Model Evaluation for Time-to-Event Machine Learning

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

21/12/2020

1 Introduction

When dealing with Time-to-Event (TTE) studies, one does not always have full information on the time when events occur for every subjects. This is a common challenge of TTE studies, and referred to as censored data. Reasons for having censored data can be manifold. One reason could be that the study ends prior to the event taking place. Having censored data makes it challenging to use classical model evaluation tools and requires some modifications. The most popular tools for model evaluation for TTE studies are the Integrated Brier Score (IBS) and the concordance-index (c-index). Both methods use a weighting scheme such as the inverse probability of the censoring weighted estimate (IPCW) to approximate the censored data. Moreover, the c-index does enjoy prominence due to it's interpretability but it only measures discrimination and neglects model calibration. When trying to measure discrimination distinct from overall performance, it is the recommended tool of choice as suggested in the literature. For overall performance evaluation, the IBS is the most prominent metric, despite being less interpretable. The trajectory of this paper is to provide a brief introduction into existing model evaluation tools for TTE studies. With respect to the structure of this paper, first of all there will 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. The section thereafter discusses considerations for these methods, also briefly followed by a short discourse to novel model evaluation tools focusing on clinical usefulness. Lastly, the conclusion will summarize core findings.

2 Fundamental Concepts

2.1 Concepts in Time-to-Event Studies

Time-to-event studies (TTE) entail some common characteristics. Firstly, we have survival time 'T', the time before an event takes places. For every TTE study, we have a hazard function 'h'. Two popular hazard model types are (1) proportional hazard models (e.g. Cox-PH-Model) and (2) the non-parametric hazard models such as the Kaplan-Meier Model (Gerds et al., 2008). Further, using the hazard function we can derive the cumulative hazard 'H'. Using 'H', we can compute the survival function 'S(t)'. The survival function defines the probability that the event has not happened at time point 't'. Graf et al. (1999) refer to this as "*...the marginal probability of being event free at time t.*" Survival is sometimes written as Pr[T > t] which is can be read as the probability of the (total) survival time 'T' being bigger or equal to our time point 't' (Graf et al., 1999).

Looking deeper into common challenges of TTE-studies, censoring is one of the biggest problems for traditional methods. There are various types of censored data. The two most granular distinctions are right censored data and left censored data. Right censored data is data where the event did not take place yet and the real event takes place in the future. Left censored data is data where the event took place at a point between our specified time thresholds, not allowing us to record the exact time accordingly and thus the real event is in the past (Steyerberg et al., 2010; Fu & Simonoff, 2016). The methods in this paper focus on working with right censored data.

When working with TTE studies, one also 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 given the imbalance of misclassification costs (Cook, 2007). Prognostics on the other hand deals with predictive modeling were also accuracy becomes an eminent consideration (Steyerberg et al., 2010). In the machine learning framework, we are predominately interested in prognostic studies.

2.2 Concepts in Model Evaluation

Further, we can disentangle the different components of model evaluation into various groups. Traditionally, 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, we are testing how strong our model discriminates between subjects that incur an event versus subjects that don't at a given time point. Perfect discrimination would imply that we classify all subjects correctly. When solely using a discrimination centered evaluation tool, our model calibration could be severely impaired, but as long as we discriminate correctly we would still measure a strong performance. The most prominent summary scores to evaluate discrimination for classification tasks are the area under the curve (AUC) and generalized version of the AUC, the concordance-statistics. (Steyerberg et al., 2010; Cook, 2007).

Calibration: Calibration deals with the accuracy of our predictions. The most prominent score to capture accuracy in classification tasks is the brier score, a modification of the mean squared error (Assel et al., 2017; Gerds et al., 2008).

Clinical usefulness: A third more recent consideration is the issue of clinical usefulness. Clinical usefulness is 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 (Kvamme & Borgen, 2019; Antolini et al., 2005). For a more detailed introduction into IPCW see Kopper and Scheipl (2020). Additionally, the c-index also changes some common assumptions used in the AUC (Antolini et al., 2005). To fully understand these methods, the focus will be on what differentiates these methods from classical tools.

3 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, building the foundation for the c-index. The ROC allows one to account for imbalanced label distribution and imbalanced misclassification costs. Boiling it down, we want to look at the model performance over various default thresholds rather than at a specific misclassification specification (Cook, 2007). Further, the ROC 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, 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 ROC looks at specificity and sensitivity at various misclassification thresholds. Subsequently, the area under the curve integrates over all possible mis-classification rates. It ranges from 0.5 (no discrimination) to the maximum value of 1 (perfect discrimination) (Uno et al., 2011). The concordance- index is the generalization of the ROC for survival data (Cook, 2007). Concordance describes consistency while discordance can be understood as inconsistency. Essentially, the AUC can be written down as done by Blanche et al. (2019):

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

In the AUC, we need uncensored data, because we need information on both subjects. (Blanche et al., 2019) Due to the fact that we deal need to be able to deal with censored data, we need to modify the AUC.

The AUC deals with different questions than the C-index (Khosla et al., 2010). 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 pairwise and therefore evaluates whether the model correctly classifies 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 consideration, 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. Paul Blanche et al. (2019) specify the C statistics as follows:

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

With Harrell's C, we only need one of two subjects to have an event taking place for the subject pair to be comparable. Harrell's C is one of the most popular concordances statistics. The c-statistic describes how well models rank cases and non-cases, using a rank correlation measure, inspired by Kendall's tau (Uno et al., 2011). What counts as a "good" score strongly depends on the setting at hand. In a clinical setting rules of thumb such as a score of ~ 0.8 is good will not always hold, given that we are frequently interested in a very high true positive rate (Cook, 2007). A concordant pair is a pair of subjects with risk and event data that is consistent, henceforth subjects with higher risks have earlier events and subjects with lower risk scores translate into later event times (Antolini et al., 2005).

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

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 when working with Harell's C. Irrespective, we cannot always use all our data because we don't necessarily have all our data in comparable pairs, henceforth we still loose some data (Gerds et al., 2013). Packages such as 'pec' use a time dependent c-index as for instance proposed by Antolini et al. (2005). Antolini et al. (2005) 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). Rather than omitting the unusable pairs as done with Harell's C, the time dependent c-index uses a weighting function, of which the IPCW is the most popular, to estimate the censored data (Wolbers et al., 2014). There are various functions that can be used to estimate the IPCW. In this paper we only look at the default method using Kaplan Meier - estimates for the censored data. Regardless of the model used, 3 assumptions for the IPCW need to hold for our estimates to be consistent:

- have conditional independent censoring
- have a correct specification of our censoring model
- our data to be right censored data

For further illustration, see Gerds et al. (2013). Another notable mention was proposed by Uno et al.(2011) which suggested 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.

4 Brier Score

The MSE is the incurred quadratic loss (Sonabend et al., 2011). Further, it is a accuracy measure used in the regression setting. Mathematically, we can define the MSE as follows:

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

Essentially, we are comparing actual $y^{(i)}$ and predicted scores $\hat{y}^{(i)}$ and square their difference, giving each observation equal weight. Now, we need to make some changes when working with classifications. The Brier Score is the MSE for Classification. For the Brier score, we are using a probability estimates $\hat{\pi}(x^i)$ rather than estimates of y.

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

Given that we are trying to get some insights into the calibration and accuracy, we need to use probabilities and cannot use raw classification labels. We use estimated predictions of the event status at the time points and values can range from 0 to 1 (Graf et al., 1999). The Brier scores is dependent on the evaluation time. Other terminology that you might encounter is the predicted error as seen in the 'pec' package. The mlr3proba package refers to the brier score as the 'surv.graf', based on Graf who initially modified the measure.

One makes the assumption that the censored data is missing data at random, or re-phased our survival times are independent. Furthermore, we introduce a weighting scheme to estimate our censored data. There are various different methods to do this, one of the most prominent being the IPCW method (Gerds and Schumacher, 2006). Gerds and Schumacher (2006) compare different estimators for the IPCW and recommend the usage of a Cox-PH model or a nonparametric Allen-model given that these two estimators illustrated the most promising results. We are not just interested in the maximum time point, but also performance at different thresholds. It is common practice to use the 25th, 50th and 75th percentile quantile (Bender and Scheipl, 2018) For further specification see Gerds and Schumacher (2006). In the 'mlr3proba' package the loss function for the individual is defined using the definition given by Graf et al. (1999).

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

L is the loss function, S is the survival function, G is IPCW estimate of the censored survival function $P(C* > t)$, S_i is the predicted survival function, t is the time of the event, t^* the time before event and δ is our indicator whether our data is censored or not. For the population score, the following notation is specified:

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

N is the number of observations, and T are all unique time points. Note that the method how the scores are integrated differs per package. E.g. in mlr3proba you can integrate via equal weights (method==1) for each unique time point, or use difference between time points (method==2). With the first method we integrate over all T unique time points, and with the second method we integrate over all N observations. The later is the more prominent method and used e.g. in the 'pec' package. (Sonabend et al., 2020)

5 Implementation

For this illustration, we are using simulation data provided by the 'SimSurv' function. We have specified 10000 observations and set a seed. 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") # Additional complementary functions for the survival package
library("survival") # Entails functions for general survival analysis setting
library("pec") # package for our prediction error curve plots,ibs and c-index scores
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.

5.1 Implementation: Integrated Brier Score

Firstly, we look at the IBS. To derive the IBS, we can first derive the Brier Score and specific the method with which we will estimate the censored data. Here, we are using the default method,the Kaplan-Meier estimates, for the censored data (Gerds, 2020).

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

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 (Gerds, 2020).

```
calPlot(models)
```

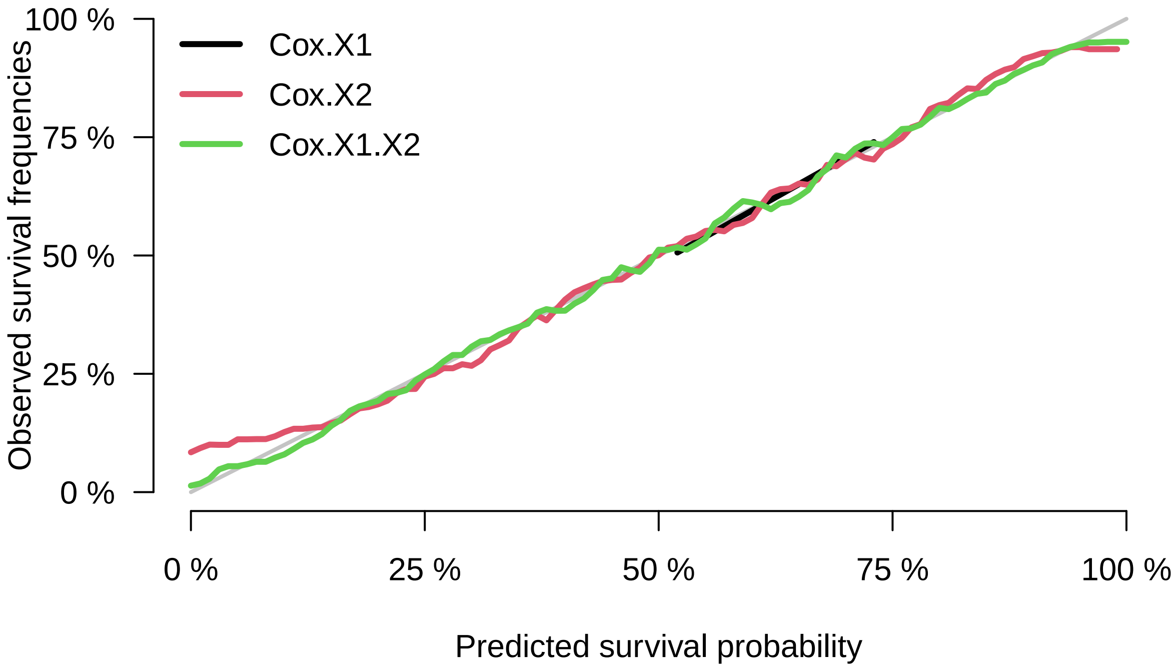

Figure 1: Calibration Plot, plotting predicted probabilities against frequencies of observed scores

Looking at the summary statistics for our prediction error curve at different thresholds, one 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 of our training data at different thresholds this depiction is not synonymous with the IBS scores at the thresholds (Gerds, 2020).

```
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 one can also look at overall performance over time visually. With respect to the interpretation, a lower prediction error is better. The model with two covariates (the blue line) has the lowest prediction error at the different thresholds and hence is the best performing model in this benchmark (Gerds, 2020).

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

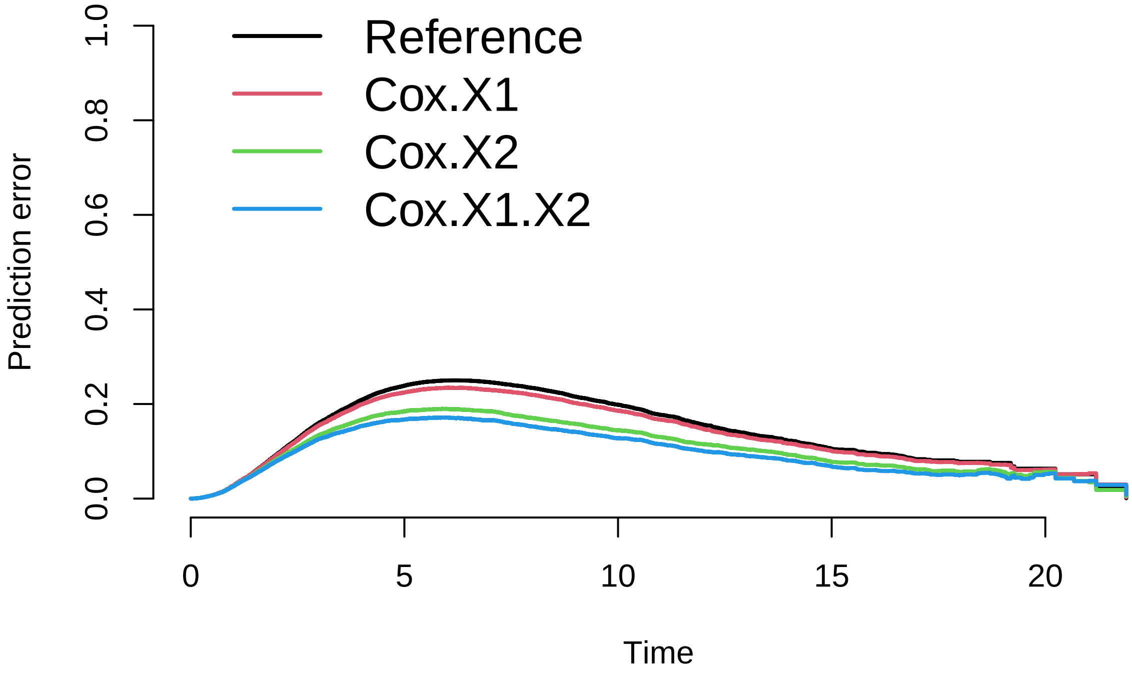

Figure 2: Prediction Error Curve over time

To 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. Again, we can display the scores at various thresholds of our training data and the same interpretation as for the brier score holds (Gerds, 2020).

```
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
```

5.2 Implementation: Concordance Statistics

We are again using the simulated data. To compute the c-index, we are using the ‘cindex’ function from the ‘pec’ package. Note, that a specification for the censored data is needed. Again, we are using the default settings, thus we are using the Kaplan-Meier-estimates for our censored data (Gerds, 2020).

```r
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)))
```

```
##
## 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(x, print = TRUE, ...): The C-index is not proper for t-year predictions. Blanche et al
##
## Consider using time-dependent AUC instead: riskRegression::Score
```

The 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 of 1 would describe a perfect model and a score of 0.5 would imply randomness (Gerds, 2020). At all time points, the model with both covariates performs best with a score ranging from 77.3 to 74.4 over the different time points.

5.3 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 evaluation tools. To specify the measure, we need to 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. Here, we focus on these two as those are most prominent. 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 (Sonabend et al., 2020).

```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)))
```

```r
autoplot(bmr, measure = measure)
```

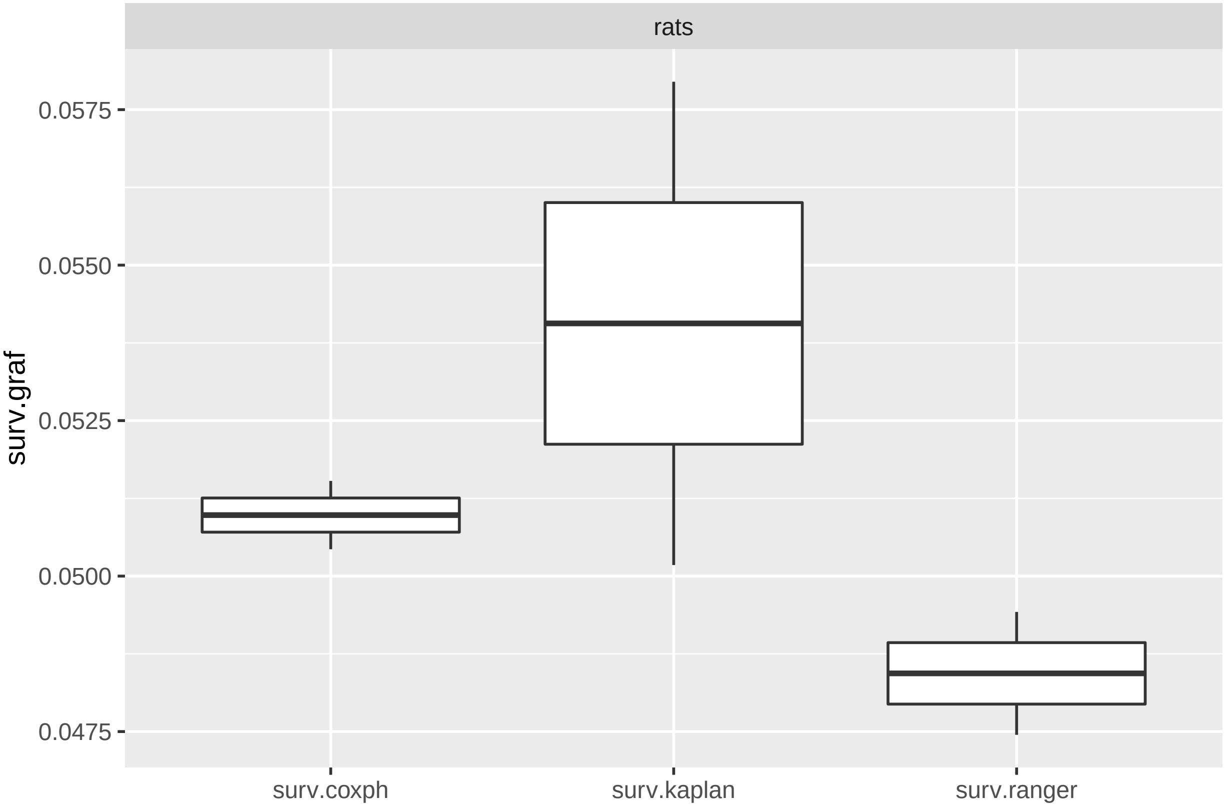

Figure 3: Evaluating Model Performance using the IBS

As we can see, the random survival forest performs best. The Kaplan Meier model has the widest spread and the worst performance. The Cox-PH model has the slimmest spread, but performs slightly worse than the random survival forest model. We are interest in the different thresholds and not only want a single score, hence the boxplot representation is suitable for visual representation.

6 Considerations

After talking about the implementation, there are some important considerations that need to be evaluated.

6.1 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 get insights into classification separately 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. With the c-index we are unable to plot various misclassification rates without knowing the true misclassification rate. Moreover, we are averaging to one single score with less information on model performance. We basically average over all misclassification rates (Hand, 2009). (2) Prognostic studies usually should account for 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.

6.2 Considerations: IBS

Advantages: (1) The integrated brier score is a measure accounting for both discrimination and calibration. Essentially we are comparing two different things, once a tool to specifically look at discrimination (c-index) and secondly an overall performance measure. Merely by keeping this distinction in mind, it should become evident that generally speaking the IBS is more suitable for model comparison. (2) Further, the IBS has the ability to differentiate between useless and harmful models. (3) Moreover, the IBS, deals with estimates specified for a time-horizon as opposed to the c-index where we cannot make t-year predictions (Kattan and Gerds, 2018).

Disadvantages:(1) 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) Further, we also need a benchmark model for evaluation. Henceforth, the interpretation of the absolute scores is problematic and less useful for e.g. clinicians. (3) 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 (Cook, 2007).

6.3 Considerations: Novel Research

Cook (2007) 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 (Cook, 2010).

Decision analysis curve enables the use of weights, allowing optimal decision making based on subjective preferences. Harm is transformed, using an exchange rate to put harm and benefit on one scale (Vickers and Elkin, 2006). This exchange rate can be obtained by asking clinicians questions based on their subjective preferences. 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. One important consideration is that decision analysis curve is a complement, not a substitute to existing models (Vickers and Elkin, 2006).

7 Conclusion

Time-to-Event studies require adjusted model evaluation tools for censored survival data. One needs to separate between model evaluation metrics 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. One should consider that the c-index is very interpretable, but simultaneously loses a lot of information compared to the AUC/ROC and ignores clinical usefulness. Blanche et al. (2019) emphasis that the c-index is not suitable for the same type of questions as the AUC. Further, the IBS provides more information on overall model performance than the c-index but is not as interpretable and neglects clinical usefulness. Irrespective, when evaluating machine learning models, it is recommended to consider both discrimination and calibration. Both of these tools are useful starting points when dealing with TTE studies.

8 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.

Bender, Andreas, and Fabian Scheipl. 2018. "Pammtools: Piece-Wise Exponential Additive Mixed Modeling Tools." *arXiv Preprint arXiv:1806.01042*.

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. 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. 2020. "Package 'Pec'."

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.

Gerds, Thomas A., Michael W. Kattan, Martin Schumacher, and Changhong Yu. 2013. "Estimating a Time-Dependent Concordance Index for Survival Prediction Models with Covariate Dependent Censoring." *Statistics in Medicine* 32 (13): 2173–84. https://doi.org/https://doi.org/10.1002/sim.5681.

Graf, Erika, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. "Assessment and Comparison of Prognostic Classification Schemes for Survival Data." *Statistics in Medicine* 18 (17-18): 2529–45.

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

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.

Kopper, Philipp, and Fabian Scheipl. 2020. "Flexible Estimation of Complex Effects in the Context of Competing Risks Survival Analysis: Exposure-Lag-Response Association of Artificial Nutrition and Patients' Length of Stay in Intensive Care Units." 132.

Kvamme, Haavard, Ørnulf Borgan, and Ida Scheel. 2019. "Time-to-Event Prediction with Neural Networks and Cox Regression." *Journal of Machine Learning Research* 20 (129): 1–30.

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.

Sonabend, Raphael, Franz J. Király, Andreas Bender, Bernd Bischl, and Michel Lang. 2020. "Mlr3proba: Machine Learning Survival Analysis in R." *arXiv Preprint arXiv:2008.08080.*

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.

Wei Fu, and Jeffrey S. Simonoff. 2016. "Survival Trees for Left-Truncated and Right-Censored Data, with Application to Time-Varying Covariate Data." *Biostatistics*, December, kxw047. https://doi.org/10.1093/biostatistics/kxw047.

Wolbers, Marcel, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. 2014. "Concordance for Prognostic Models with Competing Risks." *Biostatistics (Oxford, England)* 15 (3): 526–39. https://doi.org/10.1093/biostatistics/kxt059.