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 don't" 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.

Diagnostic and Prognostic studies

Prior to talking about different elements of model evaluation, we will briefly talk about the different type of studies relevant for survival data. In a clinical setting, studies can be separated into diagnostic and prognostic studies.

Diagnostic studies are concerned with the problem of how to classify a patient at this point in time- Prognostic studies on the other hand are dealing with

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 survival 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. MOre general ways to measure calibration are for instance the Hosmer Lemeshow test, the "goodness of fit" test. (Gerds and Schumacher, 2006)

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

The c- index is the generalization of the ROC for survival data (Cook, N., 2007). The c statistic is a rank correlation measure, focusing on Kendall's tau (Uno et al., 2011). An alternative is working with loss functions. Because c stat is based on ranks it is less sensitive than e.g. measures based on likelihood. Another shortcoming of the usage of rank correlation is ordering of survival times when we dont have a complete data set. Survival data frequently is censored. Therefore modifications of the traditional ROC is needed.

Example modification

Antolini et al. (2005) propose a time dependent c-index, where discrimination is summarized over time. Their model considered the presence of a population feature rather than a shortcoming of the sample. Irrespective, they also agree with the common consent and propose the use the c-index in symbiosis with a tool to measure calibration for model evaluation.

Uno et al.(2011) propose a modified c-statistic which is consistent for population concordance measures.

- Advantages

The c-index has gained popularity because so interpretable (Kattan and Gerds, 2018).

- Disadvantages

As mentioned above, for a more nuanced prevalence of a disease, the sensitivity is affected and henceforth problematic.(Cook,N., 2007) Specificity is also dependent on the data structure, but as suggested by Cook (2007), specificity is for instance affected by age, gender and the prevalence of concomitant risk factors.

Problem: Studies ignoring calibration (Risk prediction models in cardiovascular literature, use c-statistic, despite working with large prospective cohort studies. Nancy Coook 2007) While for diagnostic studies, discriminiation is the most important feature for a model evaluation metric, the same is not true for prognositic studies. For prognostic studies, model evaluation metrics also need to differentiate between useless and harmful models.

The c-index also does not account for clinical consequences and the subjective importance of false positives relative to false negatives. This problem holds for both the c-index and the brier score but generally speaking, clinical cost are different than specified in these method. The relative ability of a model to perform might not be the only question of importance in model evaluation for clinicians. For instance, other important questions could be whether introducing either model in the first place, or rephrased whether any of these models actually cause a net benefit.

Henceforth,

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

The brier score is dependent on the evaluation time. To get a comprehensive understanding of model performance, multiple time points have to be studied.

Usually one would use the 25, 50 and 75 percent thresholds.

By introducing a reweighing scheme, one derives quantities that are independent on the censoring distribution and hence suitable for censored data (Graf et al.,1999)

with uni-dimensional predictions the brier score is the same as the mean squared error.

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}.

sBrier Wu and Lee (2014) * mean squared error for the current model devided by the mean squared error for the null model

- which is equivalent mean squared gain for the current model devided by mean squared gain for an error free model

Advantages

IBS is a measure accounting for both discrimination and calibration. THe measure is an attempt to obtain an hollistic understanding of the model, rather than just looking at e.g. discrimination.

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

The benchmark of the different models are also dependent on the overall prevalence of the event in our data set. Henceforth, when working with data where the event rarely takes place, the benchmark becomes convoluted (Kattan and Gerds, 2018) Kattan and Gerds (2018) suggest that the evaluation is somewhat problematic with respect to numerous aspects. One pivotal shortcoming of the method is the inability to compare results independent from other models. Kattan and Gerds (2018) argue that one is only able to compare a model compared to other models and henceforth one is always at best only able to see that the one method is superior to the other models at hand. Especially in a practical setting, when undertaking a diagnosis for a patient this is rather impractical as patients don't come in pairs. Furthermore, this implies that we are unable to see whether the implementation of the model is advisable in the first place. Steyerberg et al. (2010) further extend this argument, by arguing that one is unable to detect whether the implementation will cause more harm than benefit. Therefore, some scholars have advocated for complementary tests, to assess the overall profitability of implementing such measures, accounting for clinical consequences, or in other words more realistic weights given the preferences of clinicians. One should note that as perhaps not obvious, clinicians don't value the different components of model accuracy equally. A false negative can be more detrimental than a false positive. For example, sending a sick patient home, in the believe of being healthy can do more damage than testing further on a healthy patient, that has been tested sick.

Reclassification

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

- Change is risk stratification.

- 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 (2008) argues that Net reclassification improvement (NRI) and calibration tests for cross classified categories can be used to study the clinical usefulness. While NRI is only a measure to study discrimination, it allows to account for the formation of categories based on clinical risk estimates. Therefore, this measure is also focusing on the clinical application, rather than holistic model evaluation. Henceforth, reclassification might just complement existing clinicians in practical applications as opposed to providing a dominant model evaluation tools.

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

- 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

Other proposals

Wu, Y. C., & Lee, W. C. (2014)

Lorenz Curve

Gini

Mean separation for the current model devided by the mean separation fro an error free model

Pietra

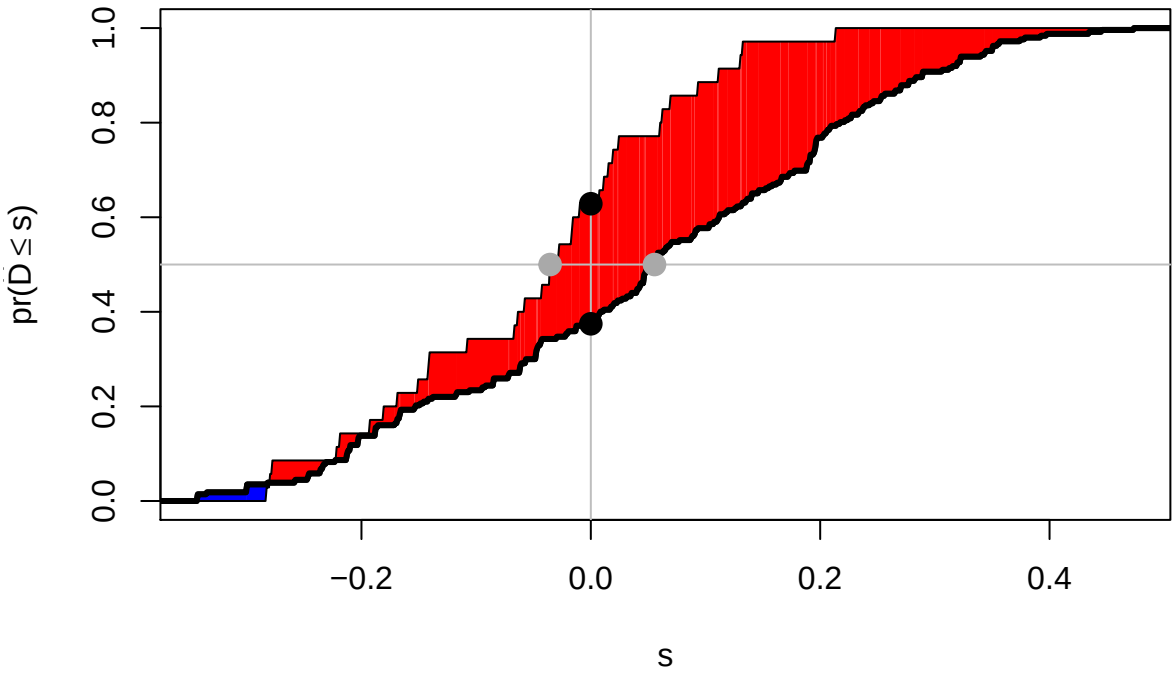
Mean gain fo current model divided by the mean gain for an error free model

Implementation

Discrimination Plot

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

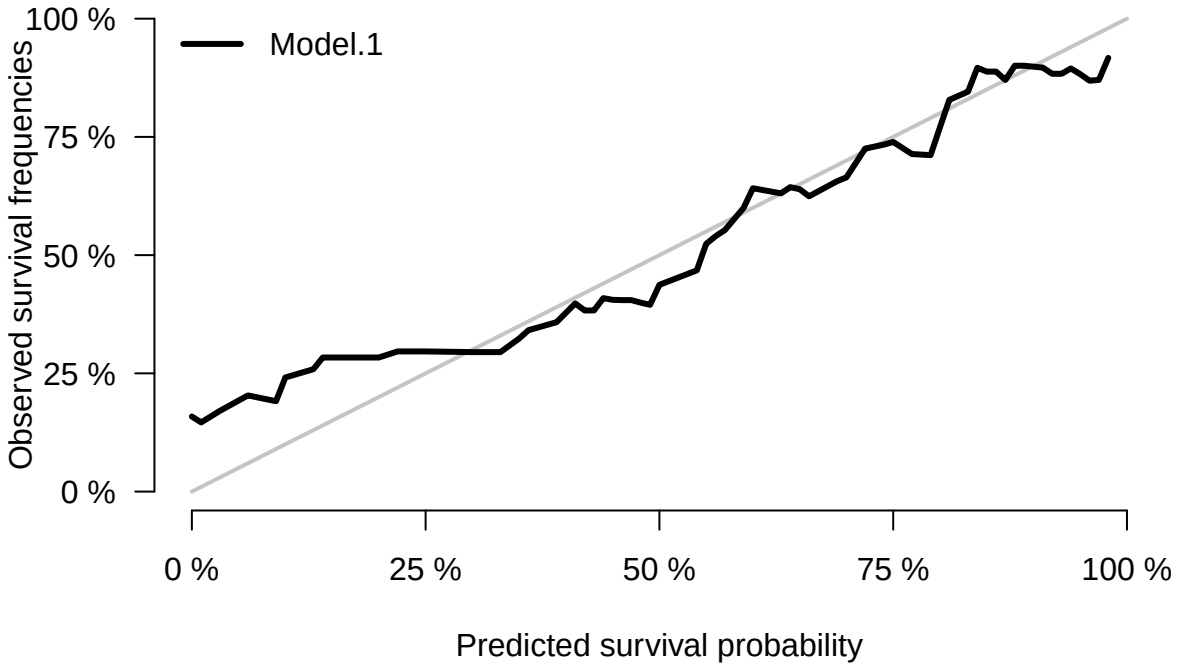
```
IDI.INF.GRAPH(res.IDI.INF)
```



Calibration Plot

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

```
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

Assel, M., Sjoberg, D. D., & Vickers, A. J. (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and prognostic research, 1(1), 1-7.

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

Wu, Y. C., & Lee, W. C. (2014). Alternative performance measures for prediction models. PLoS One, 9(3), e91249.

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).