Model Evaluation

Considerations for Time-to-Event Studies

Daniel Saggau

11/12/2020

Overview

1) Time to Event Studies
   - What differentiates a TTE Study from other studies?
2) Classical Model Evaluation: Brier Score and AUC
   - What are classical model evaluation tools & why can't we use them?
3) TTS Model Evaluation: IBS and c-index
   - How do these methods address the limitations of classical methods?
4) Discussion
   - What measure is most useful for machine learning in TTS?
5) Further Considerations
   - What other methods are coming?

Time-to Event Studies

- Analysis working with (right) censored data
- Right censored data (event after follow up) vs. left censored data (event was not recorded when it occurred initially)
- Highly relevant for clinicians in the field of medical statistics
  e.g. looking at when a patient dies or when he gets a disease (clinical/epidemiological studies)
- In Economics/Finance e.g. to examine when a subject/borrower will default or when a subject will find/lose a job
- Operations research to predict the time a machine will break

Basic Notations & Concepts

- Time T and Survival S
- From hazard to cumulative hazard to survival
- Hazard h(t,x) is the eminent probability of death a specific point in time
- Capital H is the cumulative hazard
- non-parametric hazard models (KM) vs.semi-parametric proportional hazard model

Classical Model Evaluation Tools for Classification Tasks

1. **Diagnostic vs. Prognostic Study**
2. **What elements do we consider?**
   - Discrimination: Are we able to correctly discriminate between e.g. sick and healthy patients ?
   - Calibration: How concise is our prediction accuracy ?
   - Clinical Usefulness: Will our model create more benefits than harm?
3. Working with *Label* vs. working with *Probability*
   - Brier Score (probability from true class label)
   - AUC/ROC (receiver operating characteristics)

Brier Score

Based on loss function. Other loss measures are the log loss or the integrated log loss.

MSE for Regression (L2 Loss):

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

Where: the $MSE \in [0; \infty)$

The Brier Score is the MSE for Classification:

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

The general version of the brier score looks at a specific point in time

Confusion Matrix

Sensitivity or: true positive rate

- deals with values above the threshold among the subject group which do endure an event

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

Specificity or: true negative rate

- deals with false negatives, hence patients with a disease we classify as not having any diseases

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

Why cant we use traditional model evaluation tools for time to event studies?

- Working with censored data
- Account for time dependent covariates

Early approaches: - excluding subjects with right censored data and only evaluate on the complete data

From AUC to Harell's C-index to time dependent C-index

- Advancement from AUC
- Rank correlation measure but still have to deal with censoring
- studying concordance (~consistency) and discordance (~inconsistency) pairs

Intuitively speaking the difference between AUC and c-index is as follows:

$AUC = $ $C = $ While C is defined as:

$$\frac{\#ConcordantPairs}{\#ConcordantPairs+\#DiscordantPairs}$$

In this approach, only comparable pairs are evaluated

$$C^{td} = \frac{\pi_{concordance}}{\pi_{comparable}}$$

Henceforth:

$$C^{td} = \frac{Pr(z(X_i)>z(X_j)\&T_i<T_j\&E_i=1}{Pr(Ti<Tj|E_i=1)}$$

c-index

How to deal with censoring: * addressing right censored data via inverse of the probability of censoring weighted estimate (of concordance probability) * Kendall rank correlation coefficient test as inspiration * Summary measure (over all time) based on the AUC

$$C - index = \frac{\Delta_j \times \sum_{i,j} 1_{Ti>Tj} \times 1_{\eta_i > \eta_j}}{\Delta_j \times \sum_{i,j} 1_{Ti>Tj}}$$

IBS

- called cumulative predictive error curves == continuous ranked probability score (crps)
- area under the prediction error curve
- Integral over all points in time to get one summary value henceforth called "integrated" BS
- able to build a R^2 like measure where we divide MSE of a model with a different MSE of reference model
- Where L is a loss function of the S(the probability that the event of interest has not taken place yet) and time
- t is the time of the event (death) and t* the time before death
- G(t) is the P(C>t), so where the censored time is longer than the time (in mlr3proba via survfit == KM Estimate)
- When selecting integrated == FALSE then we looking at specific time

Mean population

Mean Population:without Integration

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

Mean Population: with Integration

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

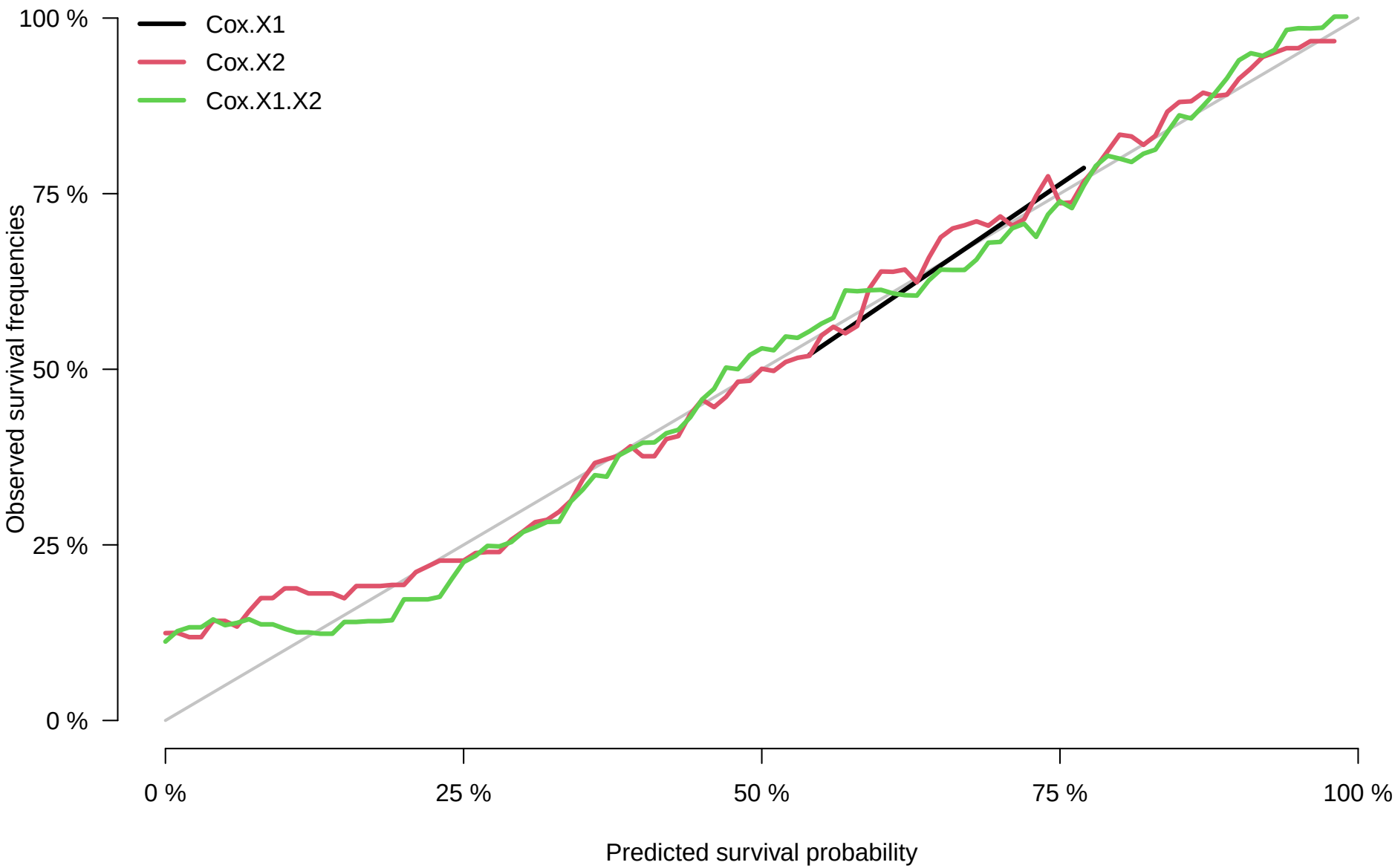
- N = Number of observations
- S_i is the predicted survival function

Coding Settup

```
set.seed(123)
dat=SimSurv(1000)
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))
perror <- pec(object=models,
                   formula=Surv(time,status)~1,
                data=dat, # formula for IPCW
                   exact=TRUE, cens.model="marginal",
                splitMethod="none",
                   B=0,# number boostrap samples
                   verbose=TRUE)
```

Calibration Plot

```
calPlot(models)
```



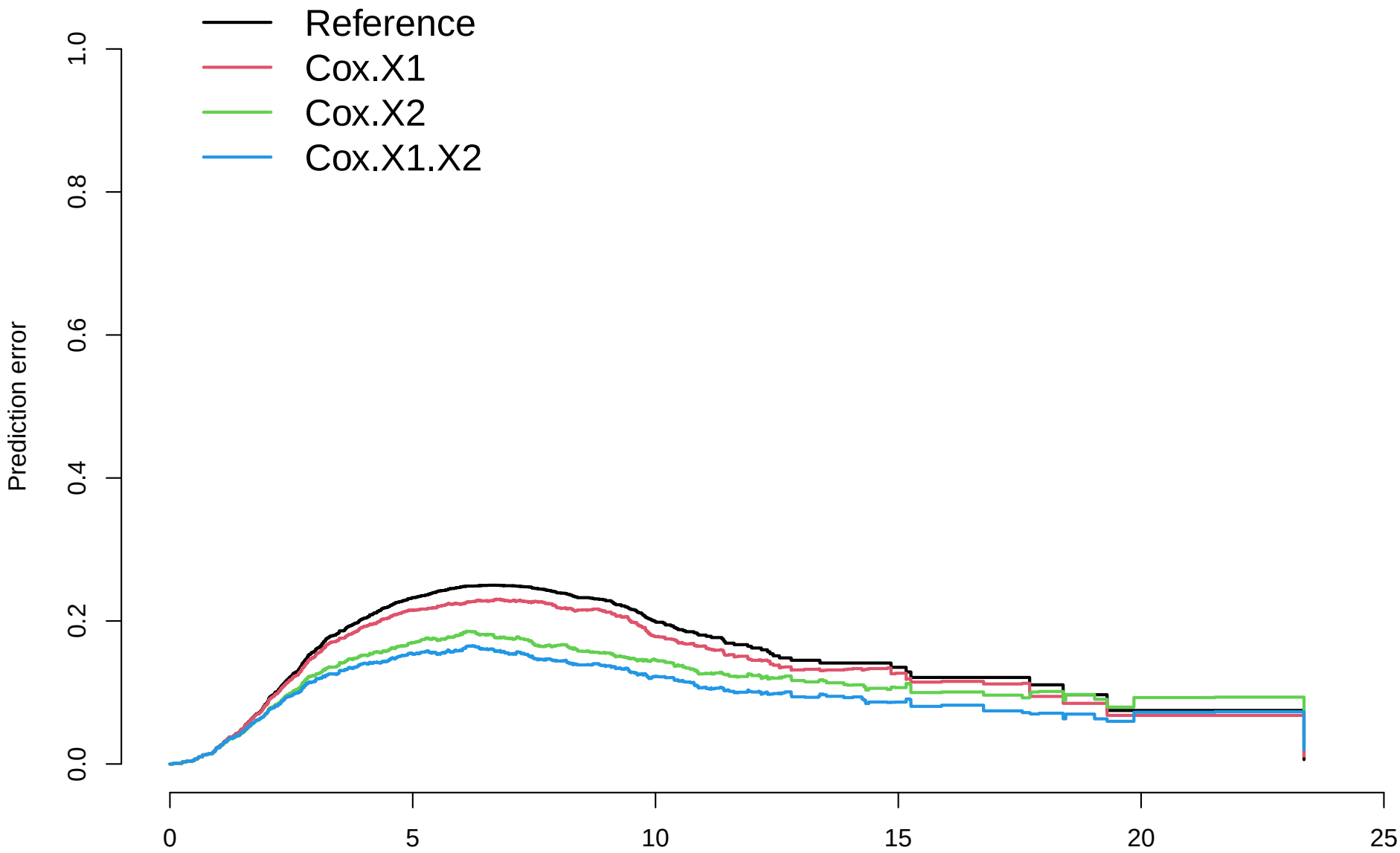
Summary Prediction Error Curves

```
summary(perror,times=seq(0,20,5))
```

```
##
## Prediction error curves
##
##
##
## No data splitting: either apparent or independent test samp
##
##  AppErr
##    time n.risk Reference Cox.X1 Cox.X2 Cox.X1.X2
## 1     0   1000     0.000  0.000  0.000     0.000
## 2     5    471     0.233  0.215  0.170     0.154
## 3    10    105     0.199  0.178  0.145     0.122
## 4    15     17     0.135  0.127  0.107     0.087
## 5    20      2     0.075  0.068  0.093     0.072
```

Plotting prediction error

```
plot(perror,xlim=c(0,25), ylim = c(0,1))
```



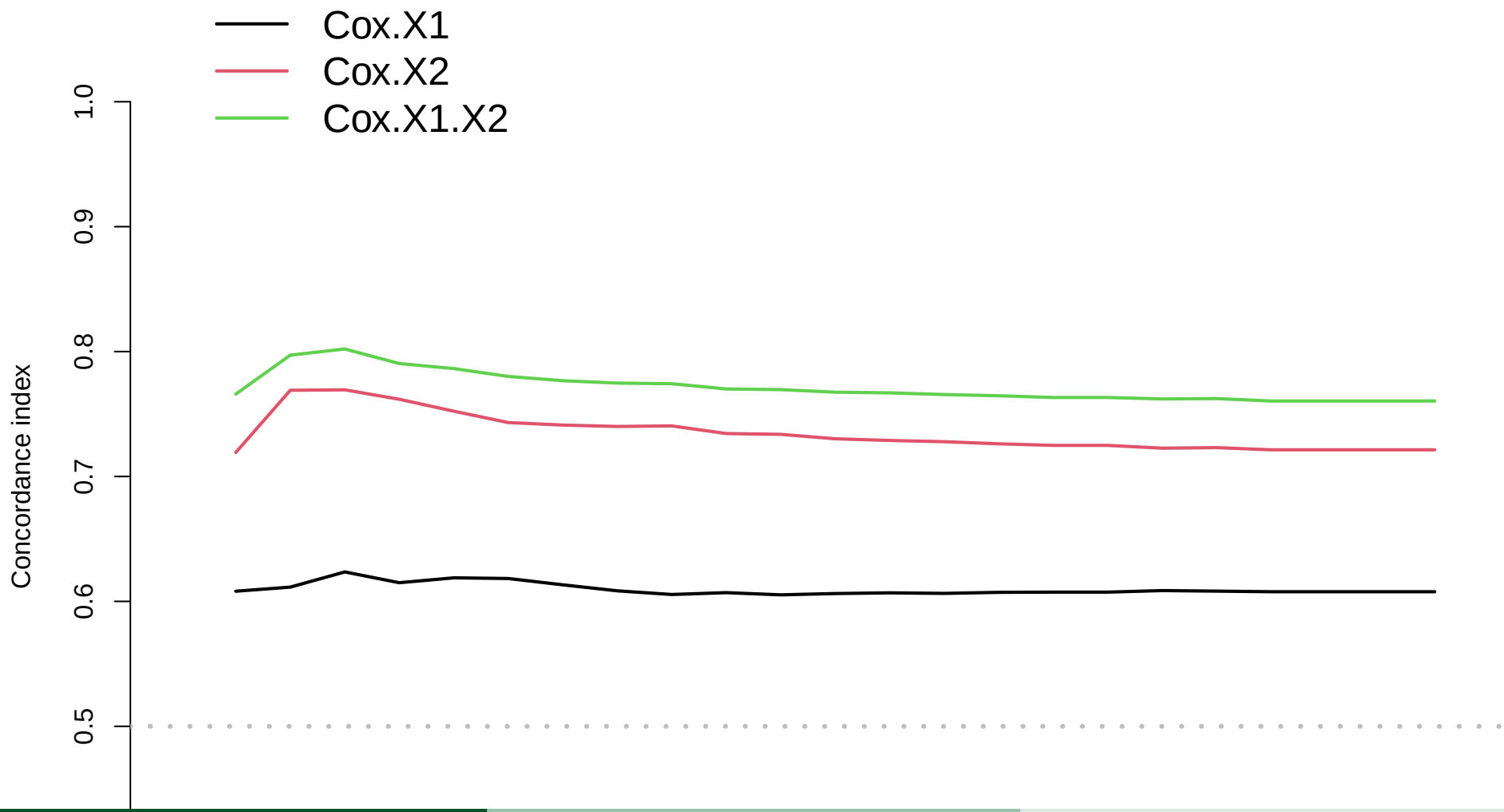
Cumulative Prediction Error

```
#crps(perror,times=seq(0,25,5))
ibs(perror,times=seq(0,23.4,5))
```

```
##
## Integrated Brier score (crps):
##
##              IBS[0;time=0) IBS[0;time=5) IBS[0;time=10) IBS[0;
## Reference                0         0.117          0.177
## Cox.X1                   0         0.111          0.164
## Cox.X2                   0         0.091          0.129
## Cox.X1.X2                0         0.085          0.116
##              IBS[0;time=20)
## Reference             0.156
## Cox.X1                0.144
## Cox.X2                0.119
## Cox.X1.X2             0.101
```

c-index plot

```
plot(cindex(models, formula = Surv(time,status) ~ 1,
                        cens.model="marginal", data = dat,
                        eval.times = seq(1,23.4,1)))
```



mlr3Proba

Methods based on the loss function:

- Integrated Graf Score (other Name for IBS based on Author Graf)
- Integrated Log Loss
- Log Loss

Further measures via survAUC package:

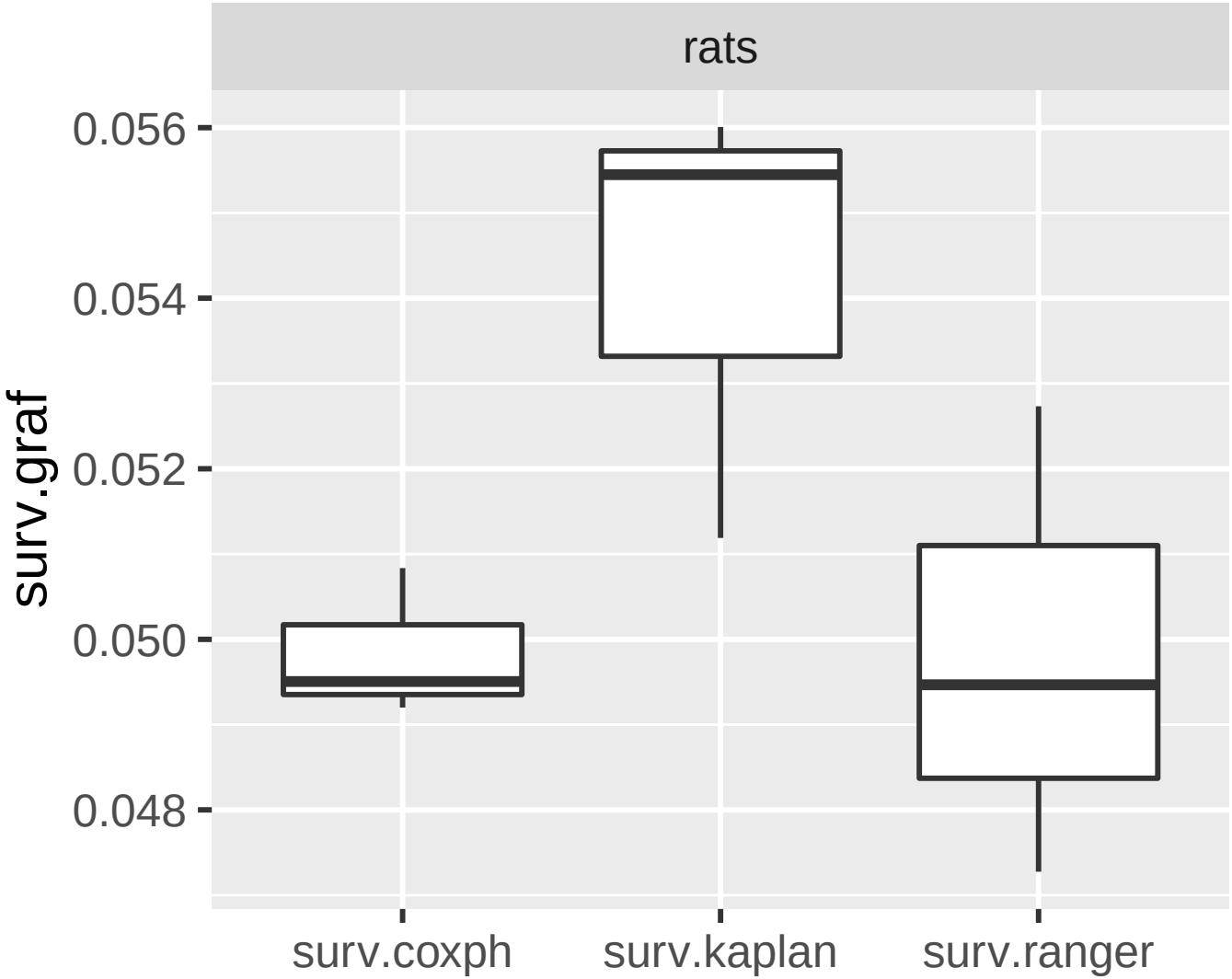
- Uno's AUC/TPR/TNR
- Song and Zhou's AUC/TNR/TPR
- Chambless and Diao's AUC
- Hung and Chiang's AUC

Others:

- van Houwelingen's Alpha Calibration
- van Houwelingen's Beta Calibration

mlr3Proba Example

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



Discussion

- c-index has gained popularity because of it's interpretability
- Integrated Brier Score accounts for both calibration and discrimination
- Irrespective, neither model accounts and leaves room for improvement
- IBS allows for differentiation of 'useless' and 'harmful'
- Estimators can be influenced by data
- Clinical consequences problematic

Novel Research

- Decision Curve Analysis
- Net Reclassification Improvement

Conclusion

- There are various different modifications for model evaluation, neither being unconditionally superior
- The Brier Score and the AUC are pivotal for many of these methods
- While there has been a lot of research on this topic, the debate is on going

Literature and Recommendations

Introduction:

- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1), 128.

Comparative Study:

- Kattan, M. W., & Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagnostic and prognostic research, 2(1), 7.

Use Cases:

https://rpubs.com/kaz_yos/survival-auc https://datascienceplus.com/time-dependent-roc-for-survival-prediction-models-in-r/
https://rdrr.io/cran/pec/ https://adibender.github.io/pammtools/