Model Evaluation for Time-to-Event Machine Learning

Author: Daniel Saggau — daniel.saggau@campus.lmu.de
Supervisor: Philipp Kopper, Andreas Bender
Department of Statistics, Ludwig Maximilian University Munich, Germany

21/12/2020

1 Introduction

We don't always have full information on the events for all subjects in our dataset. This is a common challenge of Time-to-Event studies, and referred to as censored data. Reasons for having censored data in our dataset 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 probability of the censoring weighted estimate (IPCW) for the censored data. The c-index does enjoy prominence 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. With respect to the structure of this paper, there will first of all be a brief outline of fundamental concepts within survival analysis and model evaluation. The subsequent sections devote special attention to the two dominant methods, the IBS and the c-index, also accounting for their practical implementation in R. Thereafter there is a discussion of these methods, also briefly followed by a short discourse to novel model evaluation tools focusing on clinical usefulness, namely decision curve analysis and Net Reclassification improvement. Lastly, the conclusion will summarize core findings.

2 Fundamental Ideas in Time-to-Event Studies

Time-to-event studies (TTE) entail some similar components. Firstly, we have survival time T, the time before an event takes places. For every TTE study, we have a hazard function h. Further, we can derive the cumulative hazard H. Using H, we can derive the survival function S(t). The survival function defines the probability that the event has not happened at time point t. Survival is sometimes written as Pr[T > t] which is means we are looking at the probability of the (total) survival time T being bigger or equal to our time point t. Frequently, rather than working with the hazards directly, one works with risk scores r().

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. **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 summary score to evaluate discrimination for classification tasks is the area under the curve (AUC).
Calibration: 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, the focus will be what differentiates these methods from classical tools.

2.1 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.

$$TPF = \frac{TP}{TP + FN} \tag{1}$$

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.

$$TNR = \frac{TN}{TN + FP} \tag{2}$$

The area under the curve ranges from 0.5 (no discrimination) to the maximum value of 1 (perfect discrimination). The concordance- 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:

$$\mathrm{AUC} = \Pr(\mathrm{Risk_t(i)} > \mathrm{Risk_t(j)} | \mathrm{i\ has\ event\ before\ t\ and\ j\ has\ event\ after\ t}) \tag{3}$$

In the AUC, we need uncensored data, because we need information on both subjects. On the other hand, we would specify the C score as follows:

$$\mathrm{C} = \Pr(\mathrm{Risk_t(i)} > \mathrm{Risk_t(j)} | \mathrm{i\ has\ event\ before\ t}) \tag{4}$$

With the c-index, we only need one of two subjects to have an event taking place for the subject pair to be comparable. 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.

$$\frac{\mathrm{Concordant\ Pairs}}{\mathrm{Concordant\ Pairs + Discordant\ Pairs}} \tag{5}$$

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.

2.1.1 Differentiation

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.

2.1.2 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.

2.2 Brier Score: Foundation

The Brier Score is the MSE for Classification. The MSE, the mean squared error, is a accuracy measure in a regression setting. Mathematically, we can define the MSE as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)}) - \hat{y}^{(i)})^2 \tag{6}$$

Now, we need to make some changes when working with classifications. For the brier score, we are using a probability estimates $\hat{\pi}(x^i)$ rather than estimates of y.

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}(x^{(i)}) - y^{(i)})^2 \tag{7}$$

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.

2.2.1 Brier Score: Adjustments for censored data

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 probability, which take values in interval [0,1] are viewed as prediction of event status at $t^*, I(T > t^*)$.” 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:

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

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 re-phased our survival times are independent.

Integrated population mean version:

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

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

3 Implementation

For this illustration, simulation data is used by calling the SimSurv function. In total, 3 different models are generated namely one with one variable (X1), one with a different variable (X2) and one model where we combine X1 and X2.

```
set.seed(123)
library("prodlim")
library("survival")
library("pec")
dat <- SimSurv(10000)
models <- list("Cox.X1" = coxph(Surv(time, status) ~ X1,
    data = dat, x = TRUE, y = TRUE),
  "Cox.X2" = coxph(Surv(time, status) ~ X2,
    data = dat, x = TRUE, y = TRUE),
  "Cox.X1.X2" = coxph(Surv(time, status) ~ X1 + X2,
    data = dat, x = TRUE, y = TRUE))
```

After setting a list with our different models, we can set up our model evaluation tools.

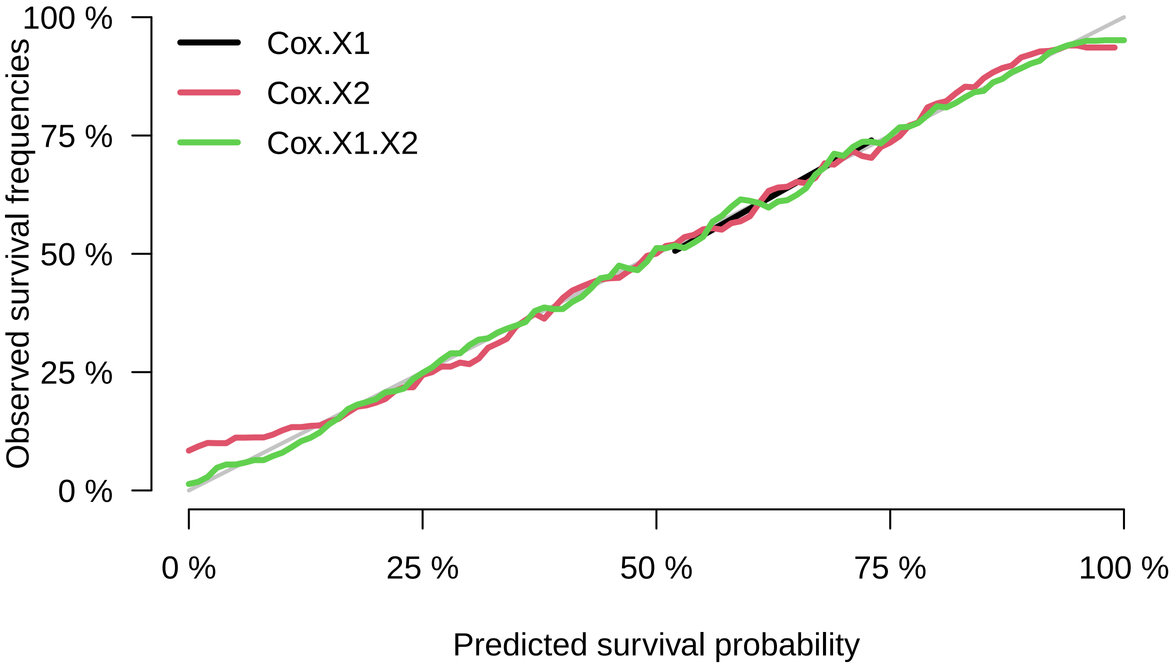
3.1 Implementation: Integrated Brier Score

Firstly, we can look at the IBS. To derive the IBS, we can separately call the Brier Score and specificy the method with which we will estimate the censored data. Here, we are using Kaplan-Meier estimates for the IPCW.

```
perror <- pec(
  object = models,
  formula = Surv(time, status) ~ 1, # ,~X1 +X2, for cox
  data = dat, exact = TRUE, cens.model = "marginal", # .model="cox"
  splitMethod = "none",
  B = 0)
```

Now, we can separately also look at the calibration of our model. The calibration plot looks at the frequencies of the survival function and compared the predicted survival probabilities. We can see that the third model is closest to the 45° line, thus the predictions are closest to the true survival frequencies. Irrespective, here the example is somewhat incomprehensible given how close the lines are and not as informative as comparing the actual scores at the different time points.

```
calPlot(models)
```



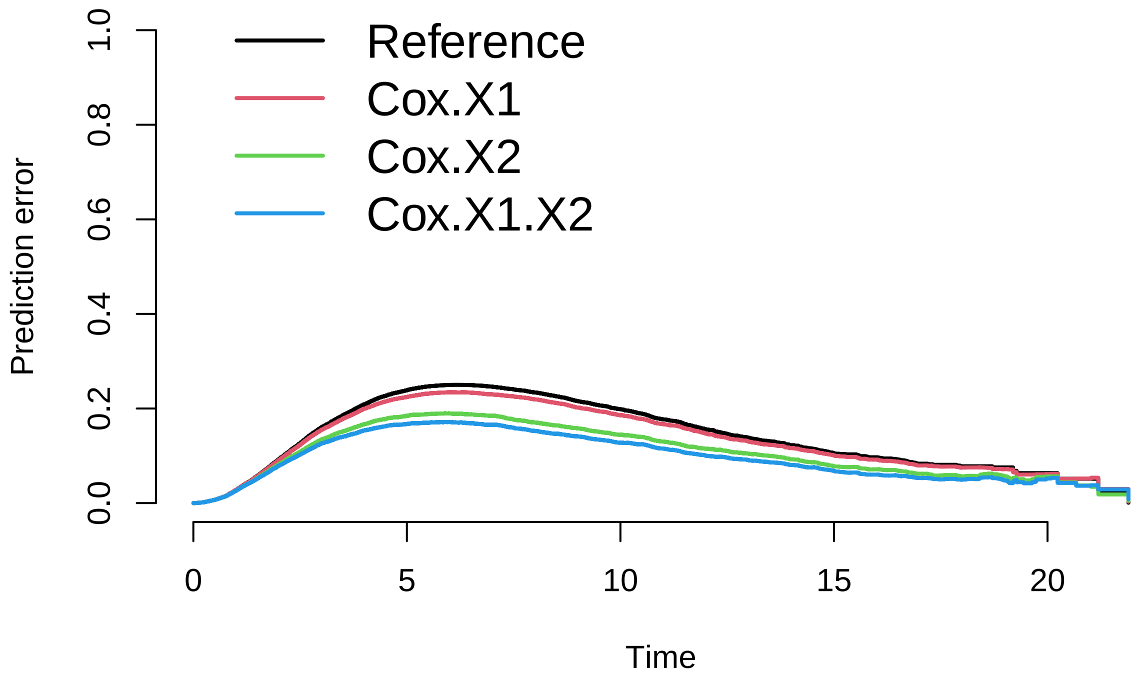
Now looking at the summary statistics for our prediction error curve at different thresholds, we can see a more detailed performance depiction of our model. Here we are examining the overall brier score and not only calibration. This is a score for calibration and discrimination combined. We can see that the third model has the lowest brier scores at all thresholds and henceforth the best performance. One should note that we looking at the score at different thresholds but this depiction is not synonymous with the IBS scores at the thresholds.

```
summary(perror, times = quantile(dat$time[dat$status == 1], c(.25, .5, .75, 1)))
```

```
##
## Prediction error curves
##
##
## No data splitting: either apparent or independent test sample performance
##
##  AppErr
##      time n.risk Reference Cox.X1 Cox.X2 Cox.X1.X2
## 1  2.568   7892     0.132  0.128  0.112     0.106
## 2  4.270   5644     0.220  0.208  0.174     0.159
## 3  6.513   3179     0.249  0.233  0.188     0.169
## 4 21.189      1     0.026  0.030  0.018     0.029
```

So, now we can visually also look at overall performance visually. One can plot our prediction error over time. A lower prediction error is better. Therefore, the third model, the blue line is lowest.

```
plot(perror, ylim = c(0,1))
```



Now, we can get a detailed look into the integrated scores, using the cumulative prediction error curves. This ‘crps' function is synonymous with the ‘ibs' function and can be used to get the integrated brier score. We can display the scores at various thresholds. The same interpretation as for the brier score holds.

```r
crps(perror, times = quantile(dat$time[dat$status == 1], c(.25, .5, .75, 1)))
```

```
##
## Integrated Brier score (crps):
##
##             IBS[0;time=2.6) IBS[0;time=4.3) IBS[0;time=6.5) IBS[0;time=21.2)
## Reference             0.051           0.102           0.150            0.142
## Cox.X1                0.050           0.099           0.143            0.134
## Cox.X2                0.046           0.086           0.120            0.108
## Cox.X1.X2             0.044           0.081           0.111            0.097
```

3.2 Implementation: Concordance Statistics

We are using the same simulated data here for comparison. Here the ‘cindex' function from the pec package is illustrated. A specification for the censored data is needed. Again, we are using the default settings, thus we are using the Kaplan-Meier-estimates.

```r
cindex <- cindex(models,
  formula = Surv(time, status) ~ 1,
  cens.model = "marginal", data = dat,
  eval.times = quantile(dat$time[dat$status == 1], c(.25, .5, .75, 1)))
```

Interpretation is reversed for the c-index. As a reminder, here we are looking at discrimination and not overall model performance. Essentially, we would do this when we want to study discrimination separately from overall performance. A score at 1 would describe a perfect model and a score of 0.5 would imply complete randomness.

```r
summary(cindex)
```

```
##
## The c-index for right censored event times
##
## Prediction models:
##
##    Cox.X1    Cox.X2 Cox.X1.X2
##    Cox.X1    Cox.X2 Cox.X1.X2
##
## Right-censored response of a survival model
##
## No.Observations: 10000
##
## Pattern:
##                 Freq
##  event          6045
##  right.censored 3955
##
## Censoring model for IPCW: marginal model (Kaplan-Meier for censoring distribution)
##
## No data splitting: either apparent or independent test sample performance
##
## Estimated C-index in %
##
## $AppCindex
##           time=2.6 time=4.3 time=6.5 time=21.2
## Cox.X1        60.5     60.2     59.6      58.8
## Cox.X2        74.8     73.2     72.1      71.0
## Cox.X1.X2     77.3     76.2     75.4      74.4
```

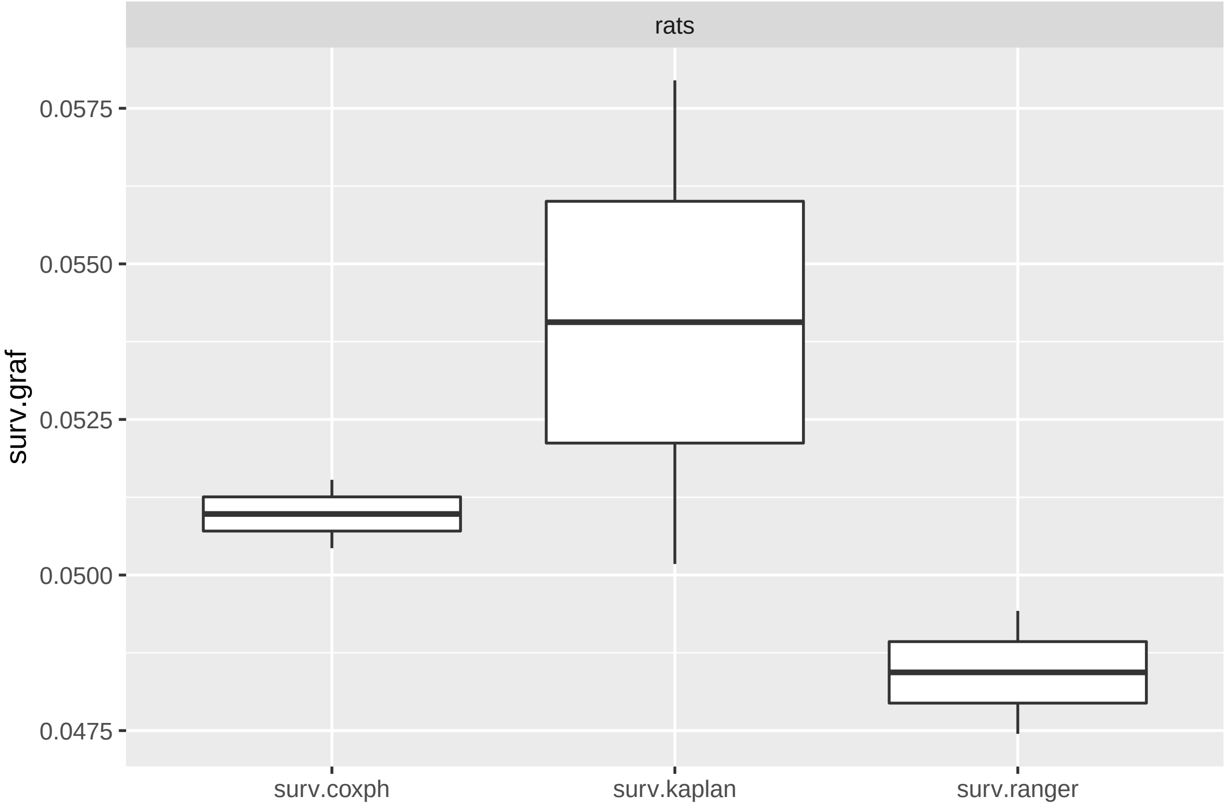
```
## Warning in summary.Cindex(cindex): The C-index is not proper for t-year predictions. Blanche et al. (2018), Bios
##
## Consider using time-dependent AUC instead: riskRegression::Score
```

3.3 Implementation: mlr3

Lastly, we can also compare model performance in the mlr3 framework, using ‘mlr3proba’ (Sonabend et al., 2020). Not going into the details of the general usage of mlr3, here the focus is the implementation of model evalation tools. To specify the measure, we define the measure. To use the IBS, we can use e.g. the ‘surv.graf’ measure and for the c-index we could use “surv.cindex”. Various different versions of these measures exist in the framework, henceforth there are further options that wont be explored at this point. Special attention should be drawn to the fact that you need to specify how censored data is treated here. E.g. the ‘surv.logloss’ function requires the user to specify how to treat the censored observations and the default is to ignore censored data. We can benchmark these results in a boxplot, using the ‘autoplot’ function.

```r
#' TaskSurv$new(
#' id = "interval_censored", backend = survival::bladder2[, -c(1, 7)],
#' time = "start", time2 = "stop", type = "interval2")
#' task <- tsk("rats")
#' learners <- lrns(c("surv.coxph", "surv.kaplan", "surv.ranger"))
#' measure <- msr("surv.graf") # for c-index you can use surv.cindex
#' bmr <- benchmark(benchmark_grid(task, learners, rsmp("cv", folds = 2)))

autoplot(bmr, measure = measure)
```



4 Considerations

4.1 Considerations: C-index

Advantages: (1) 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. (2) When using the c-index, we can separate insights into classification rather than just an aggregate score. Henceforth, this score provides insights enabling us to dissect the performance for this specific consideration.

Disadvantages: (1) 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 part of the 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). (2) 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. (3) 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. (4) Because c-statistics is based on ranks it is less sensitive than e.g. measures based on probabilities.

4.2 Considerations: IBS

Advantages: (1) 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. (2) The integrated brier score has the ability to differentiate between useless and harmful models. (3) 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)

Disadvantages:(1) 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). (2) 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. (3) 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.

5 Complemenary Model Evaluation Metrics

5.1 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 & Ridker , 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.

5.2 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).

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 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.