Decision trees for analyzing survival data with recurrent events

Juliette Murris[1] Audrey Lavenu[2] Sandrine Katsahian[3]

July 2023
Journées de Statistique
ULB, Brussel, Belgium

[1]HeKA, Inria Paris - Inserm, Université Paris Cité, Pierre Fabre,
[2]CIC-1414 Inserm, IRMAR - CNRS 6625, Université de Rennes 1,
[3]CIC-1418 HEGP, AP-HP, HeKA, Inria Paris - Inserm, Université Paris Cité

Today's talk

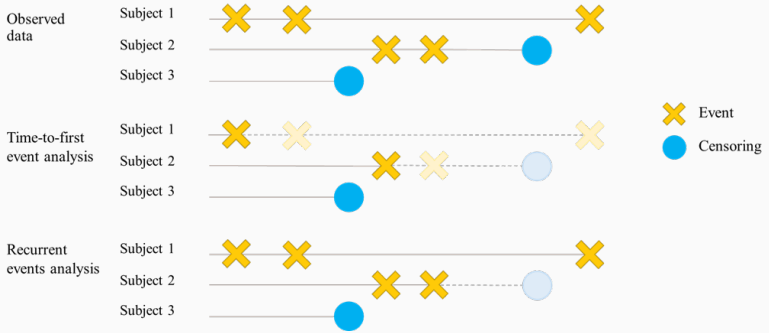
1. Introducing survival data and recurrent events
2. Developing adequate decision trees
3. A simulation study

Introducing survival data and recurrent events

What survival data are made of

In medical research, **survival endpoints** are composite:

- Binary information – did the event occur?
- Continuous time – when did it occur?
- E.g., overall survival, progression-free survival



The advent of machine learning

- Usual machine learning algorithms have been extended to account for survival data
- But **not** to account for survival data **and** recurrent events.

The objective for today is to introduce a new approach to
model recurrent events using learning techniques.

Developing adequate decision trees

Further words on recurrent events

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

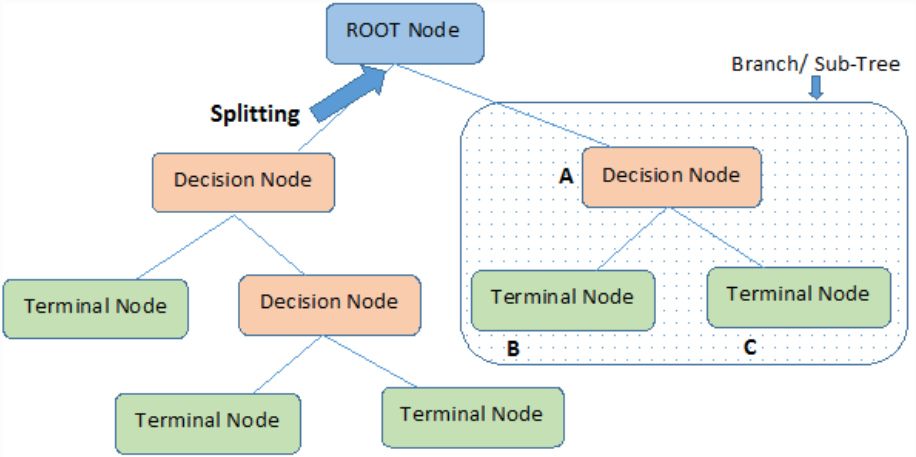
- The mean cumulative function (MCF) writes $\mu(t) = \mathbb{E}[N_i(t)]$,
- The Nelson-Aalen MCF estimator writes $\hat{\mu}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\delta(u)}$

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

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

$$U(t) = \int_0^t \frac{\delta_A(u)\delta_B(u)}{\delta_A(u) + \delta_B(u)} (d\hat{\mu}_A(u) - d\hat{\mu}_B(u)). \tag{1}$$

Growing decision trees



Note:- A is parent node of B and C.

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 individual i with x_i vector of predictors writes

$$\hat{\mu}(t|\mathbf{x}_i) = \hat{\mu}_h(t) \times \mathbb{1}_{\mathbf{x}_i \in h}, \tag{2}$$

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

Pruning

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

A simulation study

A few words on the simulation scheme

- Given the covariates z_i, the **intensity function** of time t is as follows

$$\lambda(t|z_i) = r_0(t) \times r(z_i, \beta) \tag{3}$$

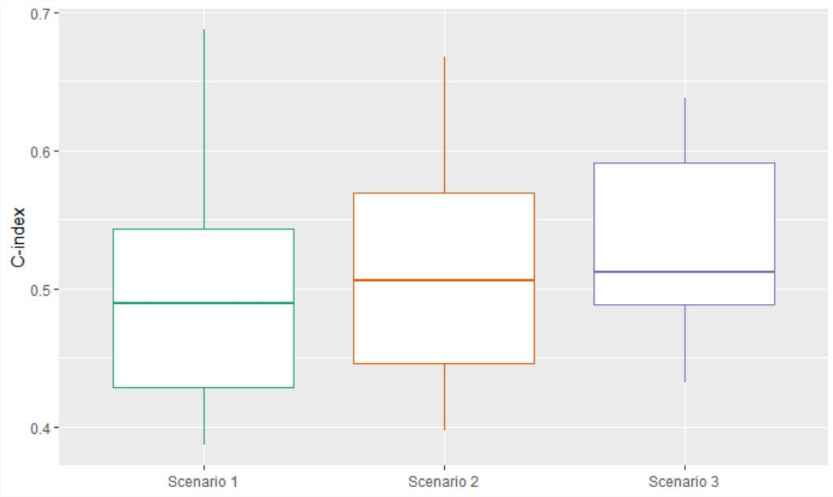
  with $r_0(t)$ the baseline hazard rate function of time t, $r(z_i, \beta)$ the relative risk function, and β the covariate coefficients

- **Homogeneous Poisson Process** (i.e., constant hazard rate over time) with the times between two successive events following exponential distribution

Today, we will go through **3 scenarii** with $n = 100$ stochastic processes and $p = 10$ predictors:

1. $\{\beta_1 = 0.5, \beta_{2:10} = 0\}$
2. $\{\beta_1 = 0.8, \beta_2 = 0.5, \beta_{3:10} = 0\}$
3. $\{\beta_1 = 0.8, \beta_2 = 0.5, \beta_3 = 0.5, \beta_{4:10} = 0\}$

Some good performance observed!



k-fold cross-validation was performed to determine best $m = \{\sqrt{p}, p\}$ number of predictors selected at each node and ξ individuals at terminal node.

We are interested in predictors too

To what extent do "good" models use the right predictors?

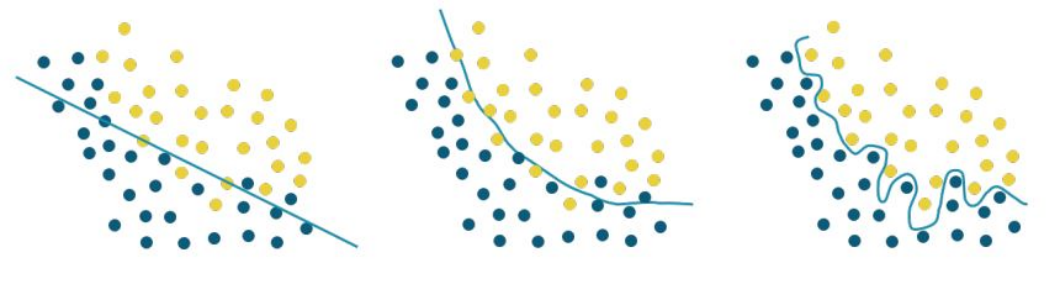
Mean $\pm$ sd	β_1	β_2	β_3
Scenario 1	0.11 ± 0.26	0.00 ± 0.00	0.18 ± 0.34
Scenario 2	0.14 ± 0.27	0.00 ± 0.00	0.00 ± 0.00
Scenario 3	0.10 ± 0.09	0.07 ± 0.02	0.11 ± 0.02

Table 1: Variable importance using permutations

Discussion & Conclusion

Main limitations

- **Overfitting** - inherent from decision trees' structure



To wrap-up

- Decision trees are **simple** and easily **accessible**
- Such an approach 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.

Thank you for your attention!

References

Andrews DF, Hertzberg AM (1985)

Breiman, L. (2001)

Cook, R. J., & Lawless, J. (2007)

Cook, R. J., Lawless, J. F., & Nadeau, C. (1996)

Cox, D. R. (1972)

Feurer, M., & Hutter, F. (2019)

Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996)

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009)

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008)

Kaplan, E. L., & Meier, P. (1958)

Kim, S., Schaubel, D. E., & McCullough, K. P. (2018)

Kvamme, H., & Borgan, Ø. (2019)

Murris, J., Charles-Nelson, A., Lavenu, A., & Katsahian, S. (2022)

Nelson, W. B. (2003)

Therneau, T., Grambsch, P., & Fleming, T. (1990)