Model Evaluation for Time to Event Studies

Daniel Saggau

11/15/2020

Introduction

Time to event studies have gained prominence in a variety of fields, but predominately found audience within the field of medical research. Nevertheless, areas of application are manifold and include e.g. financial statistics, examining the survival of an institution. One feature of time to event studies is the unique data structure. A common characteristic of survival data is the right censored nature of observations. Right censored data means that not every subject in the study experiences the event of interest. The reasons for that can be various but one frequent reason for this is that the study ends prior to the event occurring. But this could also happen because we have subjects dropping out of our study, which might not be a common feature in financial statistics but occurs with considerable frequency in clinical or epidemiological studies.

Generally speaking, there are various model evaluation metrics. The model choice is heavily dependent on the task at hand. One way to categories models is e.g. separating them by diagnostic and prognostic.

One distinguishes between discrimination, calibration, Two prominent tools are the Receiver Operating Characteristic Curve or in short the ROC, and the c-statistics. This paper will introduce a number of methods, focusing on popular extensions of these prominent tools adjusted for survival studies namely the brier score and the c-index, two methods that have gained prominence within the realms of scholarship and among clinicians. Generally speaking, both measures, the c-index and the integrated brier score have their respective advantages and their unique merits have let these methods gained more attention than other methods. While the c-index does enjoy considerably prominence among clinicians due to interpretability and the ability to make comprehensive conclusions for the individual patients. Irrespective, the c-index only considers discrimination and does not account for calibration as an evaluation criterion which i will argue is a pivotal shortcoming when assessing different prognostic models and especially when working with machine learning methods in time to event studies. Inevitably, this paper suggests that the integrated brier score is the more holistic model evaluation metric from the stance of a statistician.

The paper is structured as follows: Firstly, i will introduce the integrated brier score, providing further information on the origin, theoretical underpinning and the respective characteristics.

Secondly, i will introduce the c-index, following the same procedure as with the integrated brier score. Following the analysis of these two cornerstones of model evaluation for time to event studies, i will illustrate further competing methods and current research within the field. Subsequently, the fourth section will provide a brief example of how to implement given methods in R. Lastly, the conclusion will summarize core findings of this brief comparative study.

Model Evaluation metrics

When attempting to evaluate a model performance, there are various approaches. The optimal model evaluation metric inevitably will depend not only on the target but also the target audience at hand. Broadly speaking, various scholars suggest that the most rudimentary distinction of model evaluation metrics is dependent on the ability to capture discrimination, "do patients who have outcome have higher risk predictions than those who dont" and model calibration, **"Measure how well predicted probabilities agree with actual observed risk".**. I will briefly introduce these concepts. Additionally to these pivotal considerations, scholarship and clinicians have also developed metrics that focus on decision analytics and reclassification measures. These tools are less prominent, but will be mentioned as they also carry merit.

Discrimination

Discrimination is the ability of a model to handle patients that do not have outcomes accordingly. Inevitably, when controlling for discrimination, we are controlling for how well our model is handling subjects with outcomes as compared to subjects without outcomes. Therefore, as the name suggest, we are testing how strong our model discriminates between subjects that incur an event versus subjects that dont. As a frequent property of survival data sets is the fact that have right censured data, data that entails subjects without the outcome/event taking place. Henceforth, discrimination is an important pillar for model evaluation in survival analysis. Various measures have emerged, that deal with discrimination such as the c-index. Perfect discrimination would imply that all our subjects with the event (e.g. a disease) have higher scores than subjects that do not have an event within their time period. One should note that when using a model that only controls for discrimination, our predictive accuracy could be horrible but as long as this condition holds, inaccurate models could be evaluated falsely as superior.

Calibration

One should mention that especially when actually applying these methods, in a clinical setting one either deals with diagnostic or prognostic tasks. Diagnostic is the analysis of a given subject at that point in time. For binary classification tasks during diagnostic studies, where we need separate between e.g. patients with and without disease, discrimination is a very important concern and potentially of greater importance. Prognostic on the other hand deals with predictive modeling, predicting e.g. in our surivival analysis setting the survival of a patient.

When dealing with an prognostic analysis, calibration can become an important concern. Calibration captures the accuracy of our predictions of our model. The underlying goal is to ensure that the predictions are as accurate as possible. For this very reason, the research community has highlighted the added value of using e.g. the integrated brier score, a score that controls for both discrimination and calibration. To further understand these methods, the following section will briefly introduce the origins of these methods, namely the ROC curve and the brier score.

Origin ROC / Concordance -statistics

Introduced by elictrical engineers during ww2.

The Receiver Operating Characteristic Curve (ROC) is an important model evaluation tool, gaining substantial prominence in various fields of statistics. This method is the foundation of the c-index which is one of the most prominent tools within the field of model evaluation for survival analysis. In a nutshell, the ROC takes into account two factors namely sensitivity and specificity. Firstly, Sensitivity deals with the likelihood of positive test results, specifically it deals with values above the threshold among the subject group which do endure an event e.g. the subjects with diseases (Cook, N. 2007). Sensitivity becomes more volatile when for instance dealing with milder, nuanced cases of a disease. Another name frequently used for Sensitivity is the true positive rate. On the other hand, specificity deals with false negatives, patients with a disease we classify as not having any diseases. However, specificity is especially subject to the influence of the characteristics of a subject without disease. Another name for specificity is the true negative rate. Examples of such characteristics are age or gender. The ROC takes these two factors and plots sensitivity against 1- specificity.

- The area under the curve or the c statistic ranges from 0.5 (no discrimination) to max of 1 (perfect discrimination)
- Essentially, the c statistic is equivalent to the probability that the measure or predicted risk is higher for a case than for a non-case.
- Further, c-statistic describes how well models rank case and noncase; but not a function of actual predicted probabilities

C-index

- c- index is the generalization of the ROC for survival data.

- Because c stat is based on ranks it is less sensitive than e.g. measures based on likelihood

- Explanation of Method

Example modification

Uno et al 2011 * C-statistic by Harell at its core is a rank correlation measure, using Kendall’s tau for censored surv. data * Problem with rank correlation: **how to order survival times**

in the presence of censoring * Brown et al.: all observations, giving prob. scores based on Kaplan Meier estimate for T * Problem: KM not good when **covariates dependent on T** * Alternative: **"usable pairs"**, excluding rest- Problem: **Dependency on censuring distribution** * Solution : **modified c-statistic which is consistent for population concordance measure, free of censoring**

- Advantages

Great popularity because so interpretable:

- Disadvantages

Problem: Studies ignoring calibation (Risk prediction models in cardiovascular literature, use c-statistic, despite working with large prospective cohort studies. Nancy Coook 2007)

Performance Evaluation metrics

Integrated Brier Score

Introduction and origin story

The score brier was initially used for weather forecasting. The general version of the brier score is also called prediction error or mean squared error(Schoop et al.,2011; Gerds & Schumacher, 2006). Henceforth, some of the applications e.g. in the 'pec' package use other terminolgy for the brier score.

The mean squared error in a nutshell is the incurred quadratic loss, studying the predicted and the true event status (Schoop et al.,2011).

Explanation of Method

IBS is dependent on the evaluation time. To get a comprehensive understanding of model performance, multiple time points have to be studied. As suggested by Bender at al. (2020), one could use the 25, 50 and 75 percent threshold.

Graf et al 1999

- expected brier score may be interpreted as a mean squared error of prediction when the estimated prob, which take values in interval [0,1] are viewed as prediction of event status at t,*I(T>t)* in {0,1}.

Advantages

IBS is a measure accounting for both discrimination and calibration.

Graf et al 1999

- **Advantage**: More sophisticated to use estimated probabilities for prediction: Diagnostic test based on predictive values ; probabilities of positive or negative disease status rather than classification of diseased or not diseased.
- Therefore: brier score, measures average discrepancies between true disease status and predicted value, better than misclassification rate

Kattan and Gerds 2018

- Separate 'useless' and 'harmful' models
- They suggest that one would assume a harmful model (incorrectly predicting certainty) having worse score relative to a useless model (a model always predicting prevalence) that predicts with some level of predictive ability
- Further Harell's c index does not provide a value specific to the time of the horizon of prediction
- This paper suggests that performance prediction ought to be specific to the time horizon of the prediction

Disadvantages

Kattan and Gerds 2018

- **BUT**: less interpretable because requires that the performance of model is compared to the performance of the best of the useless models
- Data dependence of reference value complicates Brier Score interpretation
- useless benchmark depends on overall event risk
-

Reclassification

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … & Kattan, M. W. (2010)

- Change is risk stratification.
- Needs to be appropriate.
- Use observed incidence of events of the reclassification table to predicted probabilities of the orgn. model.
- Cook proposing variant og hosmer lemeshow statistic within the reclassified categories, leading to chi-squared statistic.
-

Net Reclassification improvement

Cook, N. R., & Ridker, P. M. (2009). Advances in measuring the effect of individual predictors of cardiovascular risk:

- Net reclassification improvement (NRI)
- Integrated discrimination improvement
- IDI is equivalent to testing whether the regression coefficient in a model is equl to zero (similar to R^2 or the proportion of variance explained)
- The NRI and the IDI both condition on the case-control or later disease status
- don't provide information on calibration of the estimated risk
- A limitation of NRI and other reclassification measures is that they depend on the particular categories used
- The calibration test seems to depend somewhat less on the number of categories since the degree of freedom adjust for the number of cateogries
- Suggest that reclassification calibration statistic and NRI may be useful in demonstrating the ability of new models and markers to change risk strata and alter treatment decision

Decision Analysis Curve

One fundamental problem of the methods that we have introduced is that it does not really accommodate the interests of clinicians. From the perspective of a clinician, giving false positive and false negatives the same weight does not make any sense. A false negative entails severe repercussions relative to the false positive. For stance lets say we result in a false negative for a cancer patient, the patient is harmed to a detrimental extend and deprived of the opportunity to undertake earlier action. Moreover, these methods do not really tell us whether introducing the new model creates added value. Further, preferences may differ from a clinical standpoint. E.g. sensitivity and specificity are frequently unequal in importance to a clinician. Henceforth, scholars have proposed a new complementary framework to assess the net benefit of a model, providing a tool to assess whether implementing a new model is worth it in the first place (Vickers, A., Elkin, E., 2006). Decision analysis curve enables the use of weights, allowing optimal decision making based on subjective preferences, embodied in a net benefit equation. Vickers et al. (2016) interpret those net benefits as "clinical consequences". Further Vickers et al. (2016) illustrate that harm is transformed, using an exchange rate to put harm and benefit on one scale. This exchange rate can be obtained by asking clinicians questions based on their subjective preferences such as how many patients they would have undergo a biopsy prior to finding a cancer or weighing the benefits of getting early findings as opposed to the cost of harmful further testing. Together these elements build the net-benefit equation. Plotting different exchange rates with the net benefit equation, gives us the decision analysis curve. The curves enable the practitioner the identification of the rage of threshold probabilities for when a model would be of value, providing information on the necessary benefits needed for a model to be useful and which of many models is optimal (Vickers, A., Elkin, E., 2006).

One important consideration is that decision analysis curve is a complement, not a substitute to existing models (Vickers, A., Elkin, E., 2006).

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models.

- Compare their method to AUC method claiming that:
- AUC metric focuses solely on predictive accuracy of model
- Cannot tell us whether a model is worth using at all or which of two models is preferable
- AUC does not provide insight into usefulness aka does not account in their model that clinician may have other interests
- Two general problems: require data such as on cost or quality adjusted life years, not found in the validation data set. Cannot be evaluated without further information
- Secondly: decision analysis typically requires test or prediction model evaluated give a binary result s.t. the true and false positive and negative results can be estimated (but prediciton is frequently continuously expressed)
- Interpretation requires understanding of the liking of the patient *
- The proposed method does not require obtaining information regarding treatment preferences but need theoretical relation for the threshold probabilty of diease and the realtive value of false positive and false negative results

Implementation

Discrimination Plot

A c-index above the threshold of 0,8 can be considered good (Zhang et al.,2018).

Calibration Plot

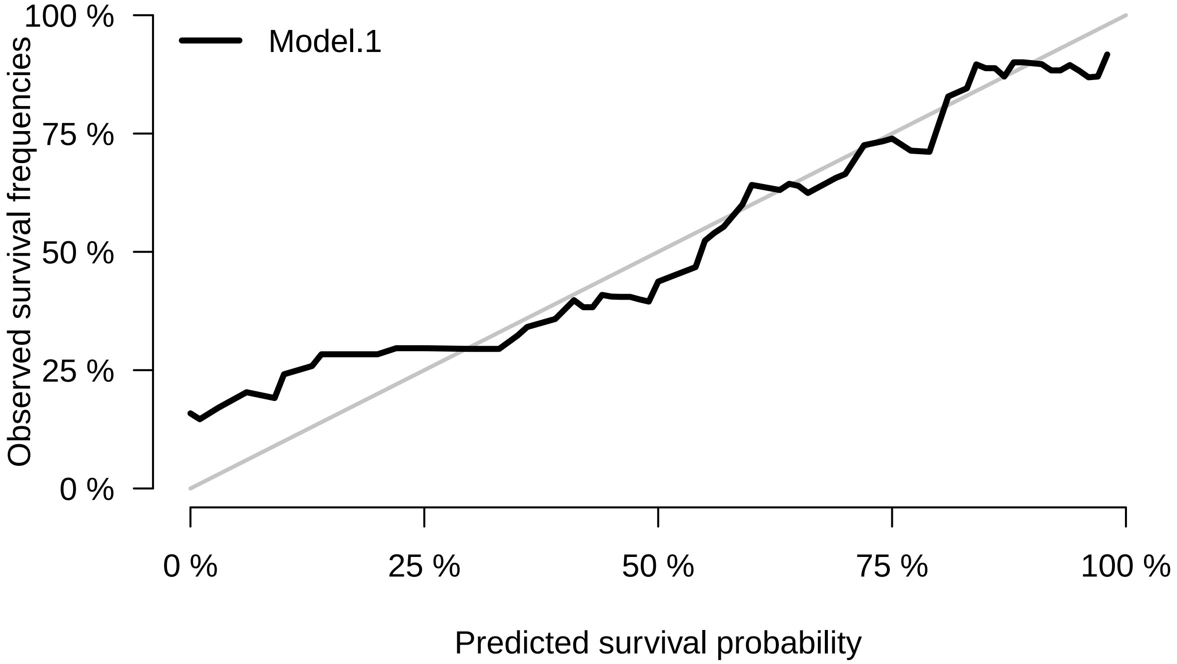
We can use calibration plots to visualize the calibration of our model. The ‘pec’ packages provides the ‘calPlot’ function.

```r
set.seed(18713)
dat=SimSurv(100)
pmodel=coxph(Surv(time,status)~X1+X2,data=dat,x=TRUE,y=TRUE)
perror=pec(list(Cox=pmodel),Hist(time,status)~1,data=dat)
```

```
## No covariates  specified: Kaplan-Meier for censoring times used for weighting.
```

```r
## cumulative prediction error
crps(perror,times=1) # between min time and 1
```

```
##
## Integrated Brier score (crps):
##
##           IBS[0;time=1)
## Reference         0.008
## Cox               0.008
## same thing:
y<- ibs(perror,times=1) # between min time and 1 crps(perror,times=1,start=0) # betwee
calPlot(pmodel)
```



A smaller brier score suggests a superior performance (Zhang et al.,2018).

Conclusion

Time to event studies require adjusted model evaluation tools for censored survival data. At the core, studies separate between models that evaluate overall performance, discrimination and calibration. New methods such as reclassification and clinical usefulness have gained prominence among scholarship within recent research, but did not achieve the same level of recognition among clinicians and in the applied research community.

References

Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. Statistics in medicine, 24(24), 3927-3944.

Bender, A., Rügamer, D., Scheipl, F., & Bischl, B. (2020). A General Machine Learning Framework for Survival Analysis. arXiv preprint arXiv:2006.15442.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation, 115(7), 928-935.

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clinical chemistry, 54(1), 17-23.

Cook, N. R., & Ridker, P. M. (2009). Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Annals of internal medicine, 150(11), 795-802.

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal, 48(6), 1029-1040.

Gerds, T. A., Cai, T., & Schumacher, M. (2008). The performance of risk prediction models. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50(4), 457-479.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine, 18(17-18), 2529-2545.

Kattan, M. W., & Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagnostic and prognostic research, 2(1), 7.

Schoop, R., Beyersmann, J., Schumacher, M., & Binder, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biometrical Journal, 53(1), 88-112.

Steyerberg, E. W., & Vickers, A. J. (2008). Decision curve analysis: a discussion. Medical Decision Making, 28(1), 146-149.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., . . . & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1), 128.

Steyerberg, E. W. (2019). Clinical prediction models. Springer International Publishing.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine, 30(10), 1105-1117.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making, 26(6), 565-574.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. bmj, 352, i6.

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 51(6), 1-36. Chicago

Zhang, Z., Cortese, G., Combescure, C., Marshall, R., Lee, M., Lim, H. J., & Haller, B. (2018). Overview of model validation for survival regression model with competing risks using melanoma study data. Annals of translational medicine, 6(16).