Model Evaluation for Time-to-Event Studies

Daniel Saggau

11/22/2020

1 Introduction

A common challenge of Time-to-Event studies is working with censored data. Censored data means that we don't have information on the event for all subjects in our data. Reasons for that can be manifold, but 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, focusing predominately on popular extensions of the loss function and the area under the curve(AUC), specifically focusing on the IBS and the (time dependent) c-index.

The c-index does enjoy considerable prominence among clinicians due to interpretability and the ability to make insightful conclusions for individual subjects. Irrespective, the c-index only considers discrimination, neglecting model calibration. We suggests that the integrated brier score is the preferable tool for model evaluation with predictive models. For completeness, there will be a brief outline of novel methods, suggesting potential for future research.

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 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 Components of Model Evaluation

When evaluating model performance, one needs to differentiate between the type of study at hand. In a clinical setting, studies can be separated into diagnostic and prognostic studies.

2.1 Diagnostic and Prognostic studies

Diagnostic studies are concerned with the problem of how to classify a patient at that very point in time. For binary classification tasks during diagnostic studies, where we need separate between e.g. patients with and without disease, optimal discrimination is a pivotal concern. Prognostics on the other hand deals with predictive modeling were also accuracy becomes an eminent consideration. Diagnostic is of less interest for the field of machine learning.

Further, we can disentangle the different components of model evaluation into various groups. Fundamentally, model evaluation focuses on discrimination and calibration. A third more recent focus is the issue of clinical usefulness. Clinical usefulness is as the name suggests only relevant for clinical research and henceforth of secondary focus here.

2.2 Discrimination

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 tools to study discrimination is the concordance statistics.

2.2.1 Concordance-statistics, Harell's C and the c-index

The Receiver Operating Characteristic Curve (ROC), the curve in the area under the curve (AUC) score, is an important model evaluation tool for discrimination. The ROC is the foundation for the concordance statistics (c-statistics), a modified AUC suitable for censored data (Antolini et al., 2005). Prior to discussing the c-index, this section will briefly recap the foundation of the ROC and AUC. In a nutshell, the ROC takes into account two factors namely sensitivity and specificity. 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 the maximum value of 1 (perfect discrimination).

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.

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

2.2.1.1 From AUC/ROC to concordance statistics The c- index is the generalization of the ROC for survival data (Cook, N., 2007). 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). Generally speaking, a c-statistics above the threshold of 0,8 can be considered good (Zhang et al.,2018).

2.2.2 Modifications of the AUC/ROC

Alternatively, there are a number of time dependent measures and modification of this method and the AUC/ROC curve, which are interesting when we have to time dependent covariates which don't have a 1-to-1 follow up.

Heagerty and Zheng (2005) introduce 3 modifications of the AUC, namely the (1) cumulative sensitivity and dynamic specificity (C/D), (2) incident sensitivity and dynamic specificity (I/D) and (3) incident sensitivity and static specificity (I/S).

Cumulative sensitivity and dynamic specificity: Cumulative sensitivity describes the likelihood of a subject to experience a higher score among those who already experienced the event prior to time t. Dynamic specificity is the counterpart, looking at the likelihood of subjects to have lower scores among the event free subjects surpassing point t (Kamarudin et al., 2017). This method is considered useful when dealing having specific points of time in mind (Kamarudin et al., 2017). **Incident sensitivity and dynamic specificity**: Here sensitivity is the likelihood of a subject to have a greater score among the individuals who have the event taking place at a the time point t. Respectively, the specificity is the likelihood of a subject to have a lower score among the individuals who dont have the event taking place in time t. This measure is less frequently used.(Kamarudin et al., 2017). **Incident sensitivity and static specificity**: The sensitivity is again the likelihood of a subject to have a greater score among the individuals who have the event taking place at a the time point t while the control is an event free individual for a fixed follow up period. I/D and I/S are rarely used.(Kamarudin et al., 2017).

2.2.2.1 Mathematical derivation

$$AUC^{I,D}(t) = P(X_i > c|T_i > t) \tag{5}$$

Resulting in:

$$C^T = \int_0^T AUC^{I,D}(t)w^T(t)dt \tag{6}$$

2.2.3 Modifications

Harell'c is the most commonly used C-index. Antolini et al. (2005) propose a time dependent c-index for time dependent covariates. Their model considered the presence of a population feature rather than a shortcoming of the sample. Uno et al.(2011) propose a modified c-statistic which is consistent for population concordance measures.

2.3 Advantages

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. Further, there are many well established packages in R to work with the AUC/ concordance statistics /c-index due to its popularity.

2.4 Disadvantages

Prognostic studies need to account for the model accuracy which is measured by model calibration. Calibration captures the accuracy of our predictions of our model which is very important for prediction models. One ways to measure calibration is for instance the Hosmer-Lemeshow test, the "goodness of fit" test. (Gerds and Schumacher, 2006). 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 always predict some level of prevalence. Henceforth, 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 affected and henceforth problematic (Cook,N., 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.

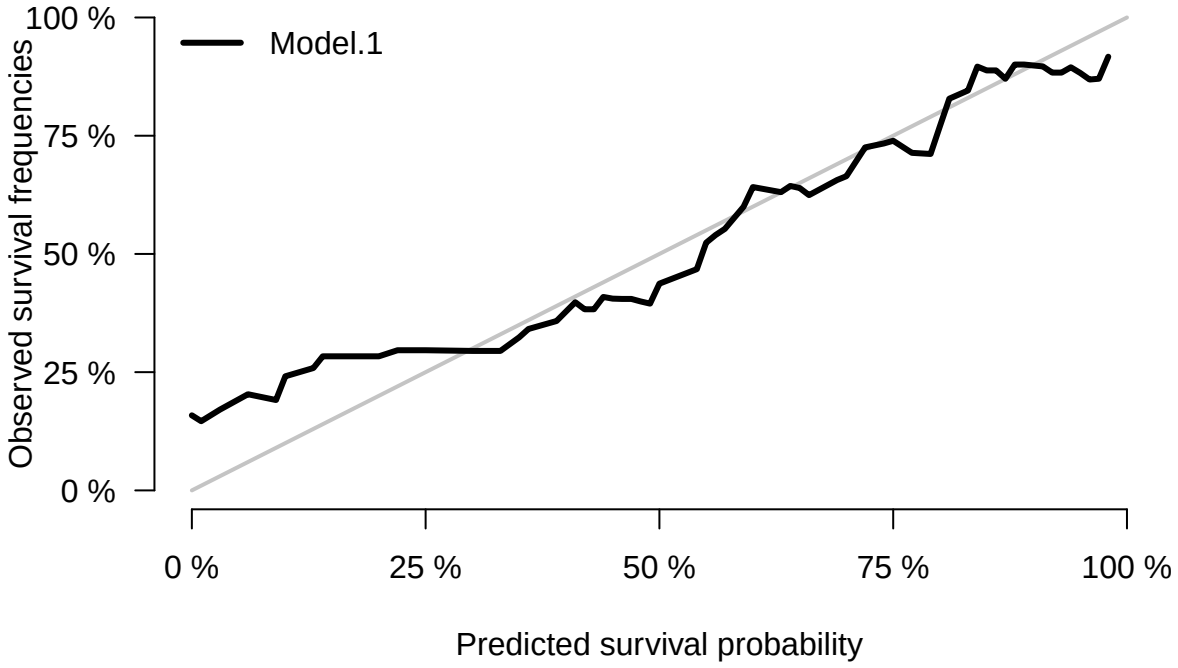
Clinical consequences: The c-index does not account for clinical consequences and the subjective importance of false positives relative to false negatives.

Another popular method is the the integrated brier score, a score that controls for both discrimination and calibration.

2.5 Calibration Plot

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

```
calPlot(pmodel)
```



3 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.

3.1 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:

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

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. ## 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:

$$L(S,t|t^*) = [(S(t^*)^2)I(t) \leq t^*, \delta = 1)(\frac{1}{G(t)})] + [((1 - S(t^*))^2)I(t > t^*)(\frac{1}{G(t^*)})] \tag{10}$$

Integrated population mean version:

$$L(S) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} L(S_i, t_i|t^*) \tag{12}$$

3.1.1 (Integrated) Log loss survival measure/ (Integrated) cross entropy

insert here

3.2 Advantages

Hollistic and Concise Statistical Underpinnings: The integrated brier score is a measure accounting for both discrimination and calibration separately. Henceforth, it is more holistic tool for model evaluation. Further, Graf et al. (1999) argue that the method is more sophisticated than the c-index because it deals with probabilities allowing us insights into the accuracy of our predictions rather than (mis-)classifications.
Differentiation of Useless and Harmful: As mentioned, the integrated brier score has the ability to differentiate between useless and harmful models.

3.3 Disadvantages

Dependency on Outcome prevalence: Kattan and Gerds (2018) suggest that the evaluation is somewhat problematic with respect to numerous aspects. The benchmark of the different models are 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).
Interpretation: One pivotal shortcoming of the method is the inability to compare results independent from other models. Hence, one is at best only able to see that the one method is superior to the other models at hand.

Clinicial 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. Two prominent tools to account for clinical consequencces are net relcassification improvement and decision analysis curves.

3.4 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:

- van Houwelingen's Alpha Calibration
- van Houwelingen's Beta Calibration
- Integrated Graf Score
- Integrated Log Loss
- Log Loss

4 Complemenary Model Evaluation Metrics

4.1 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) (somewhat similar to a R^2 score). 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.

4.1.1 Net reclassification and integrated discrimination improvement implementation

Notable packages: 'survIDINRI':

4.2 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).

5 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 discrimination and calibration, are well established tools to undertake model evaluation. New methods such as reclassification and clinical usefulness have gained prominence among scholarship.

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.

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.

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.

Pencina, Michael J., Ralph B. D' Agostino, Ralph B. D' Agostino, and Ramachandran S. Vasan. 2008a. "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.

———. 2008b. "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.