Evaluation of survival distribution predictions with discrimination measures

Raphael Sonabend[*1,2], Andreas Bender[3], and Sebastian Vollmer[2,4,5]

[1]MRC Centre for Global Infectious Disease Analysis, Jameel Institute, Imperial College London, School of Public Health, W2 1PG, London, UK
[2]Department of Computer Science, Technische Universität Kaiserslautern, Gottlieb-Daimler-Straße 47, 67663, Kaiserslautern, Germany
[3]Department of Statistics, LMU Munich, Ludwigstr. 33, 80539, Bavaria, Germany
[4]Data Science and its Application, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Trippstadter Str. 122, 67663, Kaiserslautern, Germany
[5] Mathematics Institute, University of Warwick, Zeeman Building, CV4 7AL, Coventry, UK

December 10, 2021

Abstract

In this paper we consider how to evaluate survival distribution predictions with measures of discrimination. This is a non-trivial problem as discrimination measures are the most commonly used in survival analysis and yet there is no clear method to derive a risk prediction from a distribution prediction. We survey methods proposed in literature and software and consider their respective advantages and disadvantages. Whilst distributions are frequently evaluated by discrimination measures, we find that the method for doing so is rarely described in the literature and often leads to unfair comparisons. We find that the most robust method of reducing a distribution to a risk is to sum over the predicted cumulative hazard. We recommend that machine learning survival analysis software implements clear transformations between distribution and risk predictions in order to allow more transparent and accessible model evaluation.

*raphaelsonabend@gmail.com

1. Introduction

Predictive survival models estimate the distribution of the time until an event of interest takes place. This prediction may be presented in one of three ways, as a: i) time-to-event, $Y \in \mathbb{R}_{>0}$, which represents the time until the event takes place; ii) a relative risk, $\phi \in \mathbb{R}$, which represents the risk of the event taking place compared to other subjects in the same sample; or iii) the probability distribution for the time to the event, $\mathcal{S} \in \text{Distr}(\mathbb{R}_{>0})$, where $\text{Distr}(\mathbb{R}_{>0})$ is the set of distributions over $\mathbb{R}_{>0}$. Less abstractly, consider the Cox PH [4]: $h(t) = h_0(t) \exp(X\beta)$ where h_0 is the 'baseline' hazard function, X are covariates, and β are coefficients to be estimated. In practice, software estimates coefficients, $\hat{\beta}$, and either returns a relative risk prediction as $X\hat{\beta}$ or $\exp(X\hat{\beta})$, or h_0 is also estimated and the prediction is then a distribution represented by $\hat{h}(t) = \hat{h}_0(t) \exp(X\hat{\beta})$.

The Cox PH is a special type of survival model that can naturally return both a survival distribution and a relative risk prediction, however this is not the case for all models. For example, random survival forests [14] only return distribution predictions by recursively splitting observations into increasingly homogeneous groups and then fitting the Kaplan-Meier estimator in the terminal node.

The most common method of evaluating survival models is with discrimination measures [3, 8, 21], in particular Harrell's [10] and Uno's C [26]. These measures determine if relative risk predictions are concordant with the true event time. To give a real-world example, a physician may predict that a 70-year old patient with cancer is at higher risk of death than a 12-year old patient with a broken arm. If the 70-year old dies before the 12-year old then the risk prediction is said to be concordant with the observed event times as the patient with the predicted higher risk died first. Note that we will talk about 'risks' and 'rankings' interchangeably as discrimination measures are ranking measures that treat risk predictions as abstract rankings and ignore the distance between predicted risks.

In this paper, we will consider how to evaluate methods, which only predict survival distributions, with measures of discrimination. We will consider these methods in the context of model comparison, i.e. by establishing if the measure accurately evaluates if one model can be said to be superior to another. For example, if a given random survival forest (distribution prediction only) has better discrimination than a support vector machine (risk prediction only) [27].

Despite there being no 'obvious' method of evaluating discrimination from a distribution prediction, papers that compare (or benchmark) model discrimination, frequently omit stating the software and/or method used for evaluating distribution predictions (e.g. [5, 12, 24, 29]).

We will review methods of discrimination evaluation that are discussed in the literature and discuss their advantages and disadvantages.

2. Methods

We consider how discrimination measures are utilised in the literature to evaluate the predictive performance of models that predict survival distributions.

2.1. Literature Search

A simple search of " 'discrimination' AND 'survival analysis' " on PubMed yielded 9,000 results that could not be individually reviewed. Instead we opted to review papers pertaining to models, measures, and software that are well-established in the literature.

This included literature from the models: random survival forests [14], Cox-Time [17], DeepSurv [15], DeepHit [18], Nnet-survival [7], PCHazard [16], and DNNSurv [30]. As well as the D-Calibration measure [9].

Finally, we reviewed methods for transforming distribution to risk predictions in machine learning survival software, including **mlr3proba** [23], **pysurvival** [6], and **scikit-survival** [20]. Only **mlr3proba** included distribution-to-risk transformation methods (`PipeOpCrankCompositor`).

In these papers there are two primary solutions for evaluating distribution predictions with measures of discrimination: 1. Utilising time-dependent discrimination measures; and 2. distribution reduction methods.

We will discuss these two methods in more detail but first we define some useful notation.

In practice, software for time-to-event predictions will usually return a matrix of survival probabilities. Let $[T_0, T_N]$ be the range of observed survival times in a training dataset, let M be the number of observations in the test dataset and let M^* be the number of time-points for which predictions are made, then we predict $\mathbf{S} \in [0,1]^{M \times M^*}$, which correspond to predictions of individual survival functions, $S_i(T), T \in \mathcal{T} \subseteq [T_0, T_N]$.

2.2. Time-dependent discrimination measures

Time-dependent discrimination measures evaluate discrimination over time by comparing predicted survival probabilities over time with the true event status. Popular time-dependent concordance indices have been put forward by [11], and [1]. We found Antolini's C-index to be more popular in the artificial survival network literature [16, 17, 18].

On the surface, time-dependent discrimination measures are optimal for evaluating distributions by discrimination, however they are a poor choice for model comparison. Time-dependent measures require distribution predictions, time-independent ones require risk predictions, therefore whilst they may claim to estimate the same quantity (concordance), they are evaluating separate mathematical objects. Therefore, one could not compare (say) a Cox PH with Harrell's C against a random survival forest with Antolini's C.

2.3. Distribution reduction methods

Time-independent discrimination measures for survival analysis evaluate relative risk predictions by estimating concordance. Given two patients, $i \neq j$, with predicted relative risks of event, $\phi_i > \phi_j$, these predictions are concordant with the observed event times, (T_i, T_j), if $T_i < T_j$.

Let $\mathcal{S} \subseteq \text{Distr}(\mathbb{R}_{>0})$ be a convex set of distributions over the positive Reals; then a distribution reduction method is any function of the form: $f : \mathcal{S} \to \mathbb{R}$, which map a survival distribution prediction, $\zeta \in \mathcal{S}$, to a single relative risk, $\phi \in \mathbb{R}$. In the discrete software analogue, we consider functions $f' : [0,1]^{M^*} \to \mathbb{R}$.

We will now consider the following distribution reduction methods which are utilised in the literature and in software:

1. Evaluating the survival probability at a given time-point and using this value as the relative risk [7, 19, 22, 30, 31]

2. Summarising the distribution by its mean or median [9, 23]

3. Summing the predicted cumulative hazard over observed time-points [14]

Discrimination at a given survival time
Evaluating discrimination at a given survival time is formally defined by the distribution reduction $\phi := \hat{S}(t')$ where $\hat{S}$ is the predicted survival function and $t' \in \mathbb{R}_{>0}$ is a given survival

time. This results in assessing how well a model separates patients at a single time-point. This method has several problems: 1. it is not 'proper' in the sense that the optimal model may not maximise the concordance at t' [2]; 2. it is prone to manipulation as one could select the t' that maximises the C-index for their chosen model; and 3. if predicted survival curves overlap then evaluation at different time-points will lead to contradictory results (the observed event time will always stay the same). The above issues apply even if evaluated at several time-points.

Distribution summary

The distribution summary statistic method reduces a probability distribution to one of its summary statistics. Most commonly, the mean or median of the distribution. In theory, this should provide the most meaningful reduction with a natural interpretation and ranking, however this is not the case as the presence of censoring means that the predicted survival predictions will usually result in 'improper predictions', i.e. the basic properties of the survival function are not satisfied: $\lim_{t \to +\infty} S_T(t) \neq 0$. To see why this is the case, note that the majority of survival distribution predictions make use of a discrete estimator such as the Kaplan-Meier estimator, which is defined as follows:

$$\hat{S}(t) = \begin{cases} 1 & t < t_{(1)} \\ \prod_{i:t_{(i)} \leq t} (1 - d_i/n_i) & t \geq t_{(1)} \end{cases} \tag{2.1}$$

where d_i, n_i are the number of deaths and events (death or censoring) at ordered events times time $t_{(i)}, i = 1, \ldots, n$. By definition of this estimator, unless all observations at risk in the final time-point experience the event ($d_i = n_i$), the predicted survival probability in this last point will be non-zero.

Several methods have been considered to extrapolate predictions to fix this problem, such as dropping the last predicted probability to zero either at or just after the last observed time-point [23], or by linear extrapolation from the observed range [9] (Figure 1). However these methods require unjustifiable assumptions and result in misleading quantities. For example, dropping the survival probability to zero immediately after the study end assumes that all patients (no matter their risk) instantaneously die at the same time, which will skew the distribution mean and median towards the final event time [9]. The extrapolation method has the opposite problem, if the prediction survival curves are shallow then the extrapolated predictions can easily result in impossible (or at least highly unrealistic) values (Figure 1).

However, we note that summarising a 'proper' prediction (i.e. one that doesn't violate the limit properties) by its mean or median will provide a natural relative risk but in general the predicted distributions are rarely proper for all observations.

Sum of the predicted cumulative hazard

[14] suggest obtaining a relative risk from a distribution prediction by summing over the predicted cumulative hazard, $\hat{H}$

$$\phi := \sum_{t \in \mathcal{T}} -\log \hat{S}(t) = \sum_{t \in \mathcal{T}} \hat{H}(t) \tag{2.2}$$

This sum provides a measure of expected mortality for similar individuals [13, 14] and a closely related quantity can even be used as measure of calibration [28].

This provides an interpretable quantity that is meaningful as a relative risk: the higher the expected mortality, the greater the risk of the event. Furthermore, it does not require assumptions about the survival distribution before or after the observed time period.
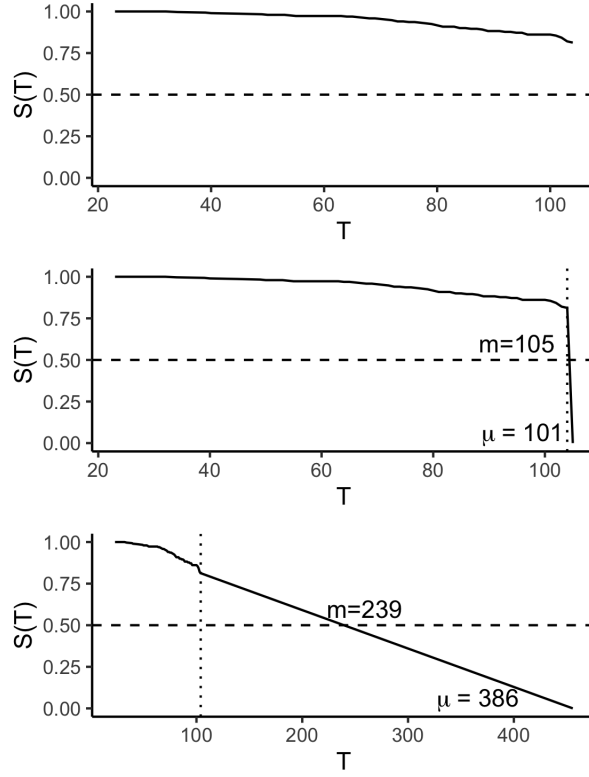
Figure 1: Extrapolation methods to 'fix' improper distribution predictions. Top: Kaplan-Meier estimator fit on the `rats` [25] dataset, which results in an improper distribution as $\lim_{T\to\infty} = 0.81 \neq 0$. Middle: Dropping the survival probability to zero at $T = 105$, just after the study end. Bottom: Dropping the survival probability to zero by linearly extrapolating from first, $(S(T) = 1, T = 0)$, and last, $(S(T) = 0.81, T = 104)$, observed survival times. Dashed horizontal lines are drawn at $S(T) = 0.5$ and dotted vertical lines at $T = 104$, where the observed data ends and the extrapolation begins. Median (m) and mean (μ) are provided for both extrapolation methods. Both methods result in quantities skewed heavily toward the final extrapolated time. For the 'dropping' method the median is exactly at the final time. Linear extrapolation results in probabilities that are unrealistically large.

5

3. Conclusions

We have surveyed the different methods used to evaluate survival distribution predictions with measures of concordance. We found that the two primary methods that are 'proper' are to either use time-dependent measures, such as that proposed by Antolini *et al.*, or to use distribution reduction measures. Whilst time-dependent measures are sensible for evaluation alone, they can not be used for model comparison between models that can and cannot make distribution predictions. For time-independent measures, several distribution reduction measures have been proposed. However, only one is interpretable whilst not requiring assumptions about predictions outside the observed time-range. Therefore, we advise that Ishwaran's method of summing over the cumulative hazard function should be utilised to provide a predicted risk value for observations. Furthermore, we believe that all open-source software should provide methods to transform distribution to risk predictions, such as the compositions in **mlr3proba** [23], in order to further transparent and accessible evaluation.

Competing interests

There is NO Competing Interest.

Author contributions statement

RS conceptualised the article. All authors contributed equally to writing and editing.

Acknowledgments

References

[1] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, dec 2005. ISSN 0277-6715. doi: 10.1002/sim.2427. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.2427.

[2] P. Blanche, M. W. Kattan, and T. A. Gerds. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy006. URL https://doi.org/10.1093/biostatistics/kxy006.

[3] G. S. Collins, J. A. De Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L. M. Yu, K. G. Moons, and D. G. Altman. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, 14(1):1–11, 2014. ISSN 14712288. doi: 10.1186/1471-2288-14-40. URL BMCMedicalResearchMethodology.

[4] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–220, 1972.

[5] T. Fernández, N. N. Rivera, and Y. W. Teh. Gaussian Processes for Survival Analysis. *Neural Information Processing Systems*, (Nips), 2016. ISSN 10495258. URL http://arxiv.org/abs/1611.00817.

[6] S. Fotso. PySurvival: Open source package for Survival Analysis modeling, 2019. URL `https://www.pysurvival.io/`.

[7] M. F. Gensheimer and B. Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. ISSN 2167-8359.

[8] M. Gönen and G. Heller. Concordance Probability and Discriminatory Power in Proportional Hazards Regression. *Biometrika*, 92(4):965–970, 2005.

[9] H. Haider, B. Hoehn, S. Davis, and R. Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020. ISSN 1533-7928.

[10] F. E. Harrell, R. M. Califf, and D. B. Pryor. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, may 1982. ISSN 0098-7484. URL `http://dx.doi.org/10.1001/jama.1982.03320430047030`.

[11] P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, 56(2):337–344, 2000. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2000.00337.x. URL `https://doi.org/10.1111/j.0006-341X.2000.00337.x`.

[12] M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *arXiv preprint arXiv:2003.03621*, 2020.

[13] D. W. Hosmer Jr, S. Lemeshow, and S. May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2011. ISBN 1118211588.

[14] B. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

[15] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. URL `https://doi.org/10.1186/s12874-018-0482-1`.

[16] H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*, 2019.

[17] H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. ISSN 1533-7928.

[18] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating Random Forests for Survival Analysis using Prediction Error Curves, 2014.

[20] S. Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1—-6, 2020. URL `http://jmlr.org/papers/v21/20-729.html`.

[21] M. S. Rahman, G. Ambler, B. Choodari-Oskooei, and R. Z. Omar. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical Research Methodology*, 17(1):1–15, 2017. ISSN 14712288. doi: 10.1186/ s12874-017-0336-2.

[22] G. Schwarzer, W. Vach, and M. Schumacher. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in Medicine*, 19(4): 541–561, feb 2000. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(20000229)19:4⟨541:: AID-SIM355⟩3.0.CO;2-V. URL https://pubmed.ncbi.nlm.nih.gov/10694735/.

[23] R. Sonabend, F. J. Király, A. Bender, B. Bischl, and M. Lang. mlr3proba: An R Package for Machine Learning in Survival Analysis. *Bioinformatics*, feb 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab039. URL https://cran.r-project.org/ package=mlr3proba.

[24] A. Spooner, E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor, and H. Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1):20410, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-77220-w. URL https://doi.org/10.1038/s41598-020-77220-w.

[25] T. M. Therneau. A Package for Survival Analysis in S, 2015. URL https://cran. r-project.org/package=survival.

[26] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, 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. ISSN 02776715. doi: 10.1002/sim.4154.

[27] V. Van Belle, K. Pelckmans, J. A. Suykens, and S. Van Huffel. Support Vector Machines for Survival Analysis. In *In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare*, number 1, 2007. doi: 10.1016/j.microrel.2005.05.002.

[28] H. C. Van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19(24):3401–3415, 2000. ISSN 02776715. doi: 10.1002/1097-0258(20001230)19:24⟨3401::AID-SIM554⟩3.0.CO;2-2.

[29] Y. Zhang, G. Wong, G. Mann, S. Muller, and J. Y. H. Yang. SurvBenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data. *bioRxiv*, page 2021.07.11.451967, jan 2021. doi: 10.1101/2021.07.11.451967. URL http://biorxiv.org/content/early/2021/09/29/2021.07.11.451967.abstract.

[30] L. Zhao and D. Feng. DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values. 2020. URL https://arxiv.org/abs/1908.02337.

[31] C. Zhong and R. Tibshirani. Survival analysis as a classification problem. sep 2019. URL http://arxiv.org/abs/1909.11171.