Super Learner for survival prediction from censored data: Extension of the R package RISCA

Yohann Foucher and Camille Sabathé

The 9th International Meeting on Statistical Methods in Biopharmacy, Paris 2022

yohann.foucher@univ-poitiers.fr

Plan

The problematic of the method choice for survival prediction.

- The prediction of the probability that a subject experienced an event is often of interest.
- Several regressions can be used for right-censored data. :
  - Most of the studies use proportional hazard (PH)-based assumption.
  - Other models such as accelerated failure time (AFT) approaches are not frequent.
- Machine learning are increasingly being used and avoid such modeling assumptions :
  - Random survival forests.
  - Survival neural networks.
  - Support-vector machines.
  - Etc.

A super learner (SL) allows us to combine regressions and algorithms.

- In 2011, Polley and van der Laan proposed a SL for right-censored data. [1]
- Two R packages are available :
  - The first one was proposed by Golmakani et al. (2020). It allows us to obtain the linear predictor of a PH regression.
  - The second one was developed by Westling et al. (2021) with additional learners : several parametric PH models, a generalized additive Cox regression, and a random survival forest.
- **We aimed to extend these packages to additional learners and loss functions.**

1. Super Learning for Right-Censored Data. In MJ van der Laan, S Rose (eds.), Targeted Learning, Springer Series in Statistics.

Plan

The SL is estimated by minimizing the cross-validated loss function.

- $S_m(.)$ is the survival function obtained by the m^{th} learner ($m = 1, ..., M$).
- w_m is the corresponding weight with respect to $\sum_1^M w_m = 1$ and $0 \leq w_m \leq 1$.
- The sample is randomly divided into V cross-validated sub-samples.
- For each of the folds, one can estimate the M learners from the training subjects and predict $\tilde{S}_m(.)$ of the leaving subjects.
- The weights $\hat{w}_m$ are then obtained by minimizing the loss function, i.e., distance between the observations and the predictions $\tilde{S}_{sl}(t \mid Z) = \sum_{m=1}^M w_m \tilde{S}_m(t \mid Z)$.
- The final SL is obtained by $\hat{S}_{sl}(t \mid Z) = \sum_{m=1}^M \hat{w}_m \hat{S}_m(t \mid Z)$, where $\hat{S}_m(t \mid Z)$ are estimated on the entire sample.

The implemented learners.

- Parametric AFT models. (Weibull, Gamma and generalized Gamma distributions).
- Parametric PH models (Exponential or Gompertz distributions).
- Semiparametric PH models with a non-parametric baseline hazard function estimated by using the Breslow estimator (with an option for covariates selection by forward AIC-based selection).
- Penalized PH models (Lasso, Ridge, or Elastic-Net). The quantitative covariates are transformed with B-splines to relax the log-linear assumption.
- Random survival forests.
- Survival neural networks. The linear predictor of the previous semi-parametric PH model is obtained by a single hidden layer network with non-linear activation functions.

The implemented loss functions.

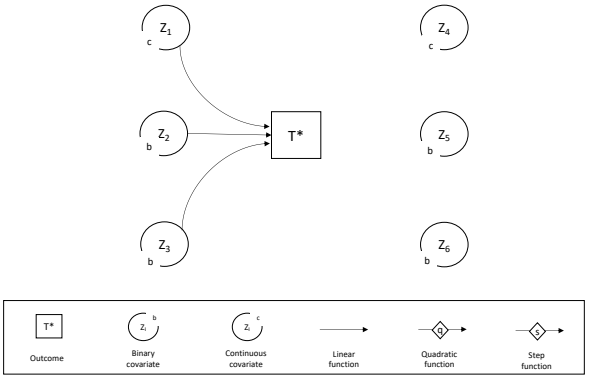
- The Brier Score (BS) for right-censored data and a prediction at time t.
- The negative binomial log-likelihood (BLL) for a prognostic at time t.
- The integrated BS and BLL up to the maximum follow-up time.
- The restricted integrated BS and BLL up to a time t.

Plan

The design.

- We generated 1000 data sets for each scenario.
- The times-to-event were obtained from Weibull distributions and the PH assumption.
- The censoring times were generated from uniform distributions to obtain a 40% censoring rate.
- We studied two sample sizes for learning (200 and 500).
- The validation samples were composed of 500 subjects.
- We proposed two contrasting scenarios.

The design of the simple scenario.



The SL performed as well as semi-parametric approaches.

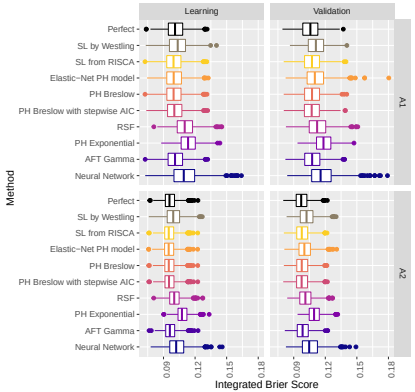
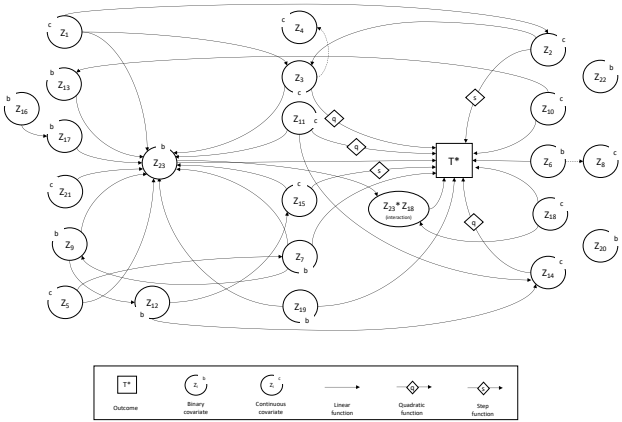


Figure 1: Simulation results in the simple context. A1-$N = 200$ for learning in the two top plots. A2-$N = 500$ for learning in the two bottom plots.

The design of the complex scenario.



The SL performed the best.

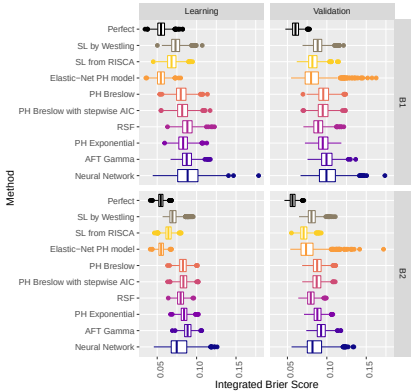


Figure 2: Simulation results in the complex context. A1-$N = 200$ for learning in the two top plots. A2-$N = 500$ for learning in the two bottom plots.

Plan

Conclusions

1. Compared to the available R-based SL in this context of survival analysis, our proposition allows a larger set of candidate learners and loss functions.

   The related functions were included in the R package RISCA :

   https://cran.r-project.org/web/packages/RISCA/index.html

2. The simulation study showed that our proposed functions performed well.