# Random survival forest for survival data with recurrent events

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

August 2023
44th Annual Conference of the International Society for Clinical Biostatistics
University of Milano Bicocca, Milan, Italy

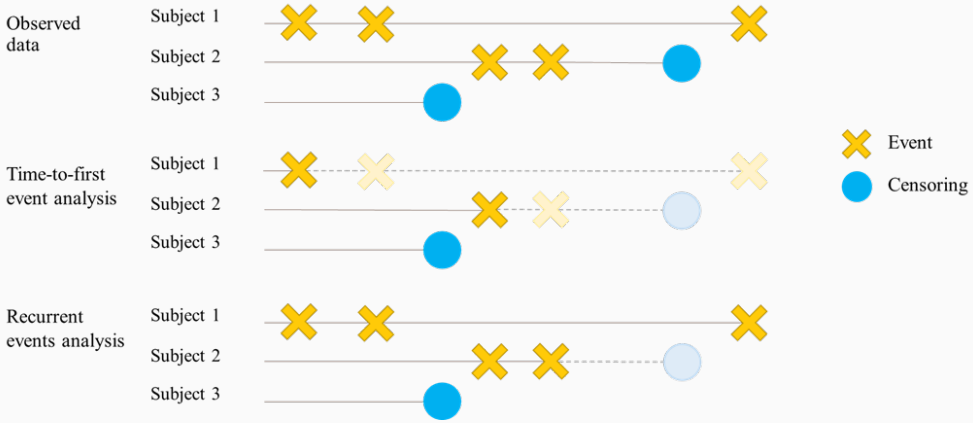
[1]HeKA, Inria Paris - Inserm, Université Paris Cité, Pierre Fabre,
[2]CIC-1414 Inserm, IRMAR - CNRS 6625, Université de Rennes 1,
[3]CIC-1418 HEGP, AP-HP, HeKA, Inria Paris - Inserm, Université Paris Cité

# Today's talk

1. Motivation for modelling recurrent events
2. Growing decision trees and ensemble random forests
3. Application based on simulation study and open-source data

# Motivation for modelling recurrent events

# What survival data are made of



# 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 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 ensemble methods**.

# Growing decision trees and ensemble random forests

# Using non-parametric principles from recurrent events analysis

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 $t$.

# Using non-parametric principles from recurrent events analysis

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

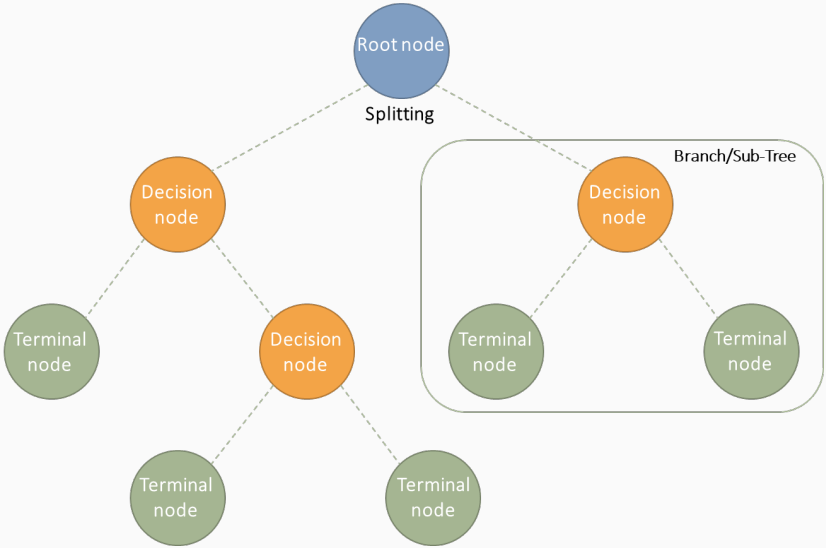
- The mean cumulative function (MCF) writes $\mu(t) = \mathbb{E}[N_i(t)]$,
- The Nelson-Aalen MCF estimator writes $\hat{\mu}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\delta(u)}$

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



## 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

### 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 $h$

### 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

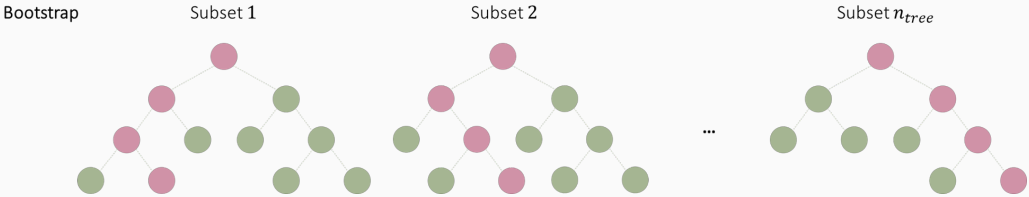
$$\hat{\mu}(t|\mathbf{x}_i) = \hat{\mu}_h(t) \times \mathbb{1}_{\mathbf{x}_i \in h}, \tag{2}$$

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

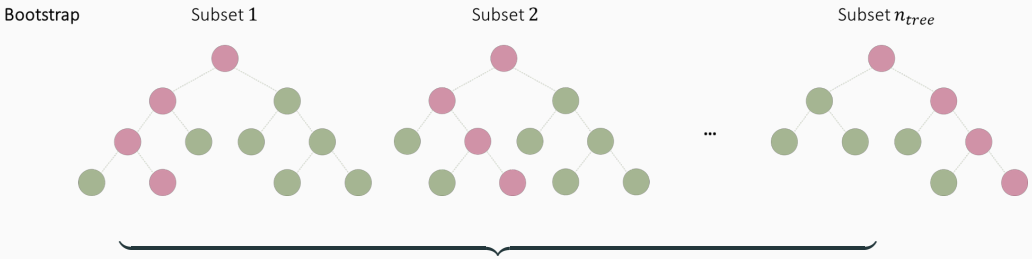
### Pruning

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

# Aggregating to build random forests



# Aggregating to build random forests




Ensemble estimator is the average of the estimates over all $n_{tree}$ trees

$$\hat{H}(t|\mathbf{x}_i) = \frac{1}{n_{tree}} \sum_{1}^{n_{tree}} \hat{\mu}(t|\mathbf{x}_i) \tag{3}$$

# Concordance error rate and evaluation

- C-index widely used as a performance metric *(Harrell, 1996)*
- Extension needed to take into account subsequent event occurrences *(Kim, 2018)*

# Concordance error rate and evaluation

- C-index widely used as a performance metric *(Harrell, 1996)*
- Extension needed to take into account subsequent event occurrences *(Kim, 2018)*

The proposed C-index is based on event occurrence rate to tackle inter-individual heterogeneity

$$\hat{\mathbb{C}}_{\text{rec}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{r_i > r_j} \times \mathbb{1}_{\hat{r}_i > \hat{r}_j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{r_i > r_j}} \tag{4}$$

with $r_i = \frac{N_i(T_i)}{T_i}$ and $\hat{r}_i = \frac{\hat{\mu}_i(T_i|\mathbf{x}_i)}{T_i}$ the observed and predicted event occurrence rates, respectively.

# Concordance error rate and evaluation

- C-index widely used as a performance metric *(Harrell, 1996)*
- Extension needed to take into account subsequent event occurrences *(Kim, 2018)*

The proposed C-index is based on event occurrence rate to tackle inter-individual heterogeneity

$$\hat{\mathbb{C}}_{\text{rec}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{r_i > r_j} \times \mathbb{1}_{\hat{r}_i > \hat{r}_j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{r_i > r_j}} \tag{4}$$

with $r_i = \frac{N_i(T_i)}{T_i}$ and $\hat{r}_i = \frac{\hat{\mu}(T_i|\mathbf{x}_i)}{T_i}$ the observed and predicted event occurrence rates, respectively.

OOB prediction error is measured by $1 - \hat{\mathbb{C}}_{\text{rec}}$.

# Application

# Simulation study

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

$$\lambda(t|z_i) = r_0(t) \times r(z_i, \beta) \tag{5}$$

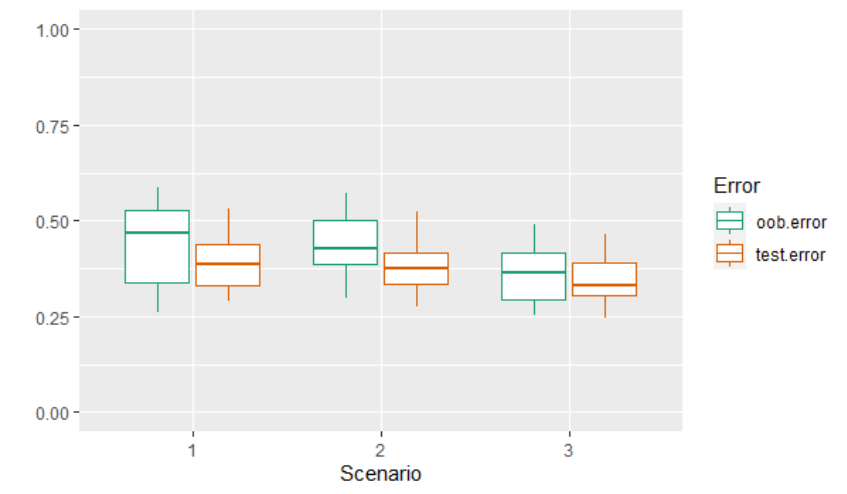
  with $r_0(t)$ the baseline hazard rate function of time $t$, $r(z_i, \beta)$ the relative risk function, and $\beta$ the covariate coefficients.

- **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 $n = 500$ stochastic processes and $p = 10$ binary predictors:

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

# Simulation study - OOB and test prediction errors

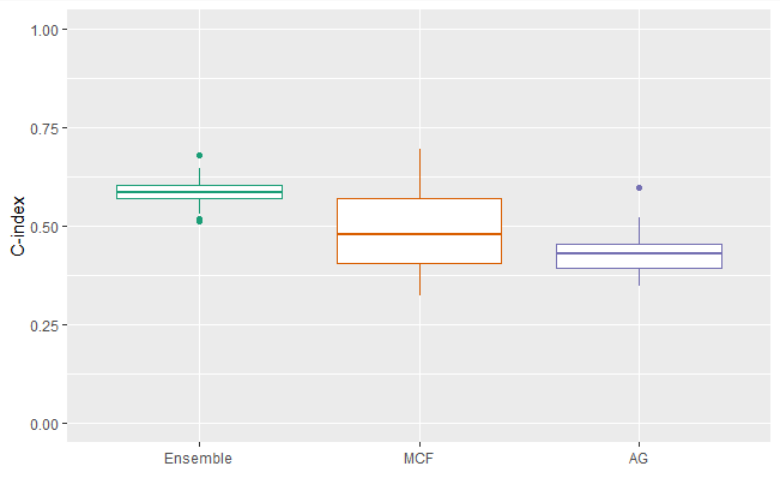



OOB and test prediction errors were estimated from 30 independent bootstrap replicates and 50 trees were grown for each random forest.

# Empirical comparison

- **Bladder** dataset from R was used
- Prediction performance was calculated using the C-index described earlier from 30 independent bootstrap replicates
- Each forest grew 50 trees
- Predictions from Nelson-Aalen MCF estimator and Andersen-Gill model were used for comparison

# Empirical comparison - prediction error




MCF = Mean cumulative function, AG = Andersen-Gill.

# Discussion & Conclusion

# To wrap-up

**Perspectives**

- Extensive experiments to be conducted on real and simulated datasets
- Feature importance
- Hyperparameter optimisation

Our approach is simple and easily accessible and constitutes a solid baseline for many extensions.

# To wrap-up

**Perspectives**

- Extensive experiments to be conducted on real and simulated datasets
- Feature importance
- Hyperparameter optimisation

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.

**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)