ISMB/ECCB 2023 CONFERENCE
23 | 27 JULY

C-133: Improving Survival Risk Prediction with Random Survival Forests for Recurrent Events in Biological Systems

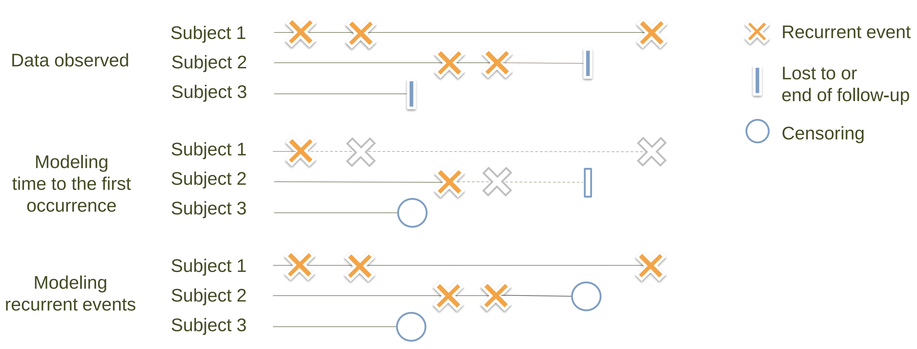
Juliette Murris[1,2,3], Audrey Lavenu[4,5,6], Sandrine Katsahian[1,2,7]

[1] Inserm, Centre de recherche des Cordeliers, Université de Paris, Sorbonne Université, Paris, France; [2] HeKA, Inria, Paris, France ; [3] R&D, Pierre Fabre, Boulogne-Billancourt, France ; [4] Université de Rennes 1, Faculté de médecine, Rennes, France ; [5] IRMAR, Institut de Recherche Mathématique de Rennes, Rennes, France ; [6] CIC Inserm CIC 1414, Université de Rennes 1, Rennes, France ; [7] CIC-1418, Hôpital Européen Georges Pompidou, Service d'informatique médicale, biostatistiques et santé publique, AP-HP, Paris, France

CONTEXT & OBJECTIVES

- Modern technologies enable data to be generated on **thousands of variables** or observations, as per genomics, medico-administrative databases, disease monitoring by intelligent medical devices
- Study individuals may face **repeated events over time**, such as hospitalizations or cancer relapses (Figure 1)
- In either clinical trials or real-world set, **survival analysis** usually focuses on modeling the time to the first occurrence of the event

Figure 1. Recurrent Event Framework



Study objectives

- To present an extension of the **random forest algorithm** for the analysis of **survival data with recurrent events**, utilizing concepts from **non-parametric** survival analysis and **statistical learning**

METHODS

Non-parametric basics

Let $N_i(t)$ the cumulative number of events for the individual $i = 1, \ldots, n$ over the interval $[0, t], t \in [0, T]$ with T the longest follow-up time overall

- The mean cumulative function (MCF) writes

$$\mu(t) = E[N_i(t)]$$

- The Nelson-Aalen MCF estimator writes

$$\hat{\mu}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\delta(u)}$$

with $\delta(t) = \sum_{i=1}^{n} \delta_i(t)$ and $\delta_i(t)$ indicates whether the individual is at risk.

Pseudo-score test from Cook, Lawless & Nadeau can be used to compare two MCFs. H_0 is no difference across MCFs. For two sub-samples A and B, the test statistic writes

$$U(t) = \int_0^t \frac{\delta_A(u)\delta_B(u)}{\delta_A(u) + \delta_B(u)} (\mathrm{d}\hat{\mu}_A(u) - \mathrm{d}\hat{\mu}_B(u))$$

Growing survival decision tree extended to recurrent events

The splitting rule

- At each node, $m \in \mathbb{N}$ predictors are randomly selected
- A greedy algorithm for optimal threshold research to **maximize** the pseudo-score test statistic

Estimates for terminal nodes

- The MCF estimator for an individual i with x_i the vector of predictors writes

$$\hat{\mu}(t|x_i) = \hat{\mu}_h(t) \times \mathbb{I}_{x_i \in h}$$

- $\hat{\mu}_h$ the MCF estimator constructed at the terminal node h

Pruning

- Trees grow up until each terminal node contains at least $\xi \in$ N individuals

Aggregating

The ensemble estimators for an individual i is the average of the estimate over all π_{tree} trees and is defined as

$$\hat{H}(t|x_i) = \frac{1}{\pi_{tree}} \sum_{\pi=1}^{\pi_{tree}} \hat{\mu}_\pi(t|x_i)$$

BIBLIOGRAPHY

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.
Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.
Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.
Murris, J., Charles-Nelson, A., Lavenu, A., & Katsahian, S. (2022). Towards Filling the Gaps around Recurrent Events in High-Dimensional Framework: Literature Review and Early Comparison. arXiv preprint arXiv:2203.15694.
Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. Automated machine learning: Methods, systems, challenges, 3-33.
Nelson, W. B. (2003). Recurrent events data analysis for product repairs, disease recurrences, and other applications. Society for Industrial and Applied Mathematics.
Cook, R. J., Lawless, J. F., & Nadeau, C. (1996). Robust tests for treatment comparisons based on recurrent event responses. Biometrics, 557-571.
Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine, 15(4), 361-387.
Kim, S., Schaubel, D. E., & McCullough, K. P. (2018). AC-index for recurrent event data: Application to hospitalizations among dialysis patients. Biometrics, 74(2), 734-743.
Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

RESULTS

Simulation scheme

- Homogenous Poisson process used with the times between two successive events following exponential distribution with following intensity function

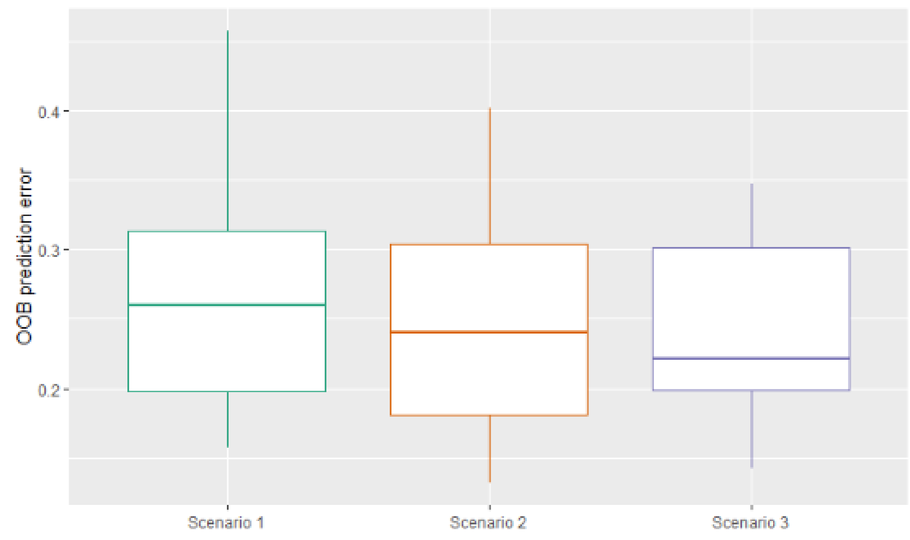
$$\lambda(t|z_i) = r_0 * r(z_i, \beta)$$

- Several scenarios explored with $n = 500$ stochastic processes, $p = 10$ binary predictors
  1. $\beta_1 = 0.5, \beta_{2:10} = 0$
  2. $\beta_1 = 0.8, \beta_2 = 0.5, \beta_{3:10} = 0$
  3. $\beta_1 = 0.8, \beta_2 = 0.5, \beta_3 = 0.5, \beta_{4:10} = 0$

Evaluation based on OOB prediction error

- Use of out-of-bag (OOB) ensemble estimator to define a predicted outcome derived from OOB data
- Extended C-index from Harrell to account for recurrence and overall follow-up
- OOB prediction error was estimated from 30 independent bootstrap replicates and in each instance 100 trees were grown (Figure 2)

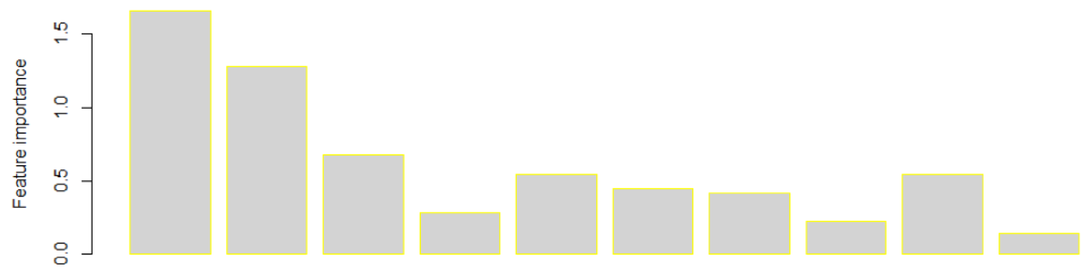
Figure 2. Performance based on OOB prediction error



Feature importance

- Assessed based on permutations and whenever prediction error < 0.50
- The feature importance for a predictor is the prediction error for the original ensemble substracted from the prediction error for the new ensemble obtained after permutation
- Large importance values indicate variables with predictive ability, whereas zero or negative values identify nonpredictive variables to be filtered (Figure 3)

Figure 3. Feature importance for scenario 3 with best performance



DISCUSSION & CONCLUSION

- Our approach is **simple** and easily **accessible**
- And constitutes a **solid baseline for many extensions**

For this reason, the approach we propose is a **valuable contribution** for analysing recurrent events in medical research.

Perspectives

- More scenarios could be explored and include variations of number of subjects and multicollinearity in predictors
- Other evaluation metrics could be used e.g., mean square error, mean absolute error, log-likelihood, feature importance

The proposed methodology has the potential to **facilitate the analysis of recurrent events in biological systems**, providing **key insights** into the underlying mechanisms of **survival outcomes**.