Model Evaluation for Time-to-Event Studies

Daniel Saggau

11/22/2020

1 Introduction

A common challenge of Time-to-Event studies is dealing with censored data. Censored data implies that we don't have information on the event for all subjects in our data. Reasons for that can be manifold. One example would be that the study ends prior to the event occurring. This paper will illustrate some model evaluation metrics and respective modifications for Time-to-Event studies. Here, the focus is predominately on popular extensions of the loss function and the area under the curve(AUC), specifically drawing attention to the IBS and the concordance-index or in short c-index. Both methods use the inverse of the probability of censoring weighted estimate (IPCW) for the censored data, allowing for model evaluation despite censoring. The c-index does enjoy prominence among clinicians due to interpretability and the ability to make insightful conclusions for individual subjects. Essentially,the c-index only considers discrimination, neglecting model calibration. When trying to measure discrimination alone, it is the most popular tool of choice. The integrated brier score is the preferable tool for model evaluation for overall performance looking specifically at machine learning methods given the relative importance of calibration in predictive modelling.

The paper is structured as follows: Firstly, there is an introduction of the different components of model evaluation. The subsequent sections outline the two dominant methods, the IBS and the c-index for Time-to-Event studies. The section thereafter will provide a brief outline of some novel complementary methods. Lastly, the conclusion will summarize core findings of this brief outline.

2 Notations in Time to event Studies

Time-to-event studies (TTE) usually all entail some similar components. For every TTE study, we have a hazard function

3 Components of Model Evaluation

When evaluating model performance, one needs to differentiate between the type of study at hand. In a clinical setting, a setting that frequently welcomes time to event studies, one distinguishes between diagnostic and prognostic studies. Diagnostic studies are concerned with the problem of how to classify a patient at that very point in time. In a clinical setting, we are often interested in having a model with very high true positive rates. Prognostics on the other hand deals with predictive modeling were also accuracy becomes an eminent consideration. In the machine learning framework, we are predominately interested in prognostic studies.

Further, we can disentangle the different components of model evaluation into various groups. Fundamentally, model evaluation focuses on discrimination and calibration.

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 don't. Perfect discrimination would imply that all our subjects with the outcome have higher scores than subjects that do not have an outcome. When solely using a discrimination centered evaluation tool, our predictive accuracy could be severely impaired, but as long as we discriminate correctly we would still measure a strong performance. The most prominent summary score to evaluate discrimination for classification tasks is the area under the curve (AUC).

Calibration deals with our predictive accuracy. The most prominent score to capture accuracy in classification tasks is the brier score, a modification of the mean squared error.

A third more recent focus is the issue of clinical usefulness. Clinical usefulness is as the name suggests relevant for clinical research and henceforth of secondary focus here.

The subsequent sections will discuss the c-index and the IBS. Both of these methods use the inverse probability of the censoring weight estimate (IPCW) for the censored data, facilitating time to event data. Additionally, the c-index also changes some common assumptions used in the AUC. To follow understand where these methods come from and what differentiates them, there will be a brief outline of the AUC and the Brier Score.

3.0.1 C-Index: Foundation

The Receiver Operating Characteristic Curve (ROC), the curve in the area under the curve (AUC) score, is an important model evaluation tool for discrimination, building the foundation for the c-index. The ROC allows one to account for imbalanced label distribution and imbalanced misclassification costs. Because of these factors, we need to account for more than solely accuracy of our model. Boiling it down, we want to look at the model performance over various default thresholds rather than at a specific misclassification specification. Further, the ROC takes evaluates two factors namely sensitivity and specificity.

Sensitivity: Firstly, sensitivity deals with values above the threshold among the subject group which do endure an event e.g. the subjects with diseases (Cook, N. 2007). Another common name for Sensitivity is the true positive rate.

$$TPF = \frac{TP}{TP + FN}$$

Specificity: Specificity deals with false negatives, hence patients with a disease we classify as not having any diseases. Another name for specificity is the true negative rate.

$$TNR = \frac{TN}{TN + FP}$$

The area under the curve or the c-statistic ranges from 0.5 (no discrimination) to the maximum value of 1 (perfect discrimination). The c- index is the generalization of the ROC for survival data (Cook, N., 2007). Concordance describes consistency while discordance can be understood as inconsistency. Essentially, the difference can be written down as follows:

$$\text{AUC} = \Pr(\text{Risk}_t(\text{i}) > \text{Risk}_t(\text{j}) | \text{i has event before t and j has event after t})$$

In the AUC, we need uncensored data, because we need information on both subjects. With the c-index, we only need one of two subjects to have an event taking place for the subject pair to be comparable.

$$\text{C} = \Pr(\text{Risk}_t(\text{i}) > \text{Risk}_t(\text{j}) | \text{i has event before t})$$

Due to the fact that we deal need to be able to deal with censored data, we need to modify the AUC. The c-statistic describes how well models rank case and non-case, using a rank correlation measure, based on Kendall's tau (Uno et al., 2011). In a general case, a score above the threshold of 0,8 would be considered a strong performance (Zhang et al.,2018). Irrespective, this strongly depends on the setting. In a clinical setting this rule of thumb wont always hold. A concordance pair is a pair that is consistent, henceforth subjects with higher risks have earlier events and subjects with lower risk scores translate in later event time points.

$$\frac{\text{Concordant Pairs}}{\text{Concordant Pairs} + \text{Discordant Pairs}}$$

Together concordant pairs (consistent) and discordant pairs(inconsistent) are classified as everything that is comparable. For subject pairs to be comparable, we need at least one of the two subjects to have an event.

3.0.1.1 Mathematical derivation The AUC deals with different questions than the C-index. Typically, the AUC deals with questions like whether an Individual is likely to have a stroke within the next t-years. The c-index on the other hand evaluates pairs and therefore evaluates whether individual A or B is more likely to have a stroke. For further information, Blanche, Kattan, and Gerds (2019) explicitly elaborate why one cannot use the c-index for t-year predictions. Their arguments boil down to the following mechanism, namely that with a concordance-index we are comparing actual event times as opposed to the (time dependent) AUC which compares binary event status at time t.

insert proper formulas here

3.0.2 Modifications

Various modifications of the c-index have been in circulation, with Harell'c attracting the most attention. You can find this version in the pec package (function: 'cindex') and the survival package (function: 'survConcordance') **Population Score:** Uno et al.(2011) propose a modified c-statistic which is consistent for population concordance measures. This method is also very popular and can be found in the survAUC and the survC1 package. Another popular method is the the integrated brier score, a score that controls for both discrimination and calibration. **Time Dependency:**For time dependent covariates, Antolini et al. (2005) propose a time dependent c-index. They use a unique definition of concordance, arguing that any event that is not in the data, is bound to take place at a later point than any event that is already in the dataset (right censoring). Their model considered the presence of a population feature rather than a shortcoming of the sample.

4 Brier Score

The score brier was initially used for weather forecasting (Graf et al., 1999). With uni-dimensional predictions the brier score is the same as the mean squared error.

4.1 Mean Squared Error/ Loss Function

MSE in a Regression setting:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)}) - \hat{y}^{(i)})^2$$

The Brier Score is the MSE for Classification:

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}(x^{(i)}) - y^{(i)})^2$$

For the brier score, we are using a probability estimates $\pi(x^i)$. Other terminology that you might encounter is the predicted error or mean squared loss function (Schoop et al.,2011; Gerds & Schumacher, 2006). The 'pec' package for instance refers to to the score as the predicted error curve. The mlr3proba package refers to the brier score as the 'surv.graf', based on Graf who initially modified the measure.

4.2 Explanation of Method

The mean squared error in a nutshell is the incurred quadratic loss, studying the predicted and the true event status (Schoop et al.,2011). Graf et al. (1999) state that the "...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}." The brier score is dependent on the evaluation time. 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). To get a comprehensive time dependent model performance, multiple time points have to be studied.

For the individual at time t:

$$L(S,t|t^*) = [(S(t^*)^2)I(t) \leq t^*, \delta = 1)(\frac{1}{G(t)})] + [((1 - S(t^*))^2)I(t > t^*)(\frac{1}{G(t^*)})]$$

Where L is the loss function, S is the survival function, G is IPCW estimate of the censored survival function P(C*>t). Typically, one makes the assumption that the censored data is missing data at random, or rephased our survival times are independent.

Integrated population mean version:

$$L(S) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} L(S_i, t_i|t^*)$$

where: N is the number of observations , S_i is the predicted survival function, t is the time of the event, t^* the time before event

4.3 mlr3 implementation

Sonabend et al. (2020) provide a package for the mlr3 framework, namely mlr3proba. An useful component is the benchmarking feature of different model evaluation measures. The mlr3proba entails 5 different measures directly namely: Integrated Graf Score, Integrated Log Loss and the Log Loss.

5 Discussion

5.1 C-index

5.1.1 Advantages

Interpretation: The c-index has gained popularity because of it's interpretability (Kattan and Gerds, 2018). Especially for the individual patient in diagnostic studies, this method has gained popularity. **Evaluation:**

5.1.2 Disadvantages

Dimensionality Reduction: Because we are reducing the ROC to a single score, we are losing the biggest advantage of the ROC, namely being able to examine the model performance for different misclassification scores. Henceforth, we are defying the entire purpose of the ROC namely plotting various misclassification rates without knowing the true misclassification rate, and averaging to one single score with less information on model performance. We basically average over all misclassification rates (Hand, 2009).
Calibration: Prognostic studies need to account for the model accuracy which is measured by model calibration. Kattan and Gerds (2018), argue that model evaluation metrics needs to be able to differentiate between useless and harmful models. Harmful models are models that make severely wrong predictions (and some right ones) while useless models could e.g. always predict some level of prevalence. Using a concordance statistic for prognostic studies is not advised.
Estimators can be influenced by data: As mentioned above, for a more nuanced prevalence of a disease, the sensitivity is impaired (Cook, 2007). Specificity is 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.

Impaired Sensitivity: Because c-statistics is based on ranks it is less sensitive than e.g. measures based on probabilities.

5.2 IBS

5.2.1 Advantages

Overall Measure: The integrated brier score is a measure accounting for both discrimination and calibration separately. Essentially we are comparing two different things, once a tool to specifically look at discrimination (c-index) and secondly an overall performance measure.
Consequences of measuring Calibration : As mentioned, the integrated brier score has the

ability to differentiate between useless and harmful models.
Time Specific Horizon: As mentioned, the c-index does not allow for t-year predictions. For the IBS, we deal estimates specifically for a time-specified horizon.(Kattan and Gerds, 2018)

5.2.2 Disadvantages

Dependency on Outcome prevalence: Kattan and Gerds (2018) suggest that the evaluation is somewhat problematic. The benchmark of the different models are dependent on the overall prevalence of the event in our data set. When working with data where the event rarely takes place, the benchmark is affected (Kattan and Gerds, 2018).
Interpetation: On the one hand, scores are affected by overall event risk. On the other hand, we also need a benchmark model for evaluation. Henceforth, the interpretation of the absolute scores is problematic.

Clinical consequences: Clinicians usually don't value the different components of model evaluation equally as their clinical consequences are not equivalent. Further, we are unable to see whether the implementation of the model is advisable in the first place. Steyerberg et al. (2010) argue that one is unable to detect whether the implementation will cause more harm than benefit. Therefore, some scholars have advocated for complementary tests accounting for clinical consequences.

6 Complemenary Model Evaluation Metrics

6.1 Net Reclassification improvement

Cook (2008) advocates for the usage of net reclassification improvement (NRI) and calibration tests for cross classified categories 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. Henceforth, reclassification complements existing clinicians in practical applications as opposed to providing a dominant model evaluation tools. Integrated discrimination improvement (IDI) is equivalent to testing whether the regression coefficient in a model is equal to zero Cook, N. R., & Ridker, P. M. (2009). Cook and Ridker (2009) point out that there is a dependency between reclassification measures and the categories used. Further they suggest that reclassification calibration statistic and NRI both may be useful to demonstrate the ability of new models and markers when altering risk strata.

6.1.1 Net reclassification and integrated discrimination improvement implementation

Notable packages: 'survIDINRI':

6.2 Decision Analysis Curve

Decision analysis curve enables the use of weights, allowing optimal decision making based on subjective preferences, embodied in a net benefit equation. 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).

7 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. Both the c-index for discrimination, and the IBS for overall performance, are well established tools to undertake model evaluation.

References

Antolini, Laura, Patrizia Boracchi, and Elia Biganzoli. 2005. "A Time-Dependent Discrimination Index for Survival Data." *Statistics in Medicine* 24 (24): 3927–44. https://doi.org/10.1002/sim.2427.

Assel, Melissa, Daniel D. Sjoberg, and Andrew J. Vickers. 2017. "The Brier Score Does Not Evaluate the Clinical Utility of Diagnostic Tests or Prediction Models." *Diagnostic and Prognostic Research* 1 (1): 19. https://doi.org/10.1186/s41512-017-0020-3.

Blanche, Paul, Michael W Kattan, and Thomas A Gerds. 2019. "The c-Index Is Not Proper for the Evaluation of t-Year Predicted Risks." *Biostatistics* 20 (2): 347–57. https://doi.org/10.1093/biostatistics/kxy006.

Cook, Nancy R. 2008. "Statistical Evaluation of Prognostic Versus Diagnostic Models: Beyond the ROC Curve." *Clinical Chemistry* 54 (1): 17–23. https://doi.org/10.1373/clinchem.2007.096529.

Cook, Nancy R. 2007. "Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction." *Circulation* 115 (7): 928–35. https://doi.org/10.1161/CIRCULATIONAHA.106.672402.

Cook, Nancy R, and Paul M Ridker. 2010. "The Use and Magnitude of Reclassification Measures for Individual Predictors of Global Cardiovascular Risk," 13.

Gerds, Thomas A., Tianxi Cai, and Martin Schumacher. 2008. "The Performance of Risk Prediction Models." *Biometrical Journal* 50 (4): 457–79. https://doi.org/10.1002/bimj.200810443.

Hand, David J. 2009. "Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve." *Machine Learning* 77 (1): 103–23.

Heagerty, Patrick J., Thomas Lumley, and Margaret S. Pepe. 2000. "Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker." *Biometrics* 56 (2): 337–44. https://doi.org/10.1111/j.0006-341X.2000.00337.x.

Kamarudin, Adina Najwa, Trevor Cox, and Ruwanthi Kolamunnage-Dona. 2017. "Time-Dependent ROC Curve Analysis in Medical Research: Current Methods and Applications." *BMC Medical Research Methodology* 17 (1): 53. https://doi.org/10.1186/s12874-017-0332-6.

Kattan, Michael W., and Thomas A. Gerds. 2018. "The Index of Prediction Accuracy: An Intuitive Measure Useful for Evaluating Risk Prediction Models." *Diagnostic and Prognostic Research* 2 (1): 7. https://doi.org/10.1186/s41512-018-0029-2.

Khosla, Aditya, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee. 2010. "An Integrated Machine Learning Approach to Stroke Prediction." In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 183–92.

Munich, LMU. 2020. "Introduction to Machine Learning · A Free Interactive Course." *Introduction to Machine Learning*. https://introduction-to-machine-learning.netlify.app/.

Pencina, Michael J., Ralph B. D' Agostino, Ralph B. D' Agostino, and Ramachandran S. Vasan. 2008. "Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond." *Statistics in Medicine* 27 (2): 157–72. https://doi.org/10.1002/sim.2929.

Steyerberg, Ewout W., Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. 2010. "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures." *Epidemiology* 21 (1): 128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. 2011. "On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data." *Statistics in Medicine* 30 (10): 1105–17. https://doi.org/10.1002/sim.4154.

Vickers, Andrew J., and Elena B. Elkin. 2006. "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models." *Medical Decision Making* 26 (6): 565–74. https://doi.org/10.1177/0272989X06295361.

Vickers, Andrew J, Ben Van Calster, and Ewout W Steyerberg. 2016. "Net Benefit Approaches to the Evaluation of Prediction Models, Molecular Markers, and Diagnostic Tests." *BMJ*, January, i6. https://doi.org/10.1136/bmj.i6.