (0.194)	−1.446	0.134	0.056

be observed in Figure B1. Detailed numerical results are given in Table B3. Therefore, based on these results we do not recommend the use of this method to search a unique cutpoint in a univariate setting.

Web Appendix C: Software implementation

To provide the biomedical researchers with an easy-to-use tool for categorising continuous variables in a Cox proportional hazards prediction model, the methodology described in the main manuscript has been implemented in the R programming language (R Core Team, 2016). Specifically, an R function, called `catpredi.survival`, was created, with the *Genetic* method being implemented using the R-package `rgenoud` (Mebane and Sekhon, 2011). This function has been implemented in the R package `CatPredi`. The `catpredi.survival` function provides the optimal cutpoints to categorise a continuous predictor variable in a Cox proportional hazards regression model.

The CatPredi package can be freely downloaded from https://sites.google.com/site/biostit/ lineas-de-investigacion/software/catpredi where the use of the function is presented in more detail.

References

Barrio, I., Arostegui, I., Rodríguez-Álvarez, M. X., and Quintana, J. M. (2016). A new approach to categorising continuous variables in prediction models: Proposal and validation. *Statistical Methods in Medical Research*, in press.

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: The Journal of the American Medical Association*, 247, 2543–2546.

Mebane, W. R. and Sekhon, J. S. (2011). Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software*, 42, 1–26.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.