



**HAL**  
open science

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

Juliette Murriss, Olivier Bouaziz, Michal Jakubczak, Sandrine Katsahian,  
Audrey Lavenu

## ► To cite this version:

Juliette Murriss, Olivier Bouaziz, Michal Jakubczak, Sandrine Katsahian, Audrey Lavenu. Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event. 2024. hal-04612431

**HAL Id: hal-04612431**





**<https://hal.science/hal-04612431v1>**

Preprint submitted on 14 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Juliette Murriss<sup>1</sup>, Olivier Bouaziz<sup>2</sup>, Michal Jakubczak<sup>3</sup>,  
Sandrine Katsahian<sup>4</sup>, and Audrey Lavenu<sup>5</sup>\*

<sup>1</sup>*HeKA, Inria, Inserm, Centre de recherche des Cordeliers, Université Paris Cité, Paris, France, & R&D, Pierre Fabre, Boulogne-Billancourt, France, e-mail:*

[juliette.murriss-ext@aphp.fr](mailto:juliette.murriss-ext@aphp.fr)

<sup>2</sup>*Université Paris Cité, CNRS, MAP5, F-75006 Paris, France, e-mail:*

[olivier.bouaziz@parisdescartes.fr](mailto:olivier.bouaziz@parisdescartes.fr)

<sup>3</sup>*R&D, Pierre Fabre, Boulogne-Billancourt, France, & Ardigen S.A., Kraków, Poland, e-mail:*

[michal.jakubczak.ext@pierre-fabre.com](mailto:michal.jakubczak.ext@pierre-fabre.com)

<sup>4</sup>*HeKA, Inria, Inserm, Centre de recherche des Cordeliers, Université Paris Cité, Paris, France, Unité de Recherche Clinique, Hôpital Européen Georges-Pompidou, Assistance Publique – Hôpitaux de Paris (AP-HP), Centre, Paris, France, & Centre d'Investigation Clinique 1418 Épidémiologie Clinique, Paris, France, e-mail:*

[sandrine.katsahian@aphp.fr](mailto:sandrine.katsahian@aphp.fr)

<sup>5</sup>*Faculté de Médecine, Université de Rennes, Rennes, France, Institut de Recherche Mathématique de Rennes (IRMAR), Rennes, France, & Centre de Investigation Clinique 1414, Inserm, Université de Rennes, Rennes, France, e-mail:*

[audrey.lavenu@univ-rennes.fr](mailto:audrey.lavenu@univ-rennes.fr)

**Abstract:** Random survival forests (RSF) have emerged as valuable tools in medical research. They have shown their utility in modelling complex relationships between predictors and survival outcomes, overcoming linearity or low dimensionality assumptions. Nevertheless, RSF have not been adapted to right-censored data with recurrent events (RE). This work introduces RecForest, an extension of RSF and tailored for RE data, leveraging principles from survival analysis and ensemble learning. RecForest adapts the splitting rule to account for RE, with or without a terminal event, by employing the pseudo-score test or the Wald test derived from the marginal Ghosh-Lin model. The ensemble estimate is constructed by aggregating the expected number of events from each tree. Performance metrics involve a concordance index (C-index) tailored for RE analysis, along with an extension of the mean squared error (MSE). A comprehensive evaluation was conducted on both simulated and open-source data. We compared RecForest against the non-parametric mean cumulative function and the Ghosh-Lin model.

---

\*SK and AL share last authorship.

Across the simulations and application, RecForest consistently outperforms, exhibiting C-index values ranging from 0.64 to 0.80 and lowest MSE metrics. As analysing time-to-recurrence data is critical in medical research, the proposed method represents a valuable addition to the analytical toolbox in this domain.

**Keywords and phrases:** Random forests, Recurrent events, Survival analyses, Terminal events, High-dimensional data.

## 1. Introduction

Recurrent events refer to instances where individuals may experience multiple occurrences of the same event over time. In medical research, patients may face recurrent disease relapses, frequent hospitalizations, or repeated surgeries. While traditional survival analyses focus solely on the first occurrence of an event, specific statistical models have been developed to capture the complexity of recurrence in a survival framework. Intensity models rely on instantaneous hazards at each time point and account for dependence amongst event occurrences captured by time-varying covariates ([Andersen and Gill \(1982\)](#); [Prentice, Williams and Peterson \(1981\)](#)). Besides, marginal models centre on the overall distribution of event times and the cumulative event counts ([Wei, Lin and Weissfeld \(1989\)](#); [Cook, Lawless and Lee \(2010\)](#)). For a more in-depth exploration of these models concerning recurrent events, comprehensive discussions can be found in works by [Amorim and Cai \(2015\)](#) and [Ozga, Kieser and Rauch \(2018\)](#).

Time-to-event analyses are systematically challenging due to the presence of censoring, i.e. when the precise timing of an event remains unknown or unobserved. Above methodologies strictly assume the censoring process to be uninformative, hence independent of the underlying event process. Nevertheless, a terminal event may occur in competition, preventing further events of interest from happening. A terminal event is then a specific type of event considered as a termination point for the study period, making the censoring process no longer uninformative. Strategies for handling terminal events include ignoring them, although this approach is acknowledged to be flawed, or accounting for competing risks. Several pertinent statistical models enable to analyse both recurrent events and competing risks ([Charles-Nelson, Katsahian and Schramm \(2019\)](#)).

Navigating medical data introduces numerous challenges, including high-dimensionality, variable selection, and multicollinearity. To address these, survival time-to-first-event approaches have integrated statistical and machine learning techniques. In practice, various algorithms now have their

survival counterparts that are effectively employed to answer medical questions in real-world applications (Huang et al. (2023)). For instance, penalized regression methods, such as LASSO (Least Absolute Shrinkage and Selection Operator), Ridge, and Elastic-Net, have been tailored for Cox models, facilitating variable selection and regularization (Cox (1972); Tibshirani (1997)). Support-vector machines introduced by Van Belle et al. (2011), renowned for their capacity to handle high-dimensional data and non-linearity, have also been extended to survival endpoints. Likewise, random survival forests (RSF) from Ishwaran et al. (2008) embody a powerful ensemble learning technique handling interactions. The RSF algorithm has been extended to model several phenomena, such as competing risks, or longitudinal data (Ishwaran et al. (2014); Devaux et al. (2023)). However, within the survival framework, no machine learning approach has hitherto been extended to recurrent events (Murriss et al. (2023)). To address these unmet needs and confront the aforementioned challenges, we introduce the first RSF capable of handling recurrent events, with or without a terminal event. Illustrated in Figure 1, our method entails a 5-step approach that i) discerns the relevance of recurrent and terminal events, ii) grows trees to construct a coherent RSF, iii) thoroughly assesses performance, iv) provides relevant variable importance, and v) enables predictions on new data.

In this paper, we consider  $n$  individuals. Let  $N^*(t)$  be the number of recurrent events that occur in the time interval  $[0, t]$ ,  $D$  the survival time and  $C$  the censoring time. The data is made of  $(N(\cdot), \Upsilon, \delta)$  where  $N(t) = N^*(t \wedge C)$ ,  $\Upsilon = D \wedge C$ ,  $\delta = I(D \leq C)$ , where  $a \wedge b = \min(a, b)$  and  $I(\cdot)$  is the indicator function. For  $i = 1, \dots, n$ ,  $(N_i(\cdot), \Upsilon_i, \delta_i)$  are assumed to be independent replicates of  $(N(\cdot), \Upsilon, \delta)$ . The marginal mean frequency function is  $\mu(t) = \mathbb{E}[N(t)]$ . An estimator of  $\mu$  in the absence of a terminal event is the Nelson-Aalen estimator from Lawless and Nadeau (1995), that writes

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

with  $N(t) = \sum_i N_i(t)$ , and  $Y(t) = \sum_i Y_i(t)$  the number of individuals at risk at time  $t$ . In presence of a terminal event, we have  $\mu(t) = \int_0^t S(u) dR(u)$  where  $S(t) = \mathbb{P}(D \geq t)$  and  $dR(t) = \mathbb{E}[dN^*(t)|D \geq t]$  (Cook and Lawless (1997); Ghosh and Lin (2000)). The associated estimator writes

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

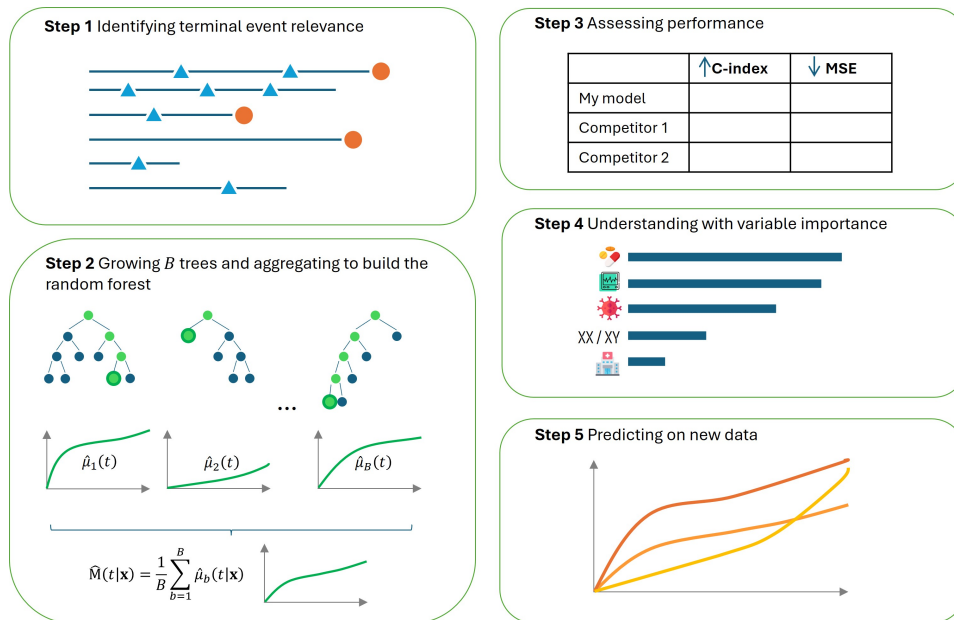


FIG 1. Scheme of the use of RecForest for survival data with recurrent events in presence or absence of a terminal event

And  $\hat{S}(t)$  is the Kaplan-Meier estimator of  $S(t)$  based on  $(Y_i, \delta_i)$  (Kaplan and Meier (1958)).

From the above considerations arises the evaluation of the provided estimations. Within survival framework, a widely common metric is an extension of the area under the ROC curve known as the concordance index (C-index). The principle of the C-index from Harrell, Lee and Mark (1996) and its derivatives as per Uno et al. (2011) is to measure the ability of a model to correctly order pairs of survival times. Recent developments from Kim, Schaubel and McCullough (2018) have expanded the application of the C-index to the recurrent event framework, incorporating the number of subsequent event occurrences. However, the number of events over time is only comparable if individuals have similar follow-up, which is hardly the case in real-world settings. Therefore, we proposed a generalized C-index by introducing event occurrence rate. Additionally, we employ the mean-square error, recently adapted to account for recurrent events by Bouaziz (2024).

Based on the non-parametric estimators and ensemble method principles, the objective of this work is to introduce a new ensemble approach, called RecForest, for the analysis of recurrent events in a survival framework, with or without a terminal event. The overall methodology based on survival decision trees and novel associated evaluation metrics is described in Section 2. Section 3 displays an extended simulation scheme for the comprehension of the proposed methodology. Illustrative examples based on open-source data are used for concrete application in Section 4.

## 2. Methodology

The proposed Algorithm 1 is an extension of the RSF introduced by Ishwaran et al. (2008). The first step is drawing bootstrap samples to prevent overfitting and capture inherent variability within the original dataset. Then, survival trees are constructed on each bootstrap sample. Unlike the original RSF, our approach accommodates for subsequent events by integrating statistical considerations tailored for recurrent events analysis. As a last step, the algorithm aggregates the results over the constructed recursive survival trees to obtain a comprehensive estimate.

Next subsections describe in further details how survival trees grow for constructing the random forest. Additionally, we provide adequate metrics for the evaluation. Finally, we expound on the computation of variable importance.

---

**Algorithm 1** Overview of RecForest algorithm

---

**Require:** Draw  $B > 0$  bootstrap samples from the learning data**for** Each node of survival tree  $b$  **do**     $mtry$  predictors are randomly selected with  $mtry \in \mathbb{N}$ ,  $mtry \leq p$ ;

A greedy algorithm for optimal threshold research is used to maximize the test statistic;

    The tree grows until the stopping rule is met based on the minimal number of events  $minsplit$  and the minimal number of individuals in terminal nodes  $nodesize$ ;    Estimate  $\hat{\mu}_b$  is computed;**end for**Estimate  $M$  is computed over the  $B$  trees.

---

**2.1. Growing trees with recurrent events****2.1.1. Splitting rules**

At each node  $h \in \mathcal{H}$ , the ongoing subsample is split into two daughter nodes denoted  $h^{(+)}$  and  $h^{(-)}$ . The aim of the split is to make the daughter nodes as different as possible with regards to the outcome. The splitting rule requires that each of the  $mtry$  randomly drawn variable is dichotomized. For continuous variables, random split points, quartiles, and deciles are considered. Let  $\mathbf{x}_h = \{A, B\}$  be the dichotomized vector of a variable inherited from  $h$ . With no terminal event, we compare the marginal mean frequency functions  $\mu_A(t)$  and  $\mu_B(t)$ . The null hypothesis is their equality. In absence of a terminal event, we use the two-sample test akin to the log-rank test from [Lawless and Nadeau \(1995\)](#). The test statistic writes  $U(t) = \int_0^t \frac{Y_A(u)Y_B(u)}{Y_A(u)+Y_B(u)} (d\hat{\mu}_A(u) - d\hat{\mu}_B(u))$ . To incorporate the presence of a terminal event, we employ the marginal Gosh-Lin (GL) model from [Ghosh and Lin \(2002\)](#) within the single variable  $\mathbf{x}_h$ . Acknowledging there are no further recurrence after the terminal event, the marginal mean up to  $t$  associated with  $\mathbf{x}_h$  is defined as  $\mu_{\mathbf{x}_h}(t) = \mathbb{E}[N^*(t)|\mathbf{x}_h] = \mu_0(t) \times \exp(\beta \mathbf{x}_h)$  with  $\mu_0$  left unspecified and  $\beta$  the regression coefficient. To accommodate longitudinal variables, the GL model considers a rate function  $d\mu_{\mathbf{x}_h}(t) = d\mu_0(t) \times \exp(\beta \mathbf{x}_h(t))$ . The Wald test statistic is then extracted from  $\mu_{\mathbf{x}_h}$  and  $d\mu_{\mathbf{x}_h}$  to test the null hypothesis of  $\beta = 0$ .

The variable selected for node  $h$  is the one that maximizes the adequate test statistic to generate  $h^{(+)}$  and  $h^{(-)}$ , based on the presence of a terminal event and/or longitudinal variables.

### 2.1.2. Terminal node estimator

Let  $b$  be a bootstrap sample from original data on which a tree is grown and  $\mathbf{x}$  a  $p$ -dimensional vector of covariates dropped down the tree. The node-specific event count  $N_b(t|\mathbf{x})$  is the number of recurrent events before censoring or a terminal event at time  $t$ . The associated number of individuals at risk  $Y_b(t|\mathbf{x})$  is the number of individuals that were not censored, or that did not encounter a terminal event by time  $t$ . We then define a tree-specific estimate as follows

$$(3) \quad \hat{\mu}_b(t|\mathbf{x}) = \hat{R}_b(t|\mathbf{x}) = \int_0^t \frac{N_b(du|\mathbf{x})}{Y_b(du|\mathbf{x})}$$

In case of the presence of a terminal event,

$$(4) \quad \hat{\mu}_b(t|\mathbf{x}) = \int_0^t \hat{S}_b(u|\mathbf{x}) d\hat{R}_b(u|\mathbf{x})$$

Individuals from the same terminal node share similar features inherited from their tree path, along with identical estimates. As per the splitting rule, the terminal node estimator depends on the presence of a terminal event in the original sample.

### 2.1.3. Pruning trees

A pruning strategy is essential to help find a trade-off to prevent overfitting and improve generalization performance of trees, within a reasonable computational time. Aligned with [Devaux et al. \(2023\)](#), we suggest two stopping rules for each terminal node: (i) a minimal number of events called *minsplit*, and (ii) a minimal number of individuals called *nodesize*. The validation of either stopping rule designates the current node  $h$  as terminal.

### 2.1.4. Handling missing data

To tackle eventual missing data, we include an adaptive-tree imputation which addresses missing data during the tree-growing stage by selectively drawing from available, non-missing, in-bag data ([Ishwaran et al. \(2008\)](#); [Chen and Xu \(2023\)](#)). At each node  $h_b$  from tree  $b$ , the method entails imputing random non-missing information specifically from the selected variables. The imputed data is then utilized for making splits within the node  $h_b$ . Imputed values are reset to missing as the tree progresses to subsequent nodes.



## 2.2. From trees to random forests

### 2.2.1. Ensemble estimates

Once all  $B$  trees are grown from the independent bootstrap samples, the ensemble estimate  $\hat{M}$  is the aggregation of all  $B$  tree-specific estimates. We define  $\hat{M}$  as

$$(5) \quad \hat{M}(t|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b(t|\mathbf{x})$$

### 2.2.2. OOB ensemble estimates

By standard bootstrap theory, each bootstrap sample leaves out circa 37% of the data (Ishwaran et al. (2008)). This is the so-called out-of-bag (OOB) sample. OOB data is used to build OOB ensembles. Let  $\mathcal{O}_i \subseteq \{1, \dots, B\}$  be the index set of trees where for  $b \in \mathcal{O}_i$ ,  $c_{i,b} = 0$ , which means the individual  $i$  is in the OOB sample. The OOB ensemble estimate  $\hat{M}^{OOB}$  of aggregated tree-specific estimates for  $i$  which is OOB writes

$$(6) \quad \hat{M}^{OOB}(t|\mathbf{x}_i) = \frac{1}{|\mathcal{O}_i|} \sum_{b \in \mathcal{O}_i} \hat{\mu}_b(t|\mathbf{x}_i)$$

OOB ensemble estimates are typically used for reporting errors.

## 2.3. Performance

Performance metrics below indicate the ability of the model to predict well from training data to unseen data. In our case, unseen data are either from the OOB sample or external validation data.

### 2.3.1. Assessing performance with relevant metrics

For the assessment of performance, we introduce an extended version of the C-index and employ the mean-square error (MSE), a derived score, and their integrated versions (Figure 2).

*Concordance index.* Kim, Schaubel and McCullough (2018) adapted the C-index to recurrent events and considered the number of events over time across individuals. This metric hence suffers from the potential bias in case of

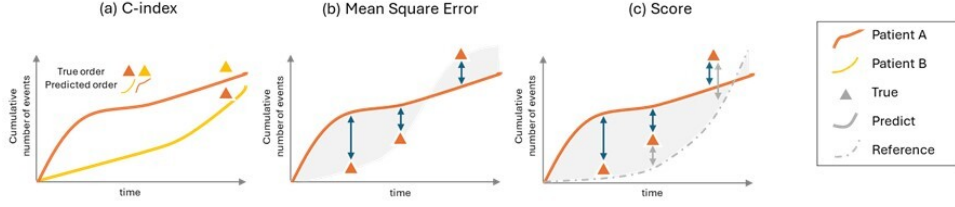


FIG 2. Illustration of the performance metrics with true and predicted cumulative number of events over time

substantial variability in the follow-up times. Individuals with longer follow-up times or a higher number of events might indeed disproportionately influence the C-index calculation. To address this issue, we suggest using occurrence rates by computing event rates per unit time.

The proposed C-index is defined as the proportion of all concordant pairs of individuals where predicted occurrence rates are correctly ordered with respect to observed occurrence rates (as shown in Figure 2a). As occurrence rates can be calculated for all individuals, including censored ones, the proposed C-index is not partial and considers all individuals in the computation. In this work, the C-index then writes

$$(7) \quad \hat{C}_{\text{rec}} = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r_i > r_j} \times \mathbb{1}_{\hat{r}_i > \hat{r}_j}}{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r_i > r_j}}$$

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. Like other C-indices, the value of the above C-index falls within the range of 0 to 1, where 1 indicates perfect concordance, and values close to 0.5 suggest randomness in the model.

*Mean-squared error and derived score.* No MSE measure has been adapted to recurrent events framework until very lately. Bouaziz (2024) filled this gap and suggested a generalization of the Brier score from Graf et al. (1999). For our problematic, for each tree  $b$ , we define

$$(8) \quad \widehat{MSE}_b(t, \hat{\mu}_b) = \frac{1}{n} \sum_{i=1}^n \left( \int_0^t \frac{dN_i(u)}{\hat{G}_c(u|\mathbf{x})} - \hat{\mu}_b(t|\mathbf{x}) \right)^2$$

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

given  $\mathbf{x}$ . We assume  $C$  and  $\mathbf{x}$  to be independent. With no terminal event,  $\hat{G}$  is the empirical cumulative distribution function of the censored variable. In the presence of a terminal event,  $\hat{G}$  is the Kaplan-Meier estimator of  $C$ . As suggested in Figure 2b), the general prediction criterion denoted  $\widehat{MSE}$  over our random forest hence writes

$$(9) \quad \widehat{MSE}(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B \widehat{MSE}_b(t, \hat{\mu}_b)$$

As pointed out in Bouaziz (2024), two different models may lead to similar MSE values over time underlining the difficulty in assessing which model is better. A score is thus introduced to represent the prediction gain compared to a reference estimator and we define for each tree  $b$ :

$$(10) \quad Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0}) = \widehat{MSE}_b(t, \hat{\mu}_{b,0}) - \widehat{MSE}_b(t, \hat{\mu}_b)$$

Where  $\hat{\mu}_b$  is the evaluated estimator and  $\hat{\mu}_{b,0}$  the reference estimator over the  $b$  samples. In our case, the reference estimator is the tree-specific non-parametric either the Nelson-Aalen or the Ghosh-Lin estimator described above. The ensemble score illustrated in Figure 2c) writes

$$(11) \quad Score(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0})$$

A higher score is associated with a better performance.

*Integrated counterparts.* Above MSE and derived score are time-dependent metrics. While they provide valuable insight of the performance for each time  $t$ , there is a need for the estimation of the expectation of single-time MSE and derived score over time (shaded areas in Figure 2). As demonstrated in Bouaziz (2024), above MSE reduces to the Brier score when individuals experience one event at most. In the spirit of the integrated version of the Brier score between two time points  $\tau_1$  and  $\tau_2$ , we integrate the MSE and the score:

$$(12) \quad \begin{cases} \widehat{IMSE}(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \widehat{MSE}(t, \hat{M}) dt \\ \widehat{IScore}(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} Score(t, \hat{M}) dt \end{cases}$$

With  $\tau_1 = 0$  and  $\tau_2$  the maximum event time on the original sample.

### 2.3.2. OOB errors

OOB errors are used for tuning hyperparameters and evaluating predictive performances and are computed on OOB samples. They are also particularly useful in the absence of external validation data or when dealing with low-dimensional original samples, where allocating a portion for validation is hardly affordable. OOB predictions are calculated by average predictions from OOB trees, and the error rate is complementary to 1. In this work, we consider the IMSE to assess the OOB error:

$$(13) \quad OOB \text{ error} = \widehat{IMSE}^{OOB}(t, \hat{M}^{OOB})$$

In this way, models exhibiting lower OOB errors are consistently favored. Of note, computing OOB errors is not recommended when the number of trees is low as each one of them may underfit.

### 2.4. Variable importance

The importance of a variable (*VImp*) is evaluated by permutation, corresponding to the impact of random perturbations in the sample on the OOB error (Breiman (2001)). To quantify the *VImp* of a covariate, a performance metric, as previously defined, is calculated following the permutation of values associated with this covariate. The *VImp* is determined as the difference between the original and permuted performance metrics. For covariate  $j$  and considering  $K$  permutations,  $VImp(j)$  writes

$$(14) \quad VImp(j) = \frac{1}{K} \sum_{k=1}^K (\hat{\theta} - \hat{\theta}_k^j)$$

With  $\hat{\theta} = \{-\widehat{IMSE}, \hat{C}\}$  the original performance metric and  $\hat{\theta}^j$  the permuted performance metric. High relative values of *VImp* indicate a loss of performance and lower/null values are interpreted as no performance for such covariates.

## 3. Simulation study

We propose the following simulation settings to illustrate the use of RecForest, inspired by Ishwaran et al. (2014); Bouaziz (2024). Simulation scenarios will cover multiple cases with associated covariates, with or without a terminal event, low- and high-dimensional data, and with or without missing

TABLE 1  
Summary of investigated scenarios

	Without a terminal event		With a terminal event	
	Low dimensional { $n = 250, p = 20$ }	High dimensional { $n = 250, p = 300$ }	Low dimensional { $n = 250, p = 20$ }	High dimensional { $n = 250, p = 300$ }
Complete	x	x	x	x
Missing	x	x	x	x
Random	x	x	x	x

data. 250 learning sets and one external validation set were generated for each scenario with  $n = 250$  individuals and  $p$  covariates. Table 1 below summarizes investigated scenarios. Next subsections further detail simulation parameters for each case. For each scenario, we grow 100 trees for RecForest. We set the following values: the minimal number of events is  $minsplitted = 5$ , the minimal number of individuals is  $nodesize = 10$ , and the number of random predictors at each node is  $mtry = \{1, \sqrt{p}, \log(p)\}$ . We compared RecForest with non-parametric estimators, as well as a semi-parametric GL model where possible. Performances were measured on the external validation set using the C-index, MSE, the score and their integrated versions.

### 3.1. Simulation scheme

#### 3.1.1. With and without a terminal event

For  $i = 1, \dots, n$ ,  $p_0$ -dimensional covariate vector,  $X_i = (X_{i,1}, \dots, X_{i,p_0})$ ,  $X_{i,1:\lfloor \frac{p_0}{2} \rfloor} \mathcal{B}(0.5)$  Bernoulli variables and  $X_{i,\lfloor \frac{p_0}{2} \rfloor+1:p_0} \mathcal{N}(2, 0.5)$  Gaussian variables are simulated. Recurrent events are generated from a non-homogenous Poisson process  $\lambda(t|X_i) = \lambda_0(t) \exp(\beta^T X_i)$  with  $\lambda_0(t) = \frac{\alpha}{\gamma} (\frac{t}{\gamma})^{\alpha-1}$  a Weibull baseline,  $\alpha = 2$  the shape parameter and  $\gamma = 0.39$  the scale parameter. The first ten associated coefficients are non-zero, with  $\beta = (\beta_1, \dots, \beta_{p_0})^T$  and  $\beta_1 = \log(5)$ ,  $\beta_{2:4} = \log(1.3)$ ,  $\beta_{5:p_0} = \log(0.7)$ . The true expected number of events is  $\mu^*(t|X_i) = \int_0^t \lambda(u|X_i) du = (\frac{t}{\gamma})^\alpha \exp(\beta^T X_i)$ . The censoring process is then simulated based on a uniform distribution  $\mathcal{U}(0, 3)$ . With a terminal event, the recurrent event process and covariates are simulated in the same way as above. The censoring process is simulated based on a uniform distribution  $\mathcal{U}(0, 8)$ . The terminal event is simulated using a Cox model with shape parameter is 8 and scale parameter is 1.8. The same covariates as the recurrent event process are included with same coefficients  $\beta$ . We set  $p_0 = 10$ .

### 3.1.2. Low- and high-dimensional scenarios

To define low- and high-dimensional scenarios, we introduce  $q$  independent noise covariates randomly drawn from a standard normal distribution and add them to simulated datasets. We set  $q = 10$  for low-dimensional scenarios, and  $q = 290$  in high-dimensional scenarios. The total number of covariates for each scenario is  $p = p_0 + q$ .

*Complete.* When we analyze scenarios that involve all  $p$  generated covariates, we refer to these as ‘complete’ datasets analyses.

*Missing.* To simulate real-world conditions where datasets may have missing values, we intentionally introduced missing data. Specifically, we randomly set 5% of the covariate  $X_1$  to *NA* across all individuals. This was done in a completely random manner, ensuring that the missing data does not follow any pattern and is not dependent on any other variables or the values of  $X_1$  itself. The missing data mechanism is completely at random. We refer to such scenarios as ‘Missing’.

*Random.* We created scenarios where the covariates are generated independently of the recurrent events to simulate a situation where no underlying factors influence the counting process. In such cases,  $q$  independent noise covariates are randomly drawn from a standard normal distribution. We created scenarios where the covariates are generated independently of the recurrent events to simulate a situation where no underlying factors influence the counting process. We set  $q = 20$  for low-dimensional scenarios, and  $q = 300$  in high-dimensional scenarios. The total number of covariates for each scenario is  $p = q$ , ensuring that all covariates are unassociated with the event data and are purely random. We refer to these scenarios as ‘Random’.

## 3.2. Results

Performances were assessed in a framework of 250 training sets and one external validation set. The non-parametric estimator uses no covariates, regardless of the dimensionality by construction. For the GL model, no variable selection was performed, meaning all  $p$  covariates were included in the model. This limits the analysis for the GL model to low-dimensional scenarios only.

TABLE 2  
Means and standard deviations of the C-index without a terminal event

Scenario\Model	Np	GL	RecForest $mtry = 1$	RecForest $mtry = \sqrt{p}$	RecForest $mtry = \log(p)$
Low dimensional $\{n = 250, p = 20\}$					
Complete		0.55 (0.12)	0.68 (0.08)	0.71 (0.04)	0.70 (0.05)
Missing	0.55 (0.04)	0.52 (0.10)	0.65 (0.14)	0.69 (0.15)	0.67 (0.15)
Random		0.49 (0.05)	0.56 (0.15)	0.54 (0.15)	0.58 (0.18)
High dimensional $\{n = 250, p = 300\}$					
Complete		/	0.67 (0.21)	0.70 (0.11)	0.70 (0.17)
Missing	0.55 (0.04)	/	0.60 (0.29)	0.64 (0.18)	0.63 (0.24)
Random		/	0.51 (0.31)	0.55 (0.25)	0.56 (0.29)

Np = non-parametric estimator; GL = Gosh-Lin model with no variable selection. RecForest was trained with fixed values for  $minsplit = 5$  and  $nodesize = 10$ . 250 learning sets and one external validation set were generated for each scenario. Values closer to 1 indicate higher performance.

### 3.2.1. Without a terminal event

On average, 62% of the individuals experienced at least one recurrent event, 46% had at least two recurrent events, 26% had at least five recurrent events, and circa four recurrent events per individual. Table 2 below reports performances in terms of C-index values. Overall, performances based on C-index values are greater in scenarios with neither missing data, nor high-dimensionality, both in average and in variability. As expected, scenarios with random inputs lead to randomness with C-index values neighboring 0.50 for each model. The non-parametric estimator provides an average C-index of 0.55. The GL model seems to suffer from not being well-specified, with average C-index values ranging from 0.49 to 0.55 where assessable. RecForest consistently outperforms, irrespective of  $mtry$  with values ranging from 0.64 up to 0.71 (random scenarios are not deemed for comparing performance). Besides, it is not impacted by the introduction of massive noisy data, as C-index values remain similar across low- and -high-dimensional scenarios. Table 3 outlines performances in terms of integrated scores. As checked with C-indices, there is no expectations in the interpretation of the random scenarios, hence there are not displayed. The non-parametric estimator is the reference model in the computation of the score. Similar conclusions may be drawn from the different scenarios with the outperformance of RecForest. Yet, higher variability is observed, especially when introducing missing data.

TABLE 3  
Means and standard deviations of the integrated score without a terminal event

Scenario\Model	GL	RecForest $mtry = 1$	RecForest $mtry = \sqrt{p}$	RecForest $mtry = \log(p)$
Low dimensional $\{n = 250, p = 20\}$				
Complete	50.45 (41.00)	208.10 (102.42)	539.49 (451.96)	161.12 (87.65)
Missing	35.78 (17.91)	325.30 (189.63)	498.75 (415.28)	258.90 (112.14)
High dimensional $\{n = 250, p = 300\}$				
Complete	/	309.47 (134.57)	388.20 (226.95)	355.65 (117.78)
Missing	/	398.70 (229.57)	574.50 (318.75)	475.20 (213.80)

GL = Gosh-Lin model with no variable selection. RecForest was trained with fixed values for  $minsplit = 5$  and  $nodesize = 10$ . 250 learning sets and one external validation set were generated for each scenario. Higher values indicate higher performance.

### 3.2.2. With a terminal event

On average, 44% of individuals experienced a terminal event during the observation period. Overall, performance discrepancies in terms of C-index values (Table 4) were observed when dealing with missing data or randomness, with performance being notably lower compared to complete data, as expected. The non-parametric estimator exhibited poor performance (C-index = 0.52 (0.03)). In both low and high-dimensional datasets, RecForest tends to perform better with higher  $mtry$  values, always reporting higher C-index values compared to non-parametric estimator and GL model. However, it is notable that in the high-dimensional scenarios, the performance drop is more significant. GL model performs better than non-parametric estimator only with complete data scenarios (C-index = 0.57 (0.07)). Integrated scores for evaluating approaches with a terminal event are displayed in Table 5. We observe similar results than without a terminal event. RecForest consistently yields integrated score values exceeding 300. The decrease in performance compared to the GL model is evident, with IScore values of 110.86 (75.14) and 112.81 (77.64) observed in complete and missing data scenarios, respectively.

In summary of the simulation study, our findings illustrate RecForest superior performance across all examined scenarios. Unlike the comparator GL model, RecForest effectively addresses both missing data and high-dimensionality. Furthermore, in random scenarios, RecForest outputs randomness, implying its reliability when the input lacks discernible patterns.



TABLE 4  
Means and standard deviations of the C-index with a terminal event

Scenario\Model	Np	GL	RecForest <i>mtry</i> = 1	RecForest <i>mtry</i> = $\sqrt{p}$	RecForest <i>mtry</i> = $\log(p)$
Low dimensional $\{n = 250, p = 20\}$					
Complete		0.57 (0.07)	0.79 (0.05)	0.80 (0.04)	0.82 (0.04)
Missing	0.52 (0.03)	0.51 (0.11)	0.73 (0.19)	0.71 (0.13)	0.75 (0.11)
Random		0.45 (0.10)	0.53 (0.16)	0.51 (0.11)	0.50 (0.19)
High dimensional $\{n = 250, p = 300\}$					
Complete		/	0.71 (0.19)	0.69 (0.13)	0.74 (0.10)
Missing	0.52 (0.03)	/	0.64 (0.20)	0.68 (0.13)	0.71 (0.11)
Random		/	0.49 (23.10)	0.48 (12.30)	0.50 (20.09)

Np = non-parametric estimator; GL = Gosh-Lin model with no variable selection. RecForest was trained with fixed values for *minsplit* = 5 and *nodesize* = 10. 250 learning sets and one external validation set were generated for each scenario. Values closer to 1 indicate higher performance.

TABLE 5  
Means and standard deviations of the integrated score with a terminal event

Scenario\Model	GL	RecForest <i>mtry</i> = 1	RecForest <i>mtry</i> = $\sqrt{p}$	RecForest <i>mtry</i> = $\log(p)$
Low dimensional $\{n = 250, p = 20\}$				
Complete	110.86 (75.14)	315.76 (119.41)	446.14 (410.88)	410.90 (115.54)
Missing	112.81 (77.64)	368.62 (211.11)	406.20 (275.57)	392.85 (138.23)
High dimensional $\{n = 250, p = 300\}$				
Complete	/	547.89 (229.37)	589.14 (472.33)	628.67 (122.41)
Missing	/	392.34 (39.16)	578.52 (336.71)	512.85 (441.29)

GL = Gosh-Lin model with no variable selection. RecForest was trained with fixed values for *minsplit* = 5 and *nodesize* = 10. 250 learning sets and one external validation set were generated for each scenario. Higher values indicate higher performance.

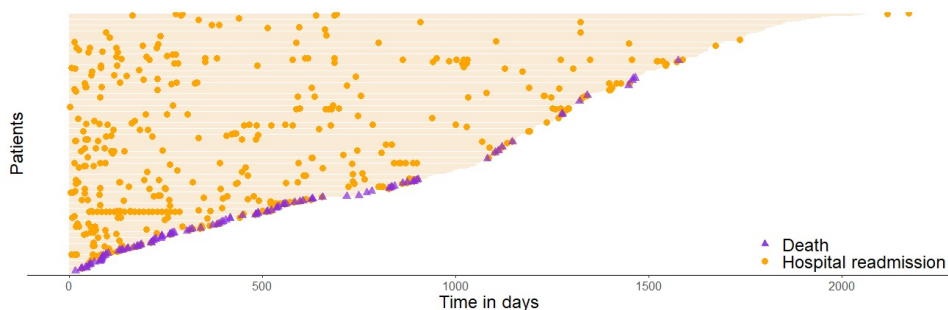


FIG 3. Event plot for readmission data

#### 4. Illustrative example: the readmission data

Readmission dataset from the `frailtypack` R package from [Rondeau, Marzroui and Gonzalez \(2012\)](#) is widely used to demonstrate methodological principles from recurrent events analysis in presence of a terminal event. The data consist of multiple rehospitalizations after surgery in 403 patients diagnosed with colorectal cancer. Available factors are sex (MF), chemotherapy treatment (YesNo), Dukes' tumoral stage (with levels A-B, C, and D), and time-dependent comorbidity Charlson's index (with levels 0, 1-2, and  $\geq 3$ ). In average, there were 1.13 (min. – max. = 0 – 22) hospital readmissions per patients, with 199 patients with no admission and a total of 106 deaths (Figure 3).

In absence of an external validation set, performances were assessed with a 10-fold cross-validation procedure. We consider the following models: four multivariate Ghosh-Lin models with arbitrary combinations of factors, and `RecForest`. The reference model is the non-parametric estimator. Hyperparameters from `RecForest` `minsplit`, `nodesize` and `mtry` were tuned on the total sample and the OOB score was minimized for  $\{ntrees = 100, minsplit = 2, nodesize = 1, mtry = 2\}$  (Figure 6 in the Supplementary).

In our analysis (results in Table 6), the non-parametric estimator registers a C-index = 0.58 (0.05). `RecForest` outperforms with C-index = 0.80 (0.04). All GL models, with one to four covariates for adjustment, maintain relatively consistent C-indices around 0.45 to 0.53. Comparing IMSE and IScore metrics, `RecForest` and the non-parametric estimator are not directly comparable due to construction. Specifically, the non-parametric reference for the integrated score in `RecForest` is constructed for each bootstrap sample. Integrated scores for GL models operate on the overall dataset from the

TABLE 6  
Means and standard deviations over the 10-fold cross-validation for *readmission* dataset

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

Np = non-parametric estimator; GL1 = Gosh-Lin model with sex; GL2 = Gosh-Lin with sex and chemotherapy; GL3 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage; GL4 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage and Charlson's index; GL\* = Ghosh-Lin model with best variables from RecForest.

Arrows indicate whether higher are lower scores lead to best performances.

ongoing fold. Consequently, IScore values from RecForest do not simply reflect the difference between IMSE values from the non-parametric estimator and RecForest, as opposed to IScores from GL models.

IMSE and IScore for RecForest indicate lower margin of errors. Among GL models, GL1 (Gosh-Lin model with sex) exhibits lower IMSE than the non-parametric estimator, resulting in a higher IScore. GL2 to GL4 (GL2 = Gosh-Lin with sex and chemotherapy; GL3 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage; GL4 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage and Charlson's index) yield negative IScore values, indicating high variability among GL models observed in our simulation study. Besides, high variability is observed across all approaches.

Variable importance for RecForest was based on both the C-index and the opposite of the integrated MSE (Figure 4). Most important variable identified by RecForest was the Charlson comorbidity index. Sex and chemotherapy did not seem to have an impact on the predictive performance. Variable selection enabled to reach better performance for GL\* model.

Prediction curves for RecForest as the expected number of recurrent events are displayed in Figure 5. Predictions were generated for two patients, one with the highest Charlson comorbidity score (in orange), and the other with the lowest (in blue). We observe for the patient in orange that the model predicted an expected number of three readmissions as the patient dies after two observed readmissions. For the patient in blue, the model predictions are in line with observed events.

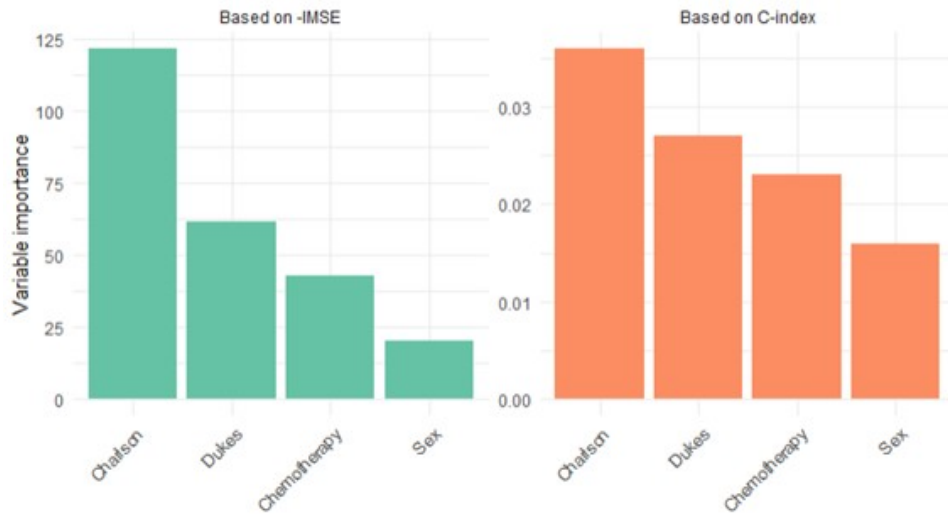


FIG 4. Variable importance of RecForest computed on the C-index and the opposite of the integrated MSE. Charlson refers to Charlson comorbidity index, Dukes refers to tumoral Dukes stage.

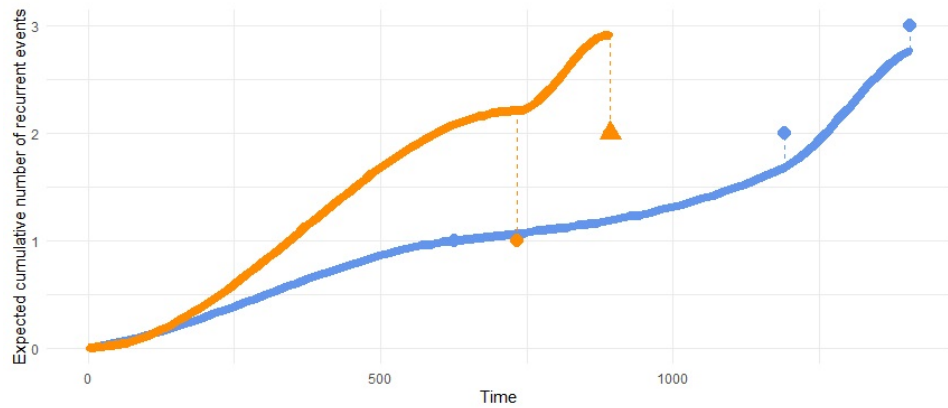


FIG 5. Expected cumulative number of recurrent events with RecForest for two patients, one in orange with the highest Charlson comorbidity score, and the other in blue with the lowest. Data points outside the prediction curves are observed data. Triangle indicates the patient died.

## 5. Discussion

We developed RecForest by extending the RSF algorithm to handle recurrent events in a survival framework, in potential presence of a terminal event, of longitudinal markers, and of missing data. To do so, the splitting rule at each node was tailormade for recurrent events analysis and mean cumulative number of events served in terminal node estimators. We characterized the performance both with discrimination and calibration by introducing a generalized C-index for recurrent event analysis and applying an innovative MSE.

In a simulation study, we compared RecForest with two baseline approaches: a non-parametric estimator and a GL model. Scenarios included variations of the presence of a terminal event, the low- or high-dimensionality of the data, and the inclusion of missing values. In instances where missing values or high-dimensional data were present, greater variability was observed across all scenarios for both metrics. In all explored cases, RecForest demonstrated higher performances both in terms of C-index and integrated score values compared with baseline. The impact was quite little when introducing massive noisy data. Besides, RecForest emerged as the sole modeling approach capable of handling high-dimensionality scenarios with no prior variable selection. Furthermore, we presented a practical application using well-known open-source data, showcasing how the fine-tuning of RecForest hyperparameters leads to a more performant model. Again, RecForest exhibited superior predictive performance, achieving a C-index of 0.80 alongside strong calibration metrics (IMSE = 706, IScore = 188). Overall, across both simulated and real-world datasets, RecForest consistently emerged as the most effective modeling approach.

In practical applications, high-dimensional problems involving recurrent events are often sidestepped by transforming the recurrent event survival framework into alternative formats, such as an event count, a time-to-first-event endpoint, or a classification problem. However, each of these transformations may lead to the voluntary omission of valuable information. In response to this, RecForest aims to bridge a recognized gap in handling such applications, ensuring a more comprehensive and nuanced analysis of recurrent event data. Additionally, our algorithm benefits from random forests features, i.e. the ability of handling missing data or multicollinearity, and reducing overfitting thanks to bagging principle.

Additional settings can be explored to integrate a terminal event within the proposed approach. For instance, [Charles-Nelson, Katsahian and Schramm \(2019\)](#) suggested working with inverse probability of survival weighting (IPSW)

to compute coefficient weights in the Ghosh-Lin model, whereas we used inverse probability of censoring weighting. IPSW is typically recommended when modeling the terminal event is also of interest. Another example would be to use frailty models, either joint or additive as per [Rondeau, Marzroui and Gonzalez \(2012\)](#). Besides, natural extensions of random forests, serving as ensemble methods, have been widely used to improve performance through boosting techniques like Gradient Boosting ([Friedman \(2001\)](#)), Extreme Gradient Boosting ([Chen and Guestrin \(2016\)](#)), or LightGBM ([Ke et al. \(2017\)](#)). Since all these methods are grounded in tree-based structures, they offer seamless extensions to the proposed approach, and would hence provide innovative tools for recurrent event analysis.

Our methodology also suffers from several drawbacks. The primary limitation of random forest-like algorithms lies in the computation time, which grows with the number of trees, the dimensionality of the data and the numbers of variables selected at each tree node. Second, we assumed the proportional hazard assumption of the Gosh model, which may not universally hold in real-world settings. Furthermore, variable importance measures provided do not account for potential correlations. To address this limitation, the implementation of grouped variable importance statistics is a promising avenue for further refinement ([Devaux et al. \(2023\)](#); [Gregorutti, Michel and Saint-Pierre \(2015\)](#)). Nevertheless, signs of associations would still be unavailable. Another potential limitation of random forests is their static usage of features. Dynamic predictions could indeed be included as per [Cottin et al. \(2022\)](#) and [Moradian et al. \(2022\)](#).

On the other hand, the issue of interpretability in machine and deep learning, particularly in digital health has garnered significant attention, as pointed out in [Farah et al. \(2023\)](#). Several explainability methods have been proposed such as Local Interpretable Model-agnostic Explanations (LIME, [Ribeiro, Singh and Guestrin \(2016\)](#)), SHapley Additive exPlanations (SHAP, [Lundberg and Lee \(2017\)](#)), and counterfactual explanations ([Guidotti \(2022\)](#); [Bhan et al. \(2023\)](#)). These interpretability techniques have been recently adapted for survival analysis ([Cottin et al. \(2024\)](#); [Kovalev, Utkin and Kasimov \(2020\)](#)). Moreover, random forest-like methods offer a valuable tool for variable selection, especially in addressing high-dimensionality or obtaining hazard ratios that are intrinsically interpretable ([Khan and Shaw \(2016\)](#); [Wang and Li \(2017\)](#)). Approaches such as permutation-based selection, variable hunting, and iterative feature elimination serve as effective means towards this purpose ([Genuer, Poggi and Tuleau-Malot \(2010\)](#); [Ishwaran et al. \(2010\)](#); [Pang et al. \(2012\)](#)).

## **6. Conclusion**

To conclude, we introduced a new algorithm based on survival theory for recurrent events with or without a terminal event and ensemble-based methodology for learning. **RecForest** is readily accessible to adequately answer further clinical needs.

## **Funding**

Author Murriss J reports a grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701.

## Supplementary Material

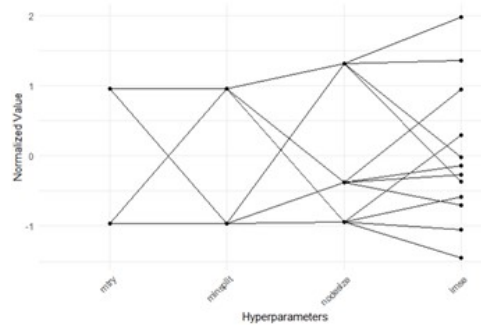


FIG 6. Hyperparameter optimization on readmission data based on out-of-bag scores

### Hyperparameter optimization on readmission data based on out-of-bag scores

Hyperparameter optimization was performed using a grid search approach.

## References

- AMORIM, L. D. A. F. and CAI, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology* **44** 324–333.
- ANDERSEN, P. K. and GILL, R. D. (1982). Cox’s Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* **10** 1100–1120.
- BHAN, M., VITTAUT, J.-N., CHESNEAU, N. and LESOT, M.-J. (2023). TIGTEC: Token Importance Guided TEXT Counterfactuals. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part III* 496–512. Springer-Verlag, Berlin, Heidelberg.
- BOUAZIZ, O. (2024). Assessing model prediction performance for the expected cumulative number of recurrent events. *Lifetime Data Analysis* **30** 262–289.
- BREIMAN, L. (2001). Random Forests. *Machine Learning* **45** 5–32.
- CHARLES-NELSON, A., KATSAHIAN, S. and SCHRAMM, C. (2019). How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery. *Statistics in Medicine* sim.8168.



- CHEN, T. and GUESTRIN, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16* 785–794. Association for Computing Machinery, New York, NY, USA.
- CHEN, S. and XU, C. (2023). Handling high-dimensional data with missing values by modern machine learning techniques. *Journal of Applied Statistics* **50** 786–804.
- COOK, R. J. and LAWLESS, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16** 911–924.
- COOK, R. J., LAWLESS, J. F. and LEE, K.-A. (2010). A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine* **29** 694–707.
- COTTIN, A., PECUCHET, N., ZULIAN, M., GUILLOUX, A. and KATSAHIAN, S. (2022). IDNetwork: A deep illness-death network based on multi-state event history process for disease prognostication. *Statistics in Medicine* **41** 1573–1598.
- COTTIN, A., ZULIAN, M., PÉCUCHET, N., GUILLOUX, A. and KATSAHIAN, S. (2024). MS-CPFI: A model-agnostic Counterfactual Perturbation Feature Importance algorithm for interpreting black-box Multi-State models. *Artificial Intelligence in Medicine* **147** 102741.
- COX, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34** 187–202.
- DEVAUX, A., HELMER, C., GENUER, R. and PROUST-LIMA, C. (2023). Random survival forests with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research* **32** 2331–2346.
- FARAH, L., MURRIS, J. M., BORGET, I., GUILLOUX, A., MARTELLI, N. M. and KATSAHIAN, S. I. M. (2023). Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence-Based Health Technologies: What Healthcare Stakeholders Need to Know. *Mayo Clinic Proceedings: Digital Health* **1** 120–138.
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29** 1189–1232.
- GENUER, R., POGGI, J.-M. and TULEAU-MALOT, C. (2010). Variable selection using random forests. *Pattern Recognition Letters* **31** 2225–2236.
- GHOSH, D. and LIN, D. Y. (2000). Nonparametric Analysis of Recurrent Events and Death. *Biometrics* **56** 554–562.
- GHOSH, D. and LIN, D. Y. (2002). Marginal Regression Models for Recurrent and Terminal Events. *Statistica Sinica* **12** 663–688.
- GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for sur-

- vival data. *Statistics in Medicine* **18** 2529–2545.
- GREGORUTTI, B., MICHEL, B. and SAINT-PIERRE, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis* **90** 15–35.
- GUIDOTTI, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*.
- HARRELL, F. E., LEE, K. L. and MARK, D. B. (1996). MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. *Statistics in Medicine* **15** 361–387.
- HUANG, Y., LI, J., LI, M. and APARASU, R. R. (2023). Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC medical research methodology* **23** 268.
- ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *The Annals of Applied Statistics* **2** 841–860.
- ISHWARAN, H., KOGALUR, U. B., GORODESKI, E. Z., MINN, A. J. and LAUER, M. S. (2010). High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association* **105** 205–217.
- ISHWARAN, H., GERDS, T. A., KOGALUR, U. B., MOORE, R. D., GANGE, S. J. and LAU, B. M. (2014). Random survival forests for competing risks. *Biostatistics (Oxford, England)* **15** 757–773.
- KAPLAN, E. L. and MEIER, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53** 457–481.
- KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q. and LIU, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems* **30**. Curran Associates, Inc.
- KHAN, M. H. R. and SHAW, J. E. H. (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing* **26** 725–741.
- KIM, S., SCHAUBEL, D. E. and MCCULLOUGH, K. P. (2018). A C-index for recurrent event data: Application to hospitalizations among dialysis patients. *Biometrics* **74** 734–743.
- KOVALEV, M. S., UTKIN, L. V. and KASIMOV, E. M. (2020). SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems* **203** 106164.
- LAWLESS, J. F. and NADEAU, C. (1995). Some Simple Robust Methods for

- the Analysis of Recurrent Events. *Technometrics* **37** 158–168.
- LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17* 4768–4777. Curran Associates Inc., Red Hook, NY, USA.
- MORADIAN, H., YAO, W., LAROCQUE, D., SIMONOFF, J. S. and FRYDMAN, H. (2022). Dynamic estimation with random forests for discrete-time survival data. *Canadian Journal of Statistics* **50** 533–548.
- MURRIS, J., CHARLES-NELSON, A., TADMOURI SELLIER, A., LAVENU, A. and KATSAHIAN, S. (2023). Towards filling the gaps around recurrent events in high dimensional framework: a systematic literature review and application\*. *Biostatistics & Epidemiology* **7** e2283650.
- OZGA, A.-K., KIESER, M. and RAUCH, G. (2018). A systematic comparison of recurrent event models for application to composite endpoints. *BMC medical research methodology* **18** 2.
- PANG, H., GEORGE, S. L., HUI, K. and TONG, T. (2012). Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **9** 1422–1431.
- PRENTICE, R. L., WILLIAMS, B. J. and PETERSON, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68** 373–379.
- RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16* 1135–1144. Association for Computing Machinery, New York, NY, USA.
- RONDEAU, V., MARZROUI, Y. and GONZALEZ, J. R. (2012). frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *Journal of Statistical Software* **47** 1–28.
- TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16** 385–395.
- UNO, H., CAI, T., PENCINA, M. J., D'AGOSTINO, R. B. and WEI, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30** 1105–1117.
- VAN BELLE, V., PELCKMANS, K., VAN HUFFEL, S. and SUYKENS, J. A. K. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in*

*Medicine* **53** 107–118.

WANG, H. and LI, G. (2017). A Selective Review on Random Survival Forests for High Dimensional Data. *Quantitative bio-science* **36** 85–96.

WEI, L. J., LIN, D. Y. and WEISSFELD, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association* **84** 1065–1073.