Machine Learning Applied to Right-Censored Survival Data

Prof. Dr. Antje Jahn
Hochschule Darmstadt

Prof. Dr. Gunter Grieser
Hochschule Darmstadt

Lukas Klein
M. Sc. Data Science

WS22/23

What is right-censored survival data?

Survival data describes time-to-event targets.

Data collection only ends when either:

- the **outcome** of interest is **observed**
- the subject is **lost to follow-up**
- the study has **ended**.

For each subject, the observed time ***T*** and a boolean indicator δ, whether an event was seen, is modeled. The observed time **T** is a censored version of the actual time-to-event ***T****, which can be only sampled for subjects, where a event was observed.

What data was used for the project?

In 2016 with the new transplantation law in Germany (**TxRegG**) a **central registry** for all data on organ transplantations in Germany was established. This registry offers data for quality assurance and research.

For **first time transplants of kidneys from deceased donors,** recipient, transplant and donor data was processed to create a dataset of **20325 subjects with 261 descriptors**.

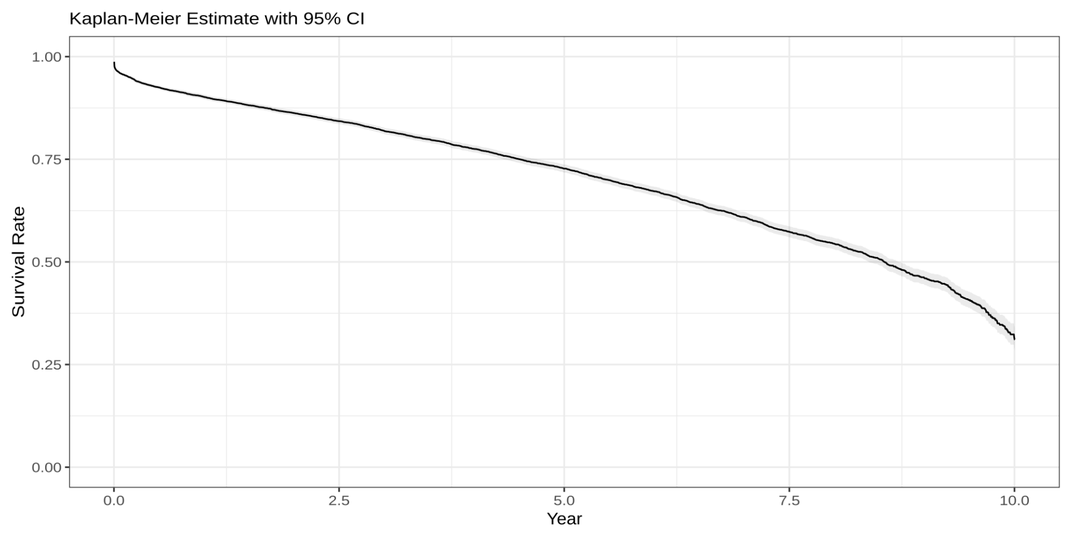
The goal was to model the time to an endpoint after transplantation. Different construction methods of this target were tested. Presented here are results, when the endpoint was either **patient death or graft failure, whichever came earlier, over 10 years**.

Results

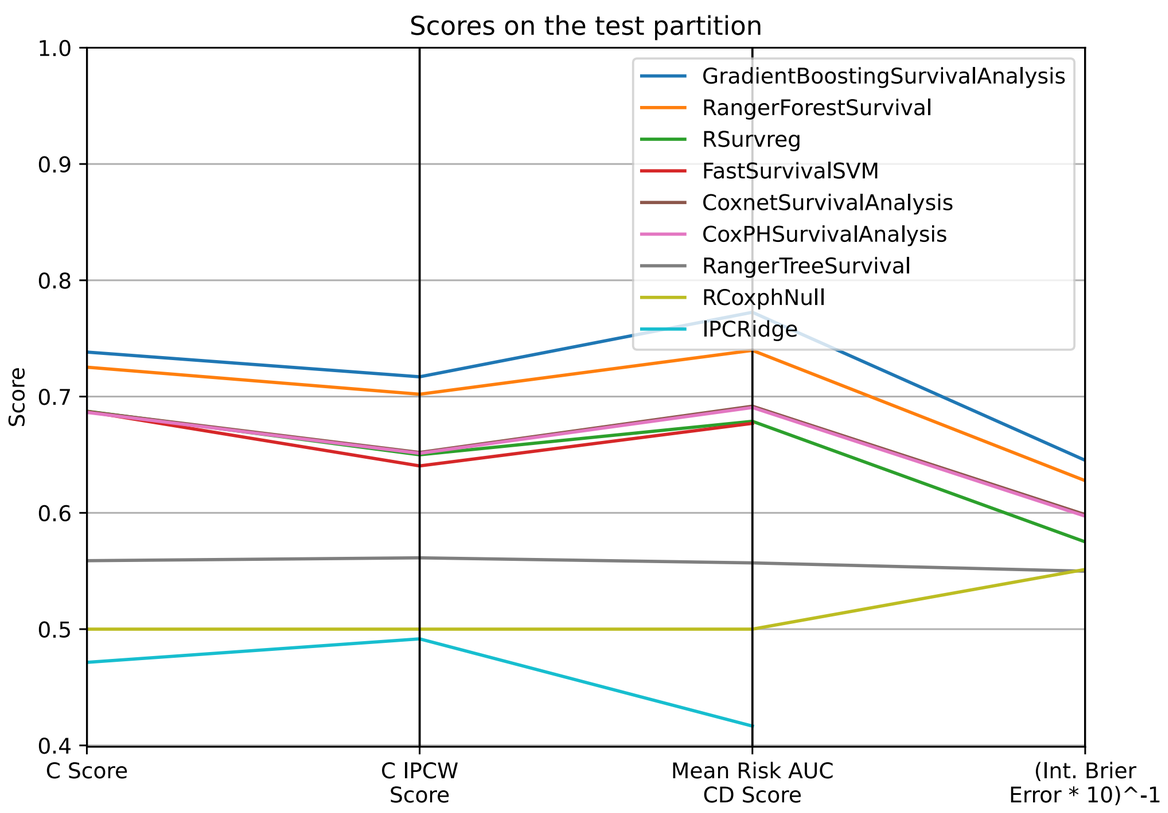
The plot below shows the test scores for the **best configurations** of each used model type. The first three scores are based on the **ranking of the subjects** based on their survival time. Higher is better. For the Integrated Brier Score, the estimated survival function is used, which is not available for the IPCRidge model.

The best models across all targets were the **boosted models** and the **random forests**.

Models from packages for R and Python were used. Performance and abilities were quite different between packages and a future project might **systematically investigate these differences.** While the registry has a scientific export function, different types of **inconsistencies** in the **relational dataset** made it difficult to use.



This plot shows the **Kaplan-Meier** curve of the target described above. A Kaplan-Meier reports the **proability** of a subject **not experiencing an event up to a time t**. It is an estimator of the survival function.



Which model types were tested?

GradientBoostingSurvivalAnalysis: Boosted survival trees and boosted regression trees with inverse-probability-of-censoring-weighting (IPCW)

RangerForestSurvival: Random forest of survival trees

RSurvreg: Linear accelerated failure time (AFT) model, models the logarithm of the time-to-event

FastSurvivalSVM: Support vector machines ranking subjects according to their risk

CoxnetSurvivalAnalysis: ElasticNet with IPCW

CoxPHSurvivalAnalysis: Cox-Regression, specialized linear regression for survival analysis

RangerTreeSurvival: Tree, where each leaf constructs a Kaplan-Meier estimator

RCoxphNull: Kaplan-Meier model ignoring descriptors

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

More Information

Transplantation Registry: *https://transplantations-register.de/ueber-das-transplantationsregister*

Python Library for Survival Analysis: *https://github.com/sebp/scikit-survival*

R Package for Survival Analysis: *https://cran.r-project.org/web/packages/survival/*