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 right censored data. Right censored data is data where not all patients have an event occurring. 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 receiver operating characteristic curve, specifically focusing on the IBS and the c-index.

The c-index does enjoy considerably prominence among clinicians due to interpretability and the ability to make comprehensive 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. For completeness, there will be a brief outline of novel methods, suggesting considering 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. After talking about some further complementary methods, the section thereafter will provide an example of how to implement given methods in R. Lastly, the conclusion will summarize core findings of this brief outline.

2 Components of Model Evaluation

When evaluating model performance, there are various components to consider. Firstly, 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 this 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 crucial concern. Prognostics on the other hand deals with predictive modeling were accuracy becomes a substantial concern. Diagnostic is of less interest for the field of machine learning.

Further, we can disentangle the different components of model evaluation into three groups. Fundamentally, model evaluation differentiates between discrimination and calibration. A third prominent evaluation criteria is clinical usefulness which will be discussed later.

2.2 Discrimination

Discrimination is the ability of a model to handle patients that do not have outcomes accordingly. 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. 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. Prominent tools to study discrimination are the concordance statistics or the net reclassification index.

2.3 Discrimination Plot

Researchers have combined the usage of reclassification tools with discrimination measures.

insert here

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

The Receiver Operating Characteristic Curve (ROC) is an important model evaluation tool for discrimination. The ROC is the foundation for the concordance statistics (c-statistics). Another name for the c-statistics is the c-index. 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).

2.3.1.1 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). Sensitivity becomes more volatile when e.g. dealing with milder, nuanced cases of a disease. Another common name for Sensitivity is the true positive rate.

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

2.3.1.2 Specificity Specificity deals with false negatives, hence 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. Examples of such characteristics are age or gender. Another name for specificity is the true negative rate.

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

2.3.1.3 From AUC/ROC to concordance statistics Due to the fact that we deal need to be able to deal with censored data, we need to modify the ROC. The c- index is the generalization of the ROC for survival data (Cook, N., 2007). 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 non-case, using a rank correlation measure, focusing on

Kendall's tau (Uno et al., 2011). The c-statistics is not a function of predicted probabilities. A c-statistics above the threshold of 0,8 can be considered good (Zhang et al.,2018).

2.3.2 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 for Time-to-Event studies.

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. As this is often the case, this method has frequently found application in clinical studies (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 and mostly not the focus of clinical studies.(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. As the second and third modification are rarely used, scholarship usually only focuses on the C/D variation.(Kamarudin et al., 2017).

2.3.2.1 Mathematical derivation

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

Resulting in:

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

2.3.3 Modifications

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.

Further measures via survAUC package:

- Uno's AUC/TPR/TNR
- Song and Zhou's AUC/TNR/TPR
- Chambless and Diao's AUC
- Hung and Chiang's AUC

Song, X., & Zhou, X. H. (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. Statistica Sinica, 947-965.

2.4 Advantages

The c-index has gained popularity because so interpretable (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. Additionally, performance is not assessed relative to a different model. Therefore evaluation does not require pairs of patients, which is more realistic.

2.5 Disadvantages

While for diagnostic studies, discrimination is the most important feature for a model evaluation metric, the same is not true for prognostic studies. Henceforth, using a concordance statistic for prognostic studies is not advised.

2.5.1 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.

2.5.2 Less sensitive

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

2.5.3 Clinical consequences

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 incorrect predictions while useless models always predict prevalence. The c-index 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.

2.6 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).

Notable mentions: * Nagelkerke's R2 * O'Quigley, Xu, and Stare's R2 * Xu and O'Quigley's R2

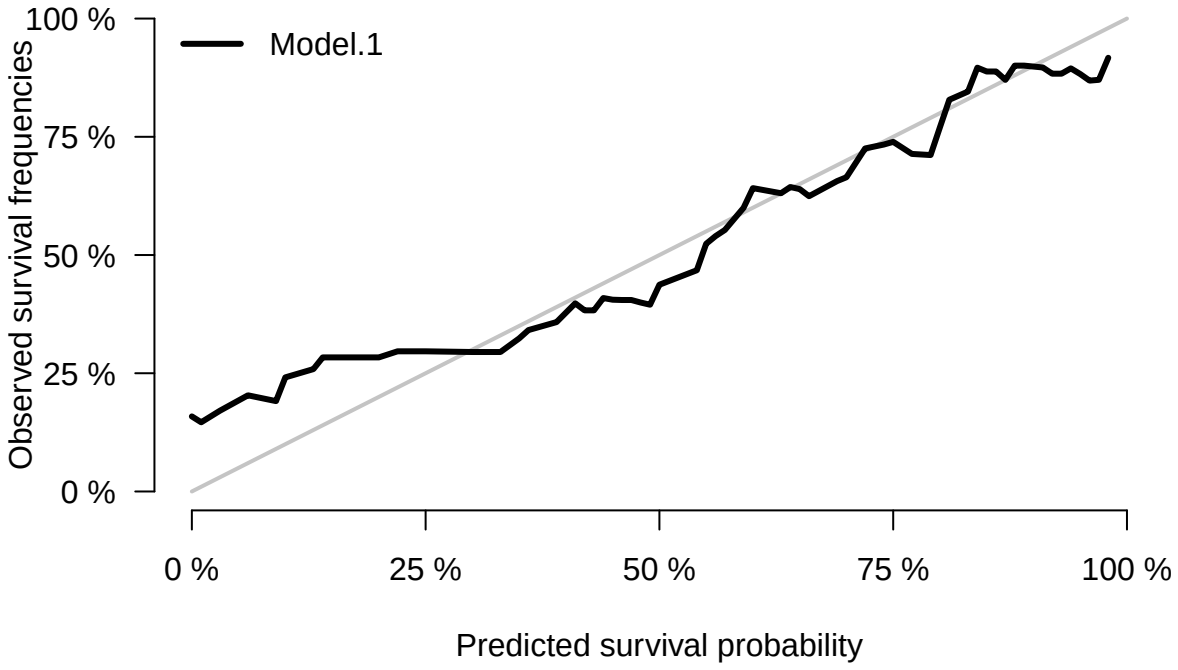
Another popular method is the the integrated brier score, a score that controls for both discrimination and calibration.

2.7 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

insert here

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.

3.2 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:

$$L(S,t|t^*) = [(S(t^*)^2)I(t) \leq t^*, \delta = 1)(\frac{1}{G(t)})] + [((1 - S(t^*))^2)I(t > t^*)(\frac{1}{G(t^*)})] \tag{8}$$

For the population mean:

$$L(S,t|t^*) = \frac{1}{N}\sum_{i=1}^{N} L(S_i, t_i|t^*) \tag{9}$$

3.3 Modifications

Integrated population mean version:

$$L(S,t|t^*) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} L(S_i, t_i|t^*) \tag{10}$$

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

insert here

3.4 Advantages

3.4.1 Hollistic Approach

The integrated brier score is a measure accounting for both discrimination and calibration separately. Henceforth, it is more holistic tool for model evaluation. One can estimate both model discrimination and calibration separately. 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.

3.4.2 Differentiation of Useless and Harmful

As mentioned, the integrated brier score has the ability to differentiate between useless and harmful models. With e.g. Harell's c index, one is unable to differentiate between the two (Kattan and Gerds, 2018).

3.5 Disadvantages

3.5.1 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).

3.5.2 Interpretation in Pairs

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. Especially in a practical setting, when undertaking a diagnosis for a patient this is rather impractical as patients don't come in pairs.

3.5.3 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.

3.6 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

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 Implementation

Notable packages: 'pec':cindex(),calPlot() 'survival': concordance() function 'survIDINRI':

6 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.