Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event

Juliette Murris

URC HEGP, AP-HP
May 2024

Today's talk

1. Motivating example
2. Growing decision trees and ensemble random forests
3. Application based on open-source data

Motivating example

What survival data are made of

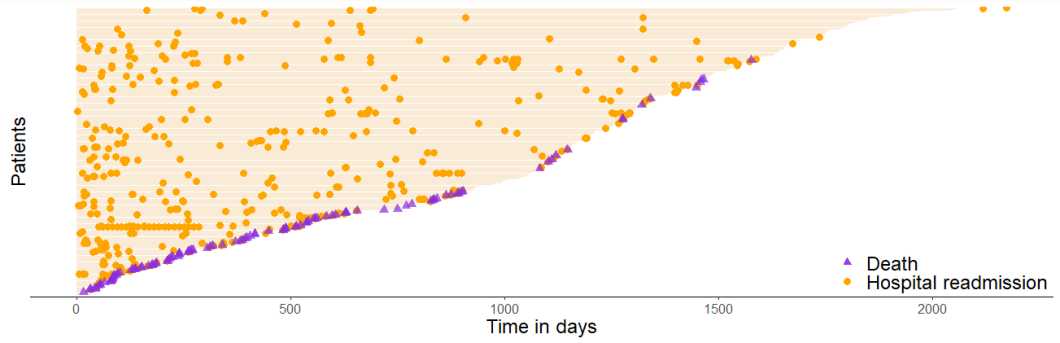


Figure 1: Readmission dataset (source: frailtypack, R)

What survival data are made of

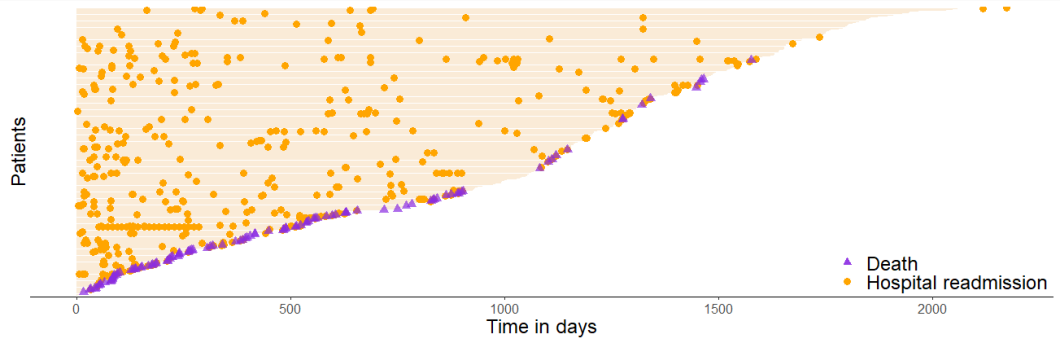


Figure 1: Readmission dataset (source: frailtypack, R)

How to predict the number of hospital readmissions over time for each patient?

What options do we have?

- Focus on first hospital readmission?
- Focus on time to death?

What options do we have?

- Focus on first hospital readmission?
- Focus on time to death?
- Focus on the number of hospital readmissions regardless of time?

What options do we have?

- Focus on first hospital readmission?
- Focus on time to death?
- Focus on the number of hospital readmissions regardless of time?
- **Focus on time to recurrent readmission**

What options do we have?

- Focus on first hospital readmission?
- Focus on time to death?
- Focus on the number of hospital readmissions regardless of time?
- **Focus on time to recurrent readmission**
- **Focus on time to recurrent readmission with a terminal event**

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, with or without a terminal event.

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, with or without a terminal event.

The objective for today is to introduce a new approach to
model recurrent events using ensemble methods.

Growing decision trees and ensemble random forests

Background on recurrent events survival analysis

Let $N(t)$ the cumulative number of events over the interval $[0, t], t \in [0, T]$ with T the longest follow-up time overall

- The mean cumulative function (MCF) writes $\mu(t) = \mathbb{E}[N(t)]$,

Background on recurrent events survival analysis

Let $N(t)$ the cumulative number of events over the interval $[0, t], t \in [0, T]$ with T the longest follow-up time overall

- The mean cumulative function (MCF) writes $\mu(t) = \mathbb{E}[N(t)]$,
- **Without** a terminal event - We use the Nelson-Aalen MCF estimator

$$\hat{\mu}(t) = \int_0^t \frac{dN(u)}{Y(u)} \tag{1}$$

with $N(t) = \sum_i N_i(t)$, and $Y(t) = \sum_i Y_i(t)$ the number of individuals at risk at time t

Background on recurrent events survival analysis

Let $N(t)$ the cumulative number of events over the interval $[0, t], t \in [0, T]$ with T the longest follow-up time overall

- The mean cumulative function (MCF) writes $\mu(t) = \mathbb{E}[N(t)]$,
- **Without** a terminal event - We use the Nelson-Aalen MCF estimator

$$\hat{\mu}(t) = \int_0^t \frac{dN(u)}{Y(u)} \tag{1}$$

  with $N(t) = \sum_i N_i(t)$, and $Y(t) = \sum_i Y_i(t)$ the number of individuals at risk at time t

- **With** a terminal event - We incorporate the Kaplan-Meier estimator of the survival function of the terminal event

$$\hat{\mu}(t) = \int_0^t \hat{S}(u) \frac{\sum_i Y_i(u) dN_i(u)}{\sum_i Y_i(u)} \tag{2}$$

Growing decision trees

Figure 2: How to grow a tree

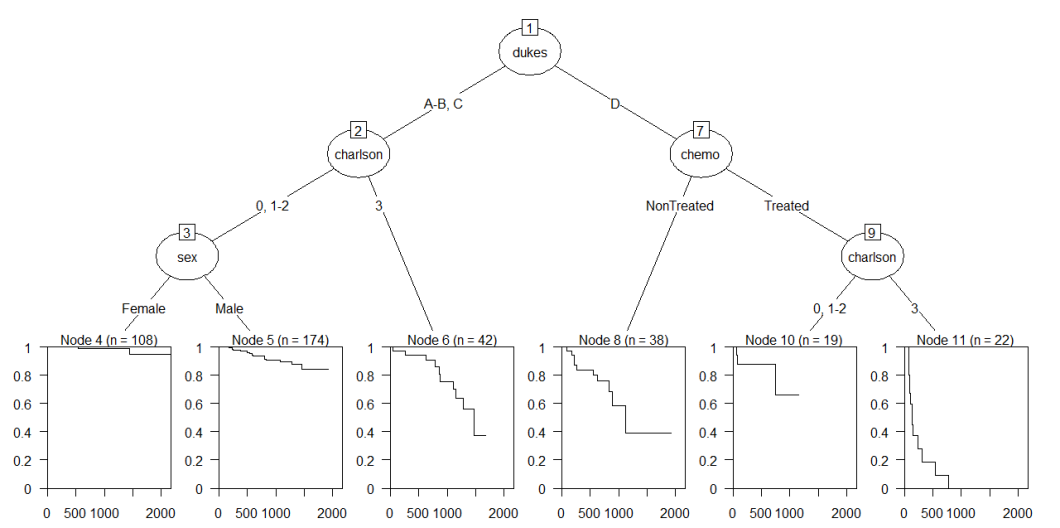
Growing decision trees

Figure 2: How to grow a tree

Bucket list:

- A **splitting** rule at each node
- A terminal node **estimator**
- A **pruning** strategy

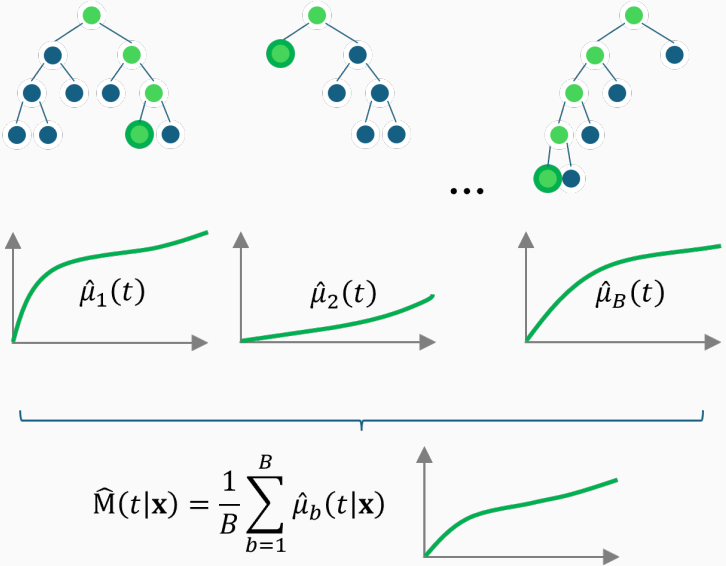
Growing survival decision trees



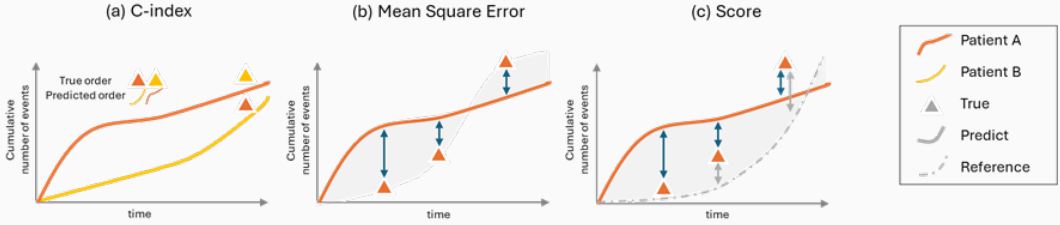
Growing survival decision trees with recurrent events

	Without a terminal event	**With** a terminal event
Splitting rule	Maximize the test statistic	
At each node, $m \in \mathbb{N}$ predictors are randomly selected	Pseudo-score test from np estimates	Wald test from Ghosh-Lin model
Terminal node estimator for tree b	$\hat{\mu}_b(t\|\mathbf{x}) = \int_0^t \frac{dN_b(u)}{Y_b(u)}$	$\hat{\mu}_b(t\|\mathbf{x}) = \int_0^t \hat{S}_b(u) \frac{\sum_i Y_{b,i}(u) dN_{b,i}(u)}{\sum_i Y_{b,i}(u)}$
Pruning strategy	A minimal number of events and/or a minimal number of individuals	

Aggregating to build random forests



Performance evaluation



Performance evaluation - (a) The concordance index

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

Performance evaluation - (a) The concordance index

- 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

$$\hat{\mathbb{C}}_{\text{rec}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{r_i > r_j} \times \mathbb{1}_{\hat{r}_i > \hat{r}_j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{r_i > r_j}} \tag{3}$$

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.

Performance evaluation - (b) The mean square error

- No MSE metric for recurrent events framework until very lately *(Bouaziz, 2023)*
- We adapted for an ensemble framework

Performance evaluation - (b) The mean square error

- No MSE metric for recurrent events framework until very lately *(Bouaziz, 2023)*
- We adapted for an ensemble framework

For each tree b,

$$\widehat{MSE}_b\left(t, \hat{\mu}_b\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\int_0^t \frac{dN_i\left(u\right)}{\hat{G}_c\left(u|\mathbf{x}\right)} - \hat{\mu}_b\left(t|\mathbf{x}\right) \right)^2 \tag{4}$$

Where $\hat{G}_c\left(u|\mathbf{x}\right) = 1 - \hat{G}\left(u - |\mathbf{x}\right)$ is an estimator of $G_c\left(u|\mathbf{x}\right) = 1 - G\left(u - |\mathbf{x}\right)$ the conditional cumulative distribution function of the censoring variable C given $\mathbf{x}$.

Performance evaluation - (b) The mean square error

- No MSE metric for recurrent events framework until very lately *(Bouaziz, 2023)*
- We adapted for an ensemble framework

For each tree b,

$$\widehat{MSE}_b\left(t, \hat{\mu}_b\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_0^t \frac{dN_i\left(u\right)}{\hat{G}_c\left(u|\mathbf{x}\right)} - \hat{\mu}_b\left(t|\mathbf{x}\right)\right)^2 \tag{4}$$

Where $\hat{G}_c\left(u|\mathbf{x}\right) = 1 - \hat{G}\left(u - |\mathbf{x}\right)$ is an estimator of $G_c\left(u|\mathbf{x}\right) = 1 - G\left(u - |\mathbf{x}\right)$ the conditional cumulative distribution function of the censoring variable C given $\mathbf{x}$.

Thus:

$$\widehat{MSE}\left(t, \hat{M}\right) = \frac{1}{B}\sum_{b=1}^{B}\widehat{MSE}_b(t, \hat{\mu}_b) \tag{5}$$

Performance evaluation - (c) The score

But

Performance evaluation - (c) The score

But

Two different models may lead to similar MSE values over time.

Performance evaluation - (c) The score

But

Two different models may lead to similar MSE values over time.

We introduce a score to represent the prediction gain compared to a reference estimator and we define for each tree b

$$Score_b\left(t, \hat{\mu}_b, \hat{\mu}_{b,0}\right) = \widehat{MSE}_b\left(t, \hat{\mu}_{b,0}\right) - \widehat{MSE}_b\left(t, \hat{\mu}_b\right) \tag{6}$$

Performance evaluation - (c) The score

But

Two different models may lead to similar MSE values over time.

We introduce a score to represent the prediction gain compared to a reference estimator and we define for each tree b

$$Score_b\left(t, \hat{\mu}_b, \hat{\mu}_{b,0}\right) = \widehat{MSE}_b\left(t, \hat{\mu}_{b,0}\right) - \widehat{MSE}_b\left(t, \hat{\mu}_b\right) \tag{6}$$

Thus:

$$Score\left(t, \hat{M}\right) = \frac{1}{B}\sum_{b=1}^{B} Score_b\left(t, \hat{\mu}_b, \hat{\mu}_{b,0}\right) \tag{7}$$

Performance evaluation - Integrated counterparts

But

Performance evaluation - Integrated counterparts

But

There is a need for the estimation of the expectation of single-time MSE and derived score over time (e.g. hyperparameter tuning, generalized metric, etc.)

Performance evaluation - Integrated counterparts

But

There is a need for the estimation of the expectation of single-time MSE and derived score over time (e.g. hyperparameter tuning, generalized metric, etc.)

$$\begin{cases} \widehat{IMSE}(\tau_1, \tau_2, \hat{M}) & = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \widehat{MSE}(t, \hat{M}) dt \\ IScore(\tau_1, \tau_2, \hat{M}) & = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} Score(t, \hat{M}) dt \end{cases}$$

(8)

With $\tau_1 = 0$ and τ_2 the maximum event time on the original sample.

Application

Empirical comparison

- **Readmission** dataset from R was used
- Multiple rehospitalizations after surgery in 403 patients diagnosed with colorectal cancer, with 199 patients with no admission and a total of 106 deaths
- Available factors: sex (M/F), chemotherapy treatment (Yes/No), Dukes' tumoral stage (with levels A-B, C, and D), and time-dependent comorbidity Charlson's index (with levels 0, 1-2, and 3)
- Predictions from np estimator and Ghosh-Lin models were used for comparison

Performance metrics

Metric	C-index ↑	IMSE ↓	IScore ↑
Np	0.58 (0.05)	7 883.50 (6 229.47)	ref.
GL1	0.53 (0.08)	7 843.99 (6 106.36)	39.41 (230.6)
GL2	0.48 (0.08)	8 361.16 (6 292.29)	-477.67 (348.48)
GL3	0.48 (0.07)	8 229.08 (6 478.35)	-345.62 (432.6)
GL4	0.45 (0.05)	9 981.50 (6 064.23)	-2 098.44 (541.59)
RecForest	**0.80 (0.04)**	**706.02 (508.96)**	**188.22 (89.00)**
GL*	0.60 (0.06)	7 934.28 (6 606.23)	51.33 (142.63)

Table 1: Means and standard deviations over the 10-fold cross-validation

Variable importance

Figure 3: Variable importance of RecForest computed on the C-index and the opposite of the integrated MSE. Charlson refers to Charlson comorbidity index, Dukes refers to tumoral Dukes stage.

Predictions

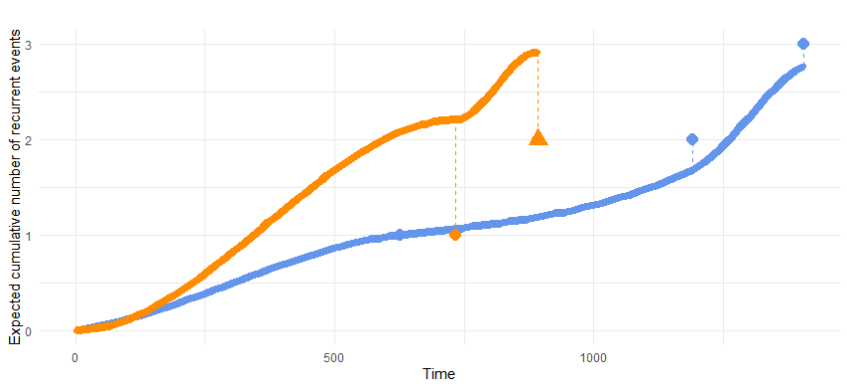


Figure 4: Expected cumulative number of recurrent events with RecForest for two patients, one in orange with the highest Charlson comorbidity score, and the other in blue with the lowest. Data points outside the prediction curves are observed data. Triangle indicates the patient died.

Discussion & Conclusion

To wrap-up

Take home messages

- Our approach is simple and easily accessible;
- **RecForest** handles longitudinal factors, terminal events, high-dimensionality, and missing data;
- Insight of interpretability with feature importance;
- Solid baseline for many extensions.

To wrap-up

Take home messages

- Our approach is simple and easily accessible;
- **RecForest** handles longitudinal factors, terminal events, high-dimensionality, and missing data;
- Insight of interpretability with feature importance;
- Solid baseline for many extensions.

For these reasons, 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)

Bouaziz, O. (2023)

Breiman, L. (2001)

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

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)