AIMS Mathematics, 7(10): 18099–18126.
DOI: 10.3934/math.2022996
Received: 25 April 2022
Revised: 25 July 2022
Accepted: 02 August 2022
Published: 09 August 2022

http://www.aimspress.com/journal/Math

*Research article*

# A survival tree for interval-censored failure time data

**Jia Chen**[1,2] **and Renato De Leone**[1,*]

[1] School of Science and Technology, University of Camerino, Camerino 62032, Italy

[2] School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

* **Correspondence:** Email: renato.deleone@unicam.it; Tel: +390737402532.

**Abstract:** Interval-censored failure time data as a general type of survival data often arises in medicine and other applied fields. Survival tree is a flexible predictive method for survival data because no specific assumptions are required.
Generalized Log-Rank Test have good power with parameters for interval-censored failure time data. We construct a special test statistic of Generalized Log-Rank Tests, and propose a new survival tree with hyper-parameter by combining the test statistic with Conditional Inference Framework for interval-censored failure time data. The effect of tuning hyper-parameter are discussed and hyper-parameter tuning allows the tree method to be more general and flexible. Thus the tree method either improve upon or remain competitive with existing tree method for interval-censored failure time data-ICtree, which is a special case of ours. An extensive simulation is executed to assess the predictive performance of our tree methods. Finally, the tree methods are applied to a tooth emergence data.

**Keywords:** survival tree; interval-censored; conditional inference framework; hyper-parameter tuning; generalized log-rank tests
**Mathematics Subject Classification:** 62-08, 62N03

## 1. Introduction

Interval-censored data is common in real world. For instance, in clinical trials or longitudinal study, patients have periodic follow-up monitored discontinuously, hence the patient's exact time of event of interest is not observed, but only known to fall in an interval (the endpoints of an interval usually represent two examination time). In those cases, interval censoring data occurs. Interval censoring is a more general type of censoring in survival analysis, and both left-censoring and right-censoring are special cases of it. When $R = \infty$ or $L = 0$ which represent the right endpoint and left endpoint of the observed interval respectively, then the interval censoring becomes a natural generalization of the

common right censoring and the less common left censoring case. We have an exact observation when $L = R$. If we can only observe each study subject once and either $L = 0$ or $R = \infty$, then it is referred to Case I interval-censored data. Otherwise if the interval-censored data include some finite intervals and $L$ is not always equal to 0, then it is referred to Case II interval-censored data [1].

Sun provides comprehensive overviews of statistical models and methods for analyzing interval-censored data. The proportional hazard model proposed by Cox is a semi-parametric model for right censored data [2], then developed by Finkelstein for interval-censored data [3]. Other semi-parametric models have been mentioned, such as the proportional odds model, the additive hazards model, the accelerated failure time model. In addition, Exponential model, Weibull model, Log-normal model as parametric models are also discussed [4].

Compared with statistical models and methods, tree-based methods flexibly and effectively handle high-dimensional problems. Moreover, no specific assumptions are required. The crucial points of a tree method are splitting and stopping criteria. The general splitting criterion is to partition the covariate space recursively by maximizing between-node heterogeneity or minimizing within-node homogeneity. Stopping criteria is to decide the final size of a tree and some tree methods have a pruning procedure to avoid overfitting [5].

A tree method for time-to-event data is called survival tree. A comprehensive review of tree-structured methods for survival data with right censoring is provided, which introduces more than 20 survival tree methods developing from 1985 up to 2008 and some ensemble methods with survival tree [6]. Some survival trees utilize logrank statistic or a parametric likelihood ratio statistic as a splitting rule [7, 8]. Some tree methods adopt Wilcoxon-Gehan statistic and Kolmogorov-Smirnov statistic to measure the heterogeneity between nodes [9, 10]. However, the majority of tree-structured methods suffer from biased variable selection, which is induced by maximizing a splitting criterion over all possible splits simultaneously [11]. Conditional Inference Framework (CIF) is also a tree method and is not biased towards covariates with many values [12]. A permutation test framework [13] is applied in CIF to guarantee unbiased variable selection for full data or censored data flexibly. An unbiased survival tree is proposed for left-truncated and right-censored data based on CIF [14].

Some new survival tree methods are also shown in the ensemble methods, which are based on a sizable set of survival trees as base learners. Ishwaran et al. proposed Random Survival Forest (RSF) for right censored data, which is based on Random Forest (RF) [15] to predict the average estimator value of cumulative hazard function at nodes of survival trees [16]. An extension of the censoring unbiased transformations to general loss function is applied to construct some new algorithms of censoring unbiased trees and ensembles for right-censored data [17].

However, only a few tree-based methods for interval censoring are proposed. A survival tree method for interval-censored data proposed by Yin suffers from bias of splitting variable selection [18]. *ICtree* is proposed as an unbiased tree method [19], which embeds the log-rank score into CIF for interval-censored data [20]. The extremely randomized trees (ERT) combining with Wilcoxon rank sum test at each iteration are proposed in a recursive forest and this forest is suitable for Case I interval-censored data which is a special type for interval-censored data [21].

We propose an unbiased interval-censored tree by combining Generalized Log-Rank Tests (GLRT) with CIF. GLRT is a test procedure for survival comparison based on interval-censored failure time data, which is a generalization of Log-Rank Test, and has stronger test ability when appropriate parameter is selected [22]. We adopt several special statistics with parameters based on GLRT, and

derive a new tree method with hyper-parameter denoted by $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$. We also discuss other tree models based on other rank tests statistics. A comparison of predictive performance among $ICtree$, $ICS_1^{\rho,\gamma}tree$, $ICS_2^{\rho,\gamma}tree$ and other interval-censored trees is implemented.

$ICS_1^{\rho,\gamma}tree$ is a competitive interval-censored tree method benefitting from the properties of GLRT and hyper-parameter $\rho, \gamma$. Appropriate hyper-parameter values taken in $ICS_1^{\rho,\gamma}tree$ improve the performance of prediction. $ICS^{\rho,\gamma}tree$ has more extensive adaptability to data because of choice of hyper-parameter. In Section 2, we first review CIF and GLRT, and then we introduce new interval-censored tree methods with hyper-parameter. The methods combine some new test statistics of GLRT or other rank tests with CIF. An extensive simulation is executed in Section 3, and hyper-parameter tuning is implemented by a grid searching cross-validation technique. The predictive accuracy of the new tree methods are evaluated. In Section 4, we apply new tree methods to analysis a tooth emergence data respectively [23].

## 2. New interval-censored trees

### 2.1. Overview of conditional inference frame

Let $X = (X_1, X_2, \ldots, X_m)^T$ denote a $m$-dimensional vector of covariates and let $Y$ denote a response variable. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_m$ be the sample space of $X$. Let $\mathcal{Y}$ be the sample space of response variable $Y$. Let $\mathcal{L}_n = \mathcal{Y} \times \mathcal{X}$ be the sample space of $n$ iid observations, specifically, $\mathcal{L}_n = \{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T; i = 1, 2, \ldots, n\}$.

CIF is a two-step tree method, because it separates the process of variable selection and splitting, thus guarantees the unbiasedness of variable selection.

The two steps are as follows:

- Step 1 is variable selection. The association between response $Y$ and covariates $X$ is measured based on $p$ value according to pre-specified significant level, and thus the component $X_{j_0}$ of $X$ with strongest association with response $Y$ as splitting variable is selected.
- Step 2 is searching for optimal splitting point by two-sample test based on the selected variable $X_{j_0}$.

In step 1, under global null hypothesis of independence between any component $X_j$ of covariates $X$ and the response $Y$, a test statistics $T_j(\mathcal{L}_n, \omega)$ is constructed, $j = 1, 2 \cdots, m$.

$$T_j(\mathcal{L}_n, \omega) = vec(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i, (Y_1, \ldots, Y_n))^T) \in \mathbb{R}^{b_j q}, \tag{2.1}$$

where $g_j : \mathcal{X}_j \to \mathbb{R}^{b_j}$ is a transformation of the covariate $X_j$, $j = 1, 2 \cdots, m$, influence function $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ depends on all the observation values of responses variable in a permutation symmetric way. 'vec' operator converts a $b_j \times q$ matrix into a $b_j q$ column vector by column-wise combination. Here $\omega = (w_1, \ldots, w_n)^T$ is a case weight vector presenting a node of the tree. $w_i$ is one when the $i$th observation fall into the node, otherwise is zero.

An extended permutation test is applied to calculate the $p$ value denoted by $p^j$ based on test statistics $T_j(\mathcal{L}_n, \omega)$ [13], and $X_{j_0}$ is selected as splitting variable if the $p^{j_0}$ is the smallest among $p^j, j = 1, 2 \cdots, m$.

In step 2, a two-sample test is executed based on the selected splitting variable $X_{j_0}$ and the test statistic $T_{j_0}^{\mathcal{A}}(\mathcal{L}_n, \omega)$. Here

$$T_{j_0}^{\mathcal{A}}(\mathcal{L}_n, \omega) = vec(\sum_{i=1}^{n} w_i I(X_{j_0 i} \in \mathcal{A}) h(Y_i, (Y_1, \ldots, Y_n))^T) \in \mathbb{R}^q \tag{2.2}$$

where $\mathcal{A}$ is an arbitrary subset taken from the sample space $\mathcal{X}_{j_0}$ and $I(\cdot)$ is the indicator function.

$T_{j_0}^{\mathcal{A}}(\mathcal{L}_n, \omega)$ is a special case of test statistic (2.1) and is used to measure the discrepancy between the samples $\{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T \in \mathcal{L}_n | X_{j_0 i} \in \mathcal{A}; i = 1, 2, \ldots, n\}$ and $\{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T \in \mathcal{L}_n | X_{j_0 i} \notin \mathcal{A}; i = 1, 2, \ldots, n\}$ in one node.

The optimal splitting $\mathcal{A}^0$ is found by exhaustive search, which is the subset maximizing the test statistic of two-sample test over all possible subsets $\mathcal{A}$ of the sample space $\mathcal{X}_{j_0}$.

The extended permutation test, two-sample test and type of data determine the way how $g_j$ and influence function $h$ are chosen, and CIF is efficiently applicable to full data and right-censored data when it takes specific $g_j$ and influence function $h$. Later we will discuss in detail the form of $g_j$ and $h$ taken for interval censored data in Subsection 2.3.

### *2.2. Overview of generalized log-rank test*

Consider $n$ independent subjects from $k$ different populations. Let $n_l$ denote the number of subjects from population $l$ with survival function $S_l$, $l = 1, 2, \ldots, k$, and let $T_i$ be survival time at which the event of interest for subject $i$ happens, $i = 1, 2, \ldots, n$, where $\sum_{l=1}^{k} n_l = n$. In addition, let $x_i$ be the $k$-dimensional indicator vector related to the subject $i$ whose $lth$ element is 1 if it belongs to population $l$, and 0 otherwise. $(L_i, R_i]$ is the censoring interval and satisfies

$$(L_i, R_i] = \begin{cases} (0, U_i] & T_i \le U_i \\ (U_i, V_i] & U_i < T_i \le V_i \\ (V_i, \infty) & T_i > V_i \end{cases} \tag{2.3}$$

where $U_i$ and $V_i$ are non-negative random variables independent of $T_i$ such that $U_i < V_i$, i.e., we assume non-informative interval censoring. Right-censoring occurs when $R_i = +\infty$. Without loss of generality, let $k = 2$ in our tree methods.

The generalized log-rank test (GLRT) is applied to interval-censored failure time data for comparison of survival function. The test statistic is $U_\xi = \sum_{i=1}^{n} x_i U_{\xi,i}$ where score statistic

$$U_{\xi,i} = \frac{\xi(\widehat{S}(L_i)) - \xi(\widehat{S}(R_i))}{\widehat{S}(L_i) - \widehat{S}(R_i)}. \tag{2.4}$$

Here $\xi(x)$ is a function over the interval (0,1) that satisfies the condition

$$\lim_{x \to 0} 1 - \xi(1 - x) = \lim_{x \to 1} 1 - \xi(1 - x) = c_0, \tag{2.5}$$

where $c_0$ is a constant. This is one of sufficient conditions for asymptotic distribution of $U_{\xi,i}$. $\widehat{S}$ is a non-parametric maximum likelihood estimator (NPMLE) of survival function, and is implemented by the improved EMICM algorithm [24, 25], which is an extended form of the early EMICM algorithm [26].

Different score statistic $U_{\xi,i}$ can be obtained by using different functions $\xi(x)$ according to (2.4). For example, $U_{\xi,i}$ is exact the logrank score statistic when $\xi(x) = xlogx$. A specific $\xi(x) = (logx)x^{\rho+1}(1-x)^{\gamma}$ with parameters $\rho$ and $\gamma$ as an example was given in [22], which inspires us to construct other new $\xi(x)$ with parameters $\rho$ and $\gamma$ to derive a new interval-censored tree with hyper-parameter.

### *2.3. A new interval-censored tree based on generalized log-rank test*

Our idea about new interval-censored tree stems from more extensive test ability of GLRT with parameters. We assign some new score statistics shown in (2.4) derived from newly constructed $\xi(x)$ to the influence function $h$ in test statistic (2.1) and (2.2).

Firstly, $\xi(x) = (logx)x^{\rho+1}(1-x)^{\gamma}$ is considered to obtain a new score statistic $U_{\xi,i}$ with parameters $\rho, \gamma$ according to (2.4), which is denoted by $U^{\rho,\gamma}_{\xi_1,i}$.

Next we construct a new $\xi_2(x) = (tan(\frac{\pi}{2}(x-1)))x^{\rho+1}(1-x)^{\gamma}$ and assign it to $U_{\xi,i}$, and then a new score statistic referred as $U^{\rho,\gamma}_{\xi_2,i}$ is derived.

Then we combine the new score statistics with the CIF to construct our interval-censored tree methods.

Considering CIF introduced in Subsection 2.1, let $Y_i^T = (L_i, R_i)$ where $L_i$ and $R_i$ are the endpoints of censoring interval specified in (2.3), and $Y_i$ is a bivariate response variable. Assume the covariates are numeric. In step 1 of CIF, let $g_j(x) = x$ where $x \in \mathcal{X}_j$, and then $\mathbb{R}^{b_j} = \mathbb{R}$. Let $h = U^{\rho,\gamma}_{\xi_1,i}$ and then $\mathbb{R}^q = \mathbb{R}$. Thus the test statistic

$$T_j(\mathcal{L}_n, \omega) = \sum_{i=1}^{n} w_i X_{ji} U^{\rho,\gamma}_{\xi_1,i} \in \mathbb{R}, \tag{2.6}$$

is used for extended permutation test.

In step 2 of CIF, let $g_{j_0}(x) = I(x \in \mathcal{A})$ where $x \in \mathcal{X}_{j_0}$ and $\mathcal{A} \subset \mathcal{X}_{j_0}$, then $\mathbb{R}^{b_j} = \mathbb{R}$. Let $h = U^{\rho,\gamma}_{\xi_1,i}$ and then $\mathbb{R}^q = \mathbb{R}$. Thus the test statistic

$$T^{\mathcal{A}}_{j_0}(\mathcal{L}_n, \omega) = \sum_{i=1}^{n} w_i I(X_{j_0 i} \in \mathcal{A}) U^{\rho,\gamma}_{\xi_1,i} \in \mathbb{R} \tag{2.7}$$

is used for two-sample test.

Then the new interval-censored tree is derived, which is denoted by $ICS_1^{\rho,\gamma}tree$. In particular, the $ICtree$ is a special case of $ICS_1^{\rho,\gamma}tree$ when $\rho = \gamma = 0$.

Similarly, we assign the $U^{\rho,\gamma}_{\xi_2,i}$ to influence function $h$, and propose another new tree with hyper-parameters $\rho$ and $\gamma$, which is called $ICS_2^{\rho,\gamma}tree$.

The Algorithm 1 explains the $ICS_1^{\rho,\gamma}tree$, and the setting of hyper-parameters is discussed in the Algorithm 2.

**Algorithm 1** $ICS_1^{\rho,\gamma}tree$

**Given:**

- $(\rho, \gamma)$: selected values of hyper-parameter.
- $X = (X_1, X_2, \ldots, X_m)^T$: $m$ dimensional covariates.
- $Y$: response variable.

**Procedure:**

For each node:

I/ Testing the global null hypothesis of independence between any component of the covariates $X$ and the response $Y$ by permutation test. The test statistic (2.6) is associated with $(\rho, \gamma)$.

(a) If the global null hypothesis can be rejected according to a pre-specified nominal level $\alpha$, then $X_{j_0}$ satisfying the $j_0$th component of covariate $X$ with the strongest association to $Y$ is selected as the splitting covariate, i.e., the index $j_0$ satisfies that the $p$ value is the smallest.

(b) If the global null hypothesis can not be rejected, then stop.

II/ Considering $X_{j_0}$ which is selected splitting variable in (i), two-sample test is utilized to search the optimal binary split $\{X_{j_0} \in \mathcal{A}\}$ and $\{X_{j_0} \notin \mathcal{A}\}$ where the null hypothesis is the two samples are statistically equivalent. Exhaustive search is executed for every allowable split, and $\mathcal{A}^0$ satisfies that the test statistic (2.7) is maximum among all the allowable subset $\mathcal{A} \subset \mathcal{X}_{j_0}$.

The step I and step II are reapplied to each of children nodes.

We also utilize other rank test as influence function to explore the predictive performance of $ICS_1^{\rho,\gamma}tree$ $ICS_2^{\rho,\gamma}tree$. Our improved tree methods are based on CIF, so the time complexity is the same with CIF.

### *2.4. Interval-censored tree methods based on other rank tests*

$G^{\rho,\gamma}$ test is a class of rank test for interval-censored data to evaluate whether the survival function under each group is equivalent or not [27]. The test statistic is $U_G = \sum_{i=1}^{n} x_i c_{\rho,\gamma,i}$, where

$$c_{\rho,\gamma,i} = \frac{\widehat{S}(L_i)B(1-\widehat{S}(L_i);\gamma+1,\rho) - \widehat{S}(R_i)B(1-\widehat{S}(R_i);\gamma+1,\rho)}{\widehat{S}(L_i) - \widehat{S}(R_i)} \tag{2.8}$$

and the incomplete beta function $B(\cdot)$ satisfies

$$-B(1-t,\gamma+1,\rho) = -\int_0^{1-t} x^{\gamma}(1-x)^{\rho-1}dx = \lambda(t).$$

Here $n$, $L_i$, $R_i$, $\widehat{S}$ and $x_i$ have the same meaning as Section 2.2. $G^{\rho,\gamma}$ test coincides with the log-rank test when $\rho = \gamma = 0$ and $\lambda(t) = log(t)t^{\rho}(1-t)^{\gamma}$. Then tree method denoted $ICG^{\rho,\gamma}tree$ is obtained when $c_{\rho,\gamma,i}$ is considered as the influence function under CIF.

A generalized Wilcoxon test for interval-censored data [28] is also applied to CIF, and

$$U_{i,W} = \widehat{S}(L_i) + \widehat{S}(R_i) - 1, \tag{2.9}$$

is assigned to the influence function, where $U_{i,W}$ is the rank score of generalized Wilcoxon test and is utilized for comparison of survival function of two groups. We refer to this tree as *ICWtree*.

### *2.5. Hyper-parameter tuning*

It is important to set an appropriate hyper-parameter value to adapt our tree methods to the data sets. We use a grid search cross validation technique [29] to explore the hyper-parameter optimization. The whole hyper-parameter tuning procedure is shown in Algorithm 2.

**Algorithm 2** Hyper-parameter tuning

**Given:**

- $A$: training set.
- $N_A$: size of training set $A$.
- $[a_\rho, b_\rho] \times [a_\gamma, b_\gamma]$: two-dimensional search space of hyper-parameter $(\rho, \gamma)$.
- $\triangle s_\rho$: step size in direction of $\rho$.
- $\triangle s_\gamma$: step size in direction of $\gamma$.
- $K$: the fold number of cross validation.
- $\phi(\cdot)$ is a measure for evaluation of the predictive performance.

**Procedure:**

For given two-dimensional search space $[a_\rho, b_\rho] \times [a_\gamma, b_\gamma]$,

- $[a_\rho, b_\rho]$ is divided into $B_1$ subinterval $[\rho_{i-1}, \rho_i]$ of equal width $\triangle s_\rho = (b_\rho - a_\rho)/B_1$, $i = 1, \cdots, B_1$, and $\rho_0 = a_\rho$.
- $[a_\gamma, b_\gamma]$ is divided into $B_2$ subinterval $[\gamma_{j-1}, \gamma_j]$ of equath width $\triangle s_\gamma = (b_\gamma - a_\gamma)/B_2$, $j = 1, \cdots, B_2$, and $\gamma_0 = a_\gamma$.
  - For $i$=1 to $B_1$, $j$=1 to $B_2$

    I/. Split the training set $A$ into $K$ non-overlapping groups randomly with the approximately same size, the $k$th group is considered as a internal test set $A_k^{test}$, then the remaining parts together are considered as a internal training set $A_k^{train}$, $k = 1, \cdots, K$.

    II/. $\phi(\cdot)$ is applied for evaluating predictive accuracy of a tree model with hyper-parameter $(\rho_i, \gamma_j)$ by $K$-fold cross validation, where the model is fitted in $A_k^{train}$ and evaluated in $A_k^{test}$, $k = 1, \cdots, K$. The outcome of $K$-fold cross validation is denoted by $\phi(\cdot)_{ij}$.

**Output**

The best hyper-parameter values $(\tilde{\rho}, \tilde{\gamma}) = (\rho_{i^*}, \gamma_{j^*})$, where $i^*, j^*$ satisfies that the value of $\{\phi(\cdot)_{i^* j^*}\}$ is optimal.

## 3. Simulation

Let $T_i$ be the failure time of interest for subject $i$, and suppose $(L_i, R_i]$ to be the censored interval to which $T_i$ belongs. Here $L_i = \tau_j$ and $R_i = \tau_{j+1}$ for some $j$, where $\tau_{j+1} > \tau_j$, $\tau_0 = 0$ and $\tau_l = +\infty$, $j = 1, \ldots, l-1$. The gap between adjacent endpoints of the interval is denoted by $len_j = \tau_j - \tau_{j-1}$.

We assume that censoring mechanism is non-informative, i.e., the failure time of interest is independent of the censoring intervals. $T_i$ is randomly generated from a specific distribution $F(t)$, $len_j$ is a random variable from a specified distribution $G(x)$ for each $j$. The type of interval-censoring is Case II. The censoring mechanism are similar to [19] for the purposes of comparison.

The observations are right-censored when the survival time $T$ fall into interval $(\tau_l, +\infty)$. The right censoring rate is controlled by adjusting the number of intervals. Three right-censoring rates are considered to explore the impact of right-censoring rate on the performance of the tree methods: 0%, 20% and 40%.

### *3.1. Evaluation of predictive accuracy for classification*

In real world, the distribution of lifetime or failure time is uncertain. A good model for classification can distinguish the observations with high hazard rates from the ones with low hazard rates, i.e., it can tell the distributions with one parameter value from the other parameter value, and correctly classify the samples into the corresponding categories from the a pool of distribution with different parameters. What's more, a survival tree method can select the important covariates [30]. To evaluate these classification performances of tree models, tree structured data sets are constructed which are synthetic data sets.

#### 3.1.1. Tree structured data

Tree structured data sets shown in Figure 1 are constructed which are easily divided into several categories according to the character of covariates.

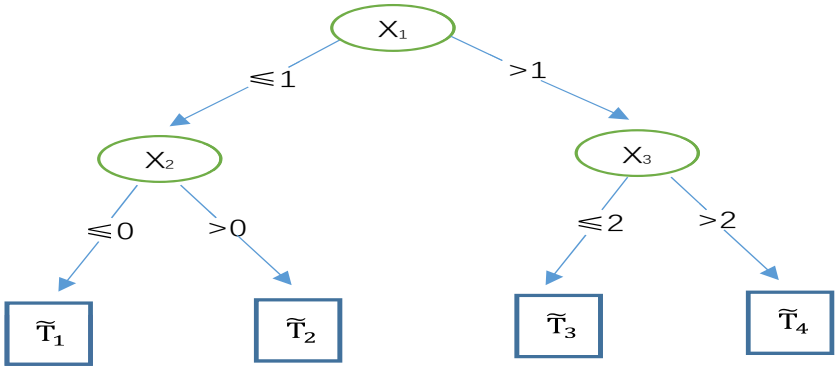



**Figure 1.** Tree structured data.

Covariates $X = \{X_1, \ldots, X_6\}^T$ is a 6 dimensional vector, and

- $X_1$ is from discrete uniform distribution $\{1, 2, 3, 4\}$;
- $X_5$ is from discrete uniform distribution $\{1, 2, 3, 4, 5\}$;

- $X_2$ and $X_6$ follow the Bernoulli distribution with $p = 0.5$;
- $X_3$ and $X_4$ are uniform in the interval $[1, 3]$.

The failure time of interest $T$ is generated from a distribution $F(t)$ determined by the values of $X_1, X_2, X_3$ according to the relationship of structural tree shown in Figure 1. More specifically, the distribution $F(t)$ involves in 4 parameters, thus $T$ is denoted by $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{T}_3$ and $\tilde{T}_4$ respectively according to the value of parameter chosen by $F(t)$, which are labelled by 4 categories. We consider four data scenarios that vary in terms of the distribution $F(t)$ as follows.

1. Exponential distribution with parameter $\lambda \in \{0.08, 0.25, 0.6, 0.95\}$;

2. Weibull distribution with shape parameter $\alpha$=7 and scale parameter $\lambda \in \{1, 6, 14, 30\}$, which have an increasing failure rate with time, we refer to this case as Weibull1;

3. Weibull distribution with shape parameter $\alpha$=0.8 and scale parameter $\lambda \in \{1, 6, 14, 30\}$, which have a decreasing failure rate with time, this case is denoted by Weibull2;

4. Log-normal distribution with parameter pairs $(\mu, \sigma)$ which is equal to $(0, 0.25)$, $(2, 0.25)$, $(3, 0.4)$, $(4, 0.1)$.

The distribution $G(t)$ of the interval length $len_j$ is [0.7,1.5] uniform distribution. Let the sample size be $N$, and the $i$th observation is denoted by $O_i = (L_i, R_i, X_1, X_2, X_3, X_4, X_5, X_6)^T \in \mathcal{L}_n$, $i = 1, \ldots, N$, where $(L_i, R_i) = Y_i$ is response variable.

Here we take the first data scenario in simulation as an example to illustrate the simulative data specifically. $\tilde{T}_1$,$\tilde{T}_2$,$\tilde{T}_3$,$\tilde{T}_4$ are from Exponential distribution with parameter $\lambda \in \{0.08, 0.25, 0.6, 0.95\}$ respectively. As we known, $\lambda$ is the constant hazard function $h(t) = \lambda$. So the simulative data has four classes which have different level of hazard rates, Class1, Class2, Class3 and Class4, The survival function of each class is $S(t) = exp(-\lambda t)$, $\lambda > 0, t > 0$. Specifically, the samples belonging to the first class (Class1) has the lowest hazard rate and the survival function is $S(t) = exp(-0.08t)$; the ones belonging to the fourth class (Class4) has the highest hazard rate, and the survival function is $S(t) = exp(-0.95t)$. The categories are associated with the values of covariate, so the class an observation should belong to is unique.

Obviously, a perfect tree method based on this simulative data should grow a recursive structure shown in Figure 2. The terminal nodes ID are 3, 4, 6 and 7 respective, and they correspond 4 classes. Each observation will fall down a terminal node of the perfect tree method according to the values of the covariate of samples and recursive structure of a tree. The terminal node ID to which an observation falls down is unique. Hence the category of each observation is labelled by node ID of perfect tree method .

For example, if an observation satisfies $X_1 = 1, X_2 = 1, X_3 = 2.3, X_4 = 1, X_5 = 5, X_6 = 0$, then the observation should be assigned to terminal node ID 4, and the category of it is Class2.

```
[[1]]
[1]root
|    [2] X1 <= 1
|    |      [3] X2 <= 0*
|    |      [4] X2 > 0*
|    [5] X1 > 1
|    |      [6] X3 <= 2*
|    |      [7] X3   > 2*
```

**Figure 2.** Recursive structure of data set.

### 3.1.2. Measure for evaluation of classification accuracy

Let $N^*(O_i)$ be the terminal node ID into which $O_i$ should fall according to Figure 2, $i = 1, \ldots, N$, $N$ is sample size. Whereas let $N^{**}(O_i)$ be the terminal node ID in which the observation $O_i$ falls according to various tree methods, $i = 1, \ldots, N$. If $N^*(O_i) = N^{**}(O_i)$, then the observation $O_i$ is correctly classified by the tree methods, $i = 1, \ldots, N$.

Now we define a performance measure $cr$ for classification task to calculate the ratio of the observations falling into a terminal node correctly, i.e., the ratio of observations labelled correctly.

$$cr = \frac{\sum_{i=1}^{N} I_{\{N^*(O_i)=N^{**}(O_i)\}}}{N},$$

where $I_A$ is an indicator function defined on set A. $cr$ describes the ratio that observations (e.g., patients) that have high (low) hazard rates are correctly classified into the corresponding class with high (low) hazard rate by a classifier ( e.g., a tree method).

$cr$ is a reasonable measure to evaluate the correctness of a tree method covering the tree-structured data sets, because the correctness of the tree structure generated by tree methods are closely related to it's value. If a tree method grows a tree structure with wrong splitting based on tree structured data, like the case shown in Figures 3 and 4, some observations will be misclassified. This in turn leads to lower value of $cr$. For example, if an observation with covariates $X_1$=2, $X_3$=1.7 should belong to the third category (Class3) and fall down terminal node 6 according to data structure shown in Figure 2. But a splitting point of the tree method (Figure 3) is inconsistent with the character of data, so this observation is incorrectly assigned to terminal node 7. And even worse in some cases that no observation has any chance to fall down the correct terminal nodes due to the wrong tree structure grown by a tree method. A very wrong tree structure grown by a tree method shown in Figure 4 has no terminal node 6 and 7, and the observations which should belong to a terminal node 6 and 7 fail to fall there, then $cr$ is much lower. Thus the more wrong splitting by tree methods, the lower $cr$ is, and $cr$ is a reasonable measure to evaluate the predictive accuracy for classification.

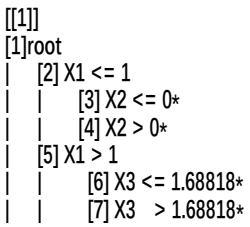



**Figure 3.** Wrong splitting structure grown by a tree method.

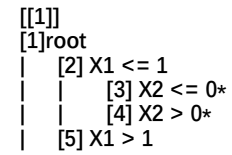



**Figure 4.** Wrong splitting structure grown by a tree method.

3.1.3. Hyper-parameter tuning

The hyper-parameter tuning is executed to explore the appropriate hyper-parameter values of tree methods for tree-structured data according to Algorithm 2. Let sample size $N$ be 1200, where the ratio of training set to test set is 5:1, i.e., the size of training set $A$ is 1000 and the size of test set $D$ is 200. The training set $A$ is used for tuning $(\rho, \gamma)$ and test set $D$ is for evaluation, which means the process of choice of hyper-parameter and evaluation is separated.

Let two-dimensional search space of hyper-parameter $(\rho, \gamma)$ be $(0.01, 0.81) \times (0.01, 0.81)$ and $B_1 = B_2 = 40$. Here $cr$ is taken as measure $\phi(\cdot)$, and 5-fold cross validation is applied to calculate $\phi(\cdot)_{ij}$. The optimal hyper-parameter is

$$(\tilde{\rho}, \tilde{\gamma}) = \arg\max_{(\rho,\gamma)} \{\phi(\cdot)_{ij}\}$$

Consider the first scenario of tree structured data where $T$ is from Exponential distribution with different right censoring rates. The optimal hyper-parameter values chosen by Algorithm 2 are partially shown in Table 1. We also discuss the optimal values of hyper-parameter in other scenarios of data sets where $T$ is from Weibull1 distribution, Weibull2 distribution and Log-normal distribution, The results are listed in Tables 9–11 in Appendix.

**Table 1.** The hyper-parameter tuning for tree-structured data under Exponential distribution.

| Type of Trees | 0% | 20% | 40% |
|---|---|---|---|
| | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ |
| $ICS_1^{\rho,\gamma}tree$ | (0.03,0.09) | (0.13,0.05) | (0.05,0.03),(0.05,0.21) |
| $ICS_2^{\rho,\gamma}tree$ | (0.69,0.55) | (0.69,0.49) | (0.03,0.03 ) |

0%,20% and 40% denote the right censoring rate of tree-structured data.

The optimal values of hyper-parameter depend on the data and the tree models, right censoring rate also affects the choice of hyper-parameter values. We find in most cases the optimal hyper-parameter values are around (0.05,0.05), (0.5,0.03), (0.05,0.21) for $ICS_1^{\rho,\gamma}tree$. While for $ICS_2^{\rho,\gamma}tree$, they are around (0.21,0.01), (0.69,0.03), (0.69,0.51) and some other values. The value of $cr$ for $ICG^{\rho,\gamma}tree$ is much lower than the one of $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$, and thus selections of hyper-parameter values are not concerned for $ICG^{\rho,\gamma}tree$.

### 3.1.4. Evaluation of predictive accuracy for classification

The predictive performance of various tree models are evaluated for classification. The mean of $cr$ from 1000 trials denoted by $\overline{cr}$ is calculated for evaluation on the test set $D$. Different values of the hyper-parameters are considered for our tree models. Especially, the values around (0.05,0.05), (0.5,0.03) and (0.5,0.21) are almost adopted for $ICS_1^{\rho,\gamma}tree$ under various distributions and they are not always the optimal hyper-parameter values but just as recommended values.

Tables 2–5 show the predictive performance. $ICS_1^{\rho,\gamma}tree$ almost performs better than $ICtree$ under all scenarios.

According to the result of Tables 2–5, $(\rho, \gamma) = (0.05, 0.03)$ is recommended for the interval-censored data of which the right censoring rate is lower than 40%, and $(\rho, \gamma) = (0.5, 0.21)$ is appropriate for the data with over 40% right censoring rate. Although $(\rho, \gamma) = (0.05, 0.03)$ and $(\rho, \gamma) = (0.05, 0.21)$ are not always the best hyper-parameter values for different data, the values of $(\rho, \gamma)$ around them make $ICS_1^{\rho,\gamma}tree$ well adaptive to the data and perform well. $ICS_2^{\rho,\gamma}tree$ performs better in particular type data than $ICS_1^{\rho,\gamma}tree$ and $ICtree$. However, as a whole, it shows instability in tree-structured data. Performance of $ICWtree$ is seriously affected by the level of right censoring rate and the distribution type of $T$, and predictive accuracy sharply deteriorate in case of high right censoring rate.

**Table 2.** Predictive accuracy under Exponential distribution.

| Type of Trees | $Hyper-parameter$ | 0% | 20% | 40% |
|---|---|---|---|---|
| | $(\rho,\gamma)$ | $\overline{cr}$ | $\overline{cr}$ | $\overline{cr}$ |
| $ICtree$ | - | 0.5130 | 0.4416 | 0.2723 |
| $ICS_1^{\rho,\gamma}tree$ | (0.05, 0.03) | 0.5155 | 0.4466 | 0.2752 |
| | (0.05, 0.05) | 0.5145 | 0.4456 | 0.2746 |
| | (0.05,0.21) | 0.5137 | 0.4465 | 0.2761 |
| | (0.69,0.03) | 0.4023 | 0.4393 | 0.2457 |
| $ICWtree$ | - | 0.4974 | 0.069 | 0.018 |
| $ICS_2^{\rho,\gamma}tree$ | (0.69, 0.49) | 0.5073 | 0.4479 | 0.2650 |
| | (0.07, 0.05) | 0.5083 | 0.3458 | 0.2744 |
| | (0.71, 0.61) | 0.5083 | 0.4437 | 0.2669 |

$\overline{cr}$ is mean of $cr$ based on 1000 trials on test set. 0%,20% and 40% denote the right censoring rate of tree structured data. $(\rho,\gamma)$ denotes the value of hyper-parameter. The boxed values indicate the higher values of $\overline{cr}$ derived from our tree methods than the one from $ICtree$.

**Table 3.** Predictive accuracy under Weibull1 distribution.

| Type of Trees | $Hyperparameter$ | 0% | 20% | 40% |
|---|---|---|---|---|
| | $(\rho,\gamma)$ | $\overline{cr}$ | $\overline{cr}$ | $\overline{cr}$ |
| $ICtree$ | - | 0.9164 | 0.9059 | 0.9212 |
| $ICS_1^{\rho,\gamma}tree$ | (0.05,0.03) | 0.9165 | 0.9076 | 0.9192 |
| | (0.05,0.05) | 0.9175 | 0.9070 | 0.9211 |
| | (0.05,0.21) | 0.8916 | 0.8843 | 0.9248 |
| | (0.57,0.33) | 0.9225 | 0.9097 | 0.9152 |
| | (0.53,0.37) | 0.9160 | 0.9132 | 0.9169 |
| $ICWtree$ | - | 0.9202 | 0.9211 | 0.4210 |
| $ICS_2^{\rho,\gamma}tree$ | (0.21,0.01) | 0.9440 | 0.9413 | 0.9324 |
| | (0.29,0.21) | 0.9364 | 0.9337 | 0.9348 |

$\overline{cr}$ is mean of $cr$ based on 1000 trials on test set. 0%,20% and 40% denote the right censoring rate of tree structured data. $(\rho,\gamma)$ denotes the value of hyper-parameter. The boxed values indicate the higher values of $\overline{cr}$ derived from our tree methods than the one from $ICtree$.

**Table 4.** Predictive accuracy under Weibull2 distribution.

| Type of Trees | *Hyperparameter* | 0% | 20% | 40% |
|---|---|---|---|---|
| | $(\rho,\gamma)$ | $\overline{cr}$ | $\overline{cr}$ | $\overline{cr}$ |
| *ICtree* | - | 0.6358 | 0.5871 | 0.4997 |
| $ICS_1^{\rho,\gamma}tree$ | (0.05,0.03) | 0.6383 | 0.5881 | 0.4994 |
| | (0.05,0.05) | 0.6385 | 0.5884 | 0.5002 |
| | (0.05,0.21) | 0.6264 | 0.5827 | 0.49325 |
| | (0.09,0.09) | 0.6397 | 0.5874 | 0.4993 |
| | (0.41,0.01) | 0.5972 | 0.5739 | 0.5029 |
| *ICWtree* | - | 0.5810 | 0.4352 | 0.3175 |
| $ICS_2^{\rho,\gamma}tree$ | (0.61, 0.45) | 0.6276 | 0.5812 | 0.4973 |
| | (0.41, 0.17) | 0.5963 | 0.5972 | 0.5074 |
| | (0.21, 0.01) | 0.1433 | 0.3366 | 0.5101 |

$\overline{cr}$ is mean of $cr$ based on 1000 trials on test set. 0%,20% and 40% denote the right censoring rate of tree structured data. $(\rho,\gamma)$ denotes the value of hyper-parameter. The boxed values indicate the higher values of $\overline{cr}$ derived from our tree methods than the one from *ICtree*.

**Table 5.** Predictive accuracy under Log-Normal distribution.

| Type of Trees | *Hyperparameter* | 0% | 20% | 40% |
|---|---|---|---|---|
| | $(\rho,\gamma)$ | $\overline{cr}$ | $\overline{cr}$ | $\overline{cr}$ |
| *ICtree* | - | 0.9181 | 0.9215 | 0.9265 |
| $ICS_1^{\rho,\gamma}tree$ | (0.05,0.03) | 0.9232 | 0.9235 | 0.9291 |
| | (0.05,0.05) | 0.9200 | 0.9193 | 0.9267 |
| | (0.05,0.21) | 0.6264 | 0.8965 | 0.9275 |
| | (0.17,0.01) | 0.9234 | 0.9196 | 0.9239 |
| *ICWtree* | - | 0.9294 | 0.9335 | 0.4126 |
| $ICS_2^{\rho,\gamma}tree$ | (0.21, 0.01) | 0.5514 | 0.9510 | 0.9448 |
| | (0.29, 0.37) | 0.5061 | 0.8593 | 0.9395 |
| | (0.41, 0.01) | 0.9495 | 0.9404 | 0.9304 |
| | (0.71, 0.61) | 0.9219 | 0.9188 | 0.9267 |

$\overline{cr}$ is mean of $cr$ based on 1000 trials on test set. 0%,20% and 40% denote the right censoring rate of tree structured data. $(\rho,\gamma)$ denotes the value of hyper-parameter. The boxed values indicate the higher values of $\overline{cr}$ derived from our tree methods than the one from *ICtree*.

All in all, $ICS_1^{\rho,\gamma}tree$ slightly improves the predictive accuracy with appropriate values of hyper-parameter. Although the algorithm 2 searches out the optimal values of hyper-parameter for tree methods depending on data, later the $(\rho, \gamma) = (0.05, 0.03)$ and $(\rho, \gamma) = (0.05, 0.21)$ as recommended values are assigned to hyper-parameter and parameter tuning can be omitted to save time cost, and the right censoring rate should be considered for the choice of hyper-parameter values.

### *3.2. Comparison of predictive performance for regression*

The synthetic tree-structured data is constructed to explore the performance of classification of tree models in Subsection 3.1. Next we discuss the predictive performance of tree models for survival time, and consider more general data structure for simulation.

#### 3.2.1. Complex structure data

A simulation setup of complex structure data is executed to evaluate the predictive performance of the new survival trees with hyper-parameter, The setup is similar to [19] for comparison. The dimension of the covariates are six, and $X_1, \ldots, X_6$ are independent, where

- $X_1$ is from discrete uniform distribution { -1, 0};
- $X_2$ is uniform in the interval [-0.5,0];
- $X_3$ and $X_5$ is from Bernoulli distribution with p = 0.5;
- $X_4$ and $X_6$ are uniform in the interval [0,1].

The distribution $F(t)$ of survival time $T$ is

1) Exponential with parameter $\lambda = e^{\vartheta}$;
2) Weibull with increasing hazard, scale parameter $\lambda = 4e^{\vartheta}$ and shape parameter $\alpha = 7$(denoted as Weibull1);
3) Weibull with decreasing hazard, scale parameter $\lambda = 4e^{\vartheta}$ and shape parameter $\alpha = 0.7$(denoted as Weibull2);

Here the value of location parameter $\vartheta$ depends on the covariates $X_1$ and $X_2$. Specifically,

$$\vartheta = -sin[(X_1 + 2X_2) \cdot \pi] + \frac{3}{2}\sqrt{-X_1 - 2X_2}.$$

The censoring mechanism is introduced at the beginning of this section. The gap of interval is randomly generated from $G(x)$ which is uniform [0.5,1], and different right-censoring rates are also considered.

#### 3.2.2. A measure of predictive accuracy for regression

Mean square error (MSE) of prediction has been ever considered as a measure to evaluate predictive accuracy of survival time, which is calculated by

$$\frac{1}{N}\sum_{i=1}^{N}(T_i - \hat{T}(X_i))^2.$$

where $N$ is the sample size, $T_i$ is the true failure time of interest of subject $i$ and $\hat{T}_i$ is predictive failure time for the $i$th subject, $i = 1, 2, \cdots, N$. However, the results given are almost inevitably inaccurate and unsatisfactory based on MSE [31].

Integrated Brier Score ($IBS$) is a popular method measuring average discrepancies between true status of event of interest and estimated predictive value, which was first proposed for right censoring data [31]. Later it is extended to the form

$$IBS = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{max(T)}\int_0^{max(T)}[\hat{S}_i(t) - I(T_i > t)]^2 dt, \tag{3.1}$$

for interval-censored data [32], where $T_i$ is the true failure time of interest of subject $i$, $\hat{S}_i(\cdot)$ is an estimator of survival function for the $i$th subject, $max(T) = max\{T_j\}_{j=1}^n$. Here $I(T_i > t)$ are calculated by the formula

$$I(T_i > t) = \frac{\hat{S}_i(t) - \hat{S}_i(R_i)}{\hat{S}_i(L_i) - \hat{S}_i(R_i)}$$

where $L_i < t \leq R_i$. Otherwise, $I(T_i > t) = 1$ when $t \leq L_i$, and $I(T_i > t) = 0$ when $t > R_i$.

Mean integrated squared error (MIE) is another measure for survival data [33], which is calculated by integrating the square difference between the two curves with respect to time and averaging over all observations

$$MIE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{max(T)}\int_0^{max(T)}[\hat{S}_i(t) - S_i(t)]^2 dt,$$

where $S_i(t)$ is true survival function for the $i$th subject. The meanings of $T_i$ and $\hat{S}_i(\cdot)$ are same with the ones in formula of IBS.

#### 3.2.3. Evaluation of predictive accuracy

The size of simulation data set is 1000 and the data set is divided into training set and test set by 4:1. Here our tree models and $ICtree$ grow in training set $A$ and then we evaluate their predictive performance in test data $D$.

The boxplots of $IBS$ are drawn by 1000 trials in test data $D$. The predictive methods are numbered in Table 6, here we just assign $(\rho, \gamma) = (0.05, 0.05)$ to $ICS_1^{\rho,\gamma}tree$, $(\rho, \gamma) = (0.69, 0.5)$ to $ICS_2^{\rho,\gamma}tree$, $(\rho, \gamma) = (0.01, 0.01)$ to $ICG^{\rho,\lambda}tree$ for comparison, these values are recommended in Subsection 3.1.4. In addition to $ICtree$, Proportional Hazards model for interval censored data (extended form of Cox model) denoted by $ICPH$ is also utilized for comparison. In order to evaluate the amount of information loss caused by interval-censoring, we also consider Conditional Inference Frame (CIF) and Proportional Hazards model using true failure time $T$, which are denoted by $Ttree$ and $TPH$ respectively.

**Table 6.** Numbering of the various predictive methods in Figure 5.

| No. | model | No. | model | No. | model | No. | model |
|---|---|---|---|---|---|---|---|
| 1 | $Ttree$ | 2 | $ICtree$ | 3 | $ICS_1^{\rho,\gamma}tree$ | 4 | $ICWtree$ |
| 5 | $ICG^{\rho,\gamma}tree$ | 6 | $ICS_2^{\rho,\gamma}tree$ | 7 | $TPH$ | 8 | $ICPH$ |

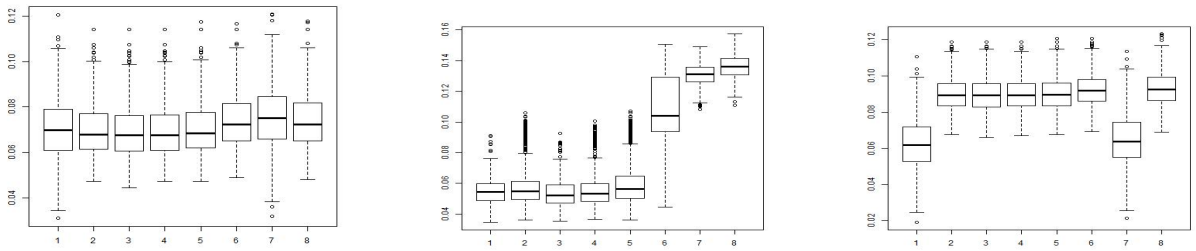
The results of comparison are shown in Figure 5. $ICS_1^{\rho,\gamma}tree$ has slightly better performance than $ICtree$ in this setting, especially for the Weibull1 distribution with increasing hazards. $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$ show similar performance which outperform $ICPH$, which demonstrates the flexibility of tree methods. $ICWtree$ and $ICG^{\rho,\gamma}tree$ seem to be more sensitive to right censoring data. Although there is no noticeable difference with $ICS^{\rho,\gamma}$ when the right-censoring rate is zero, the value of $IBS$ increases rapidly with increase of the right-censoring rate. Their performance are worse than $ICS^{\rho,\gamma}$'s with over 20% right-censoring rate. While the value of $IBS$ for $Ttree$ is much lower than tree methods because it has no informative loss from interval censoring.

MIS are also applied to evaluate the predictive performance of tree methods. The boxplots are drawn by 1000 trials to compare our tree methods to $ICtree$. The outcomes are shown in Figure 6, where the tree methods are specified in Table 7.

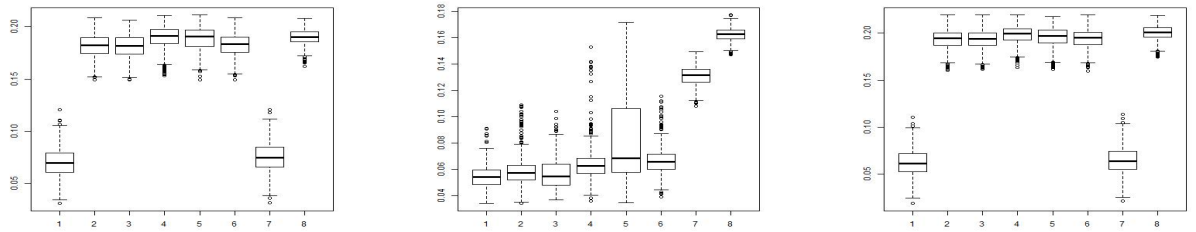
The results of the predictive performance evaluated by MIE validate that $ICS_1^{\rho,\gamma}tree$ has advantage in predictive accuracy than other tree methods mentioned in Table 7. Besides, the values of MIE increase with the increase of the right censoring rates of survival data, which means the higher right censoring rates lead to lower predictive accuracy. These results are consistent with those provided by $IBS$.

**Table 7.** Numbering of the various predictive methods in Figure 6.

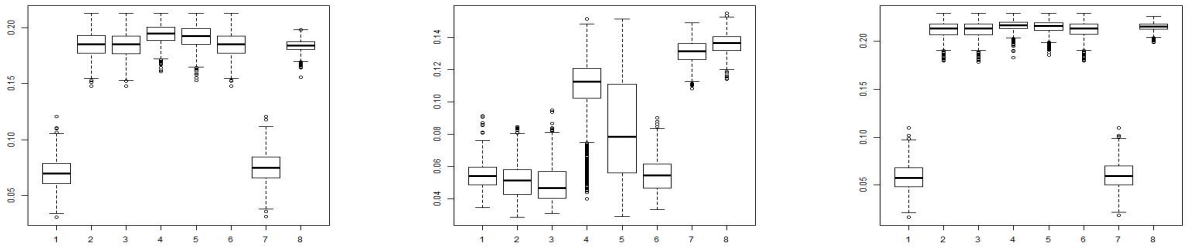
| No. | model | No. | model | No. | model |
|---|---|---|---|---|---|
| 1 | $ICtree$ | 2 | $ICS_1^{\rho,\gamma}tree$ | 3 | $ICWtree$ |
| 4 | $ICG^{\rho,\gamma}tree$ | 5 | $ICS_2^{\rho,\gamma}tree$ | | |



(a) Exponential (b) Weibull1 (c) Weibull2

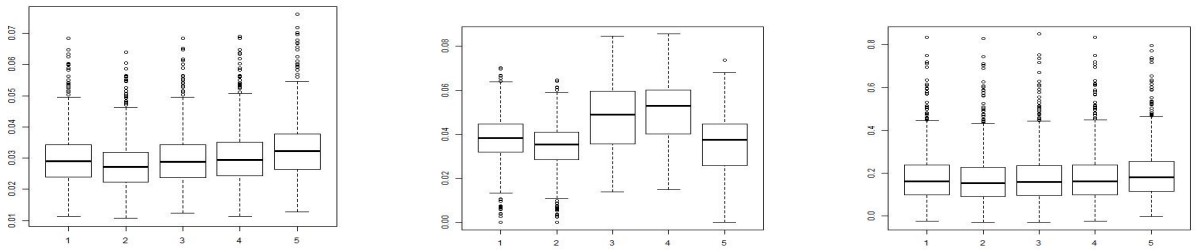


(d) Exponential (e) Weibull1 (f) Weibull2

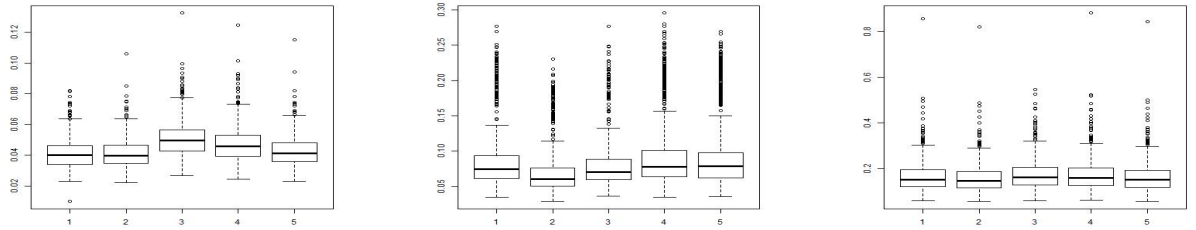


(g) Expoential (h) Weibull1 (i) Weibull2

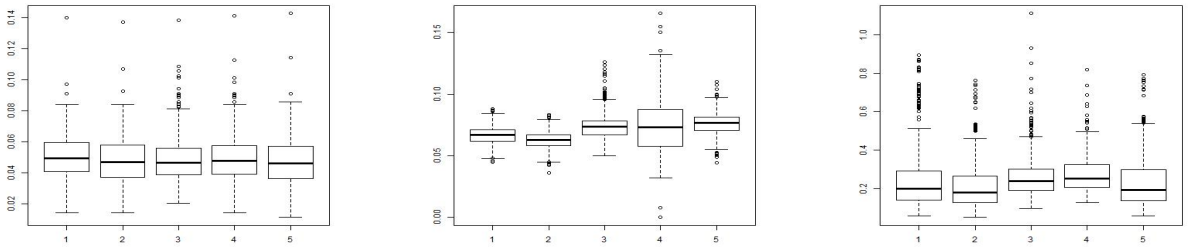
**Figure 5.** Settings: IBS boxplots with sample size $N = 200$ and $G(t)$=U[0.5,1]). First row corresponding to the result of 0% right censoring rate, second row corresponding to 20% right censoring rate and third row corresponding to 40% right censoring rate.



(a) Exponential (b) Weibull1 (c) Weibull2



(d) Exponential (e) Weibull1 (f) Weibull2



(g) Expoential (h) Weibull1 (i) Weibull2

**Figure 6.** Settings: MIE boxplots with sample size $N = 200$ and $G(t)$=U[0.5,1]). First row corresponding to the result of 0% right censoring rate, second row corresponding to 20% right censoring rate and third row corresponding to 40% right censoring rate.

## 4. Application

A real interval-censored data set is provided by the Signal $Tandmobiel^{®}$ study [23]. The study carried out a longitudinal survey of 4468 primary school pupils in Flanders (North of Belgium) about emergence times of permanent teeth, caries development from 1996 to 2001. The data set denoted by tandmob2 is available in R package bayesSurv containing the information on 28 teeth. The data set is interval-censored, because the event of interest in study is the emergence of the permanent teeth, and annual examinations for each child only record the interval that the emergence time belongs to rather than the exact time of emergence. We shift the time origin to 5 years of age, as proposed by [35] firstly. Covariates include

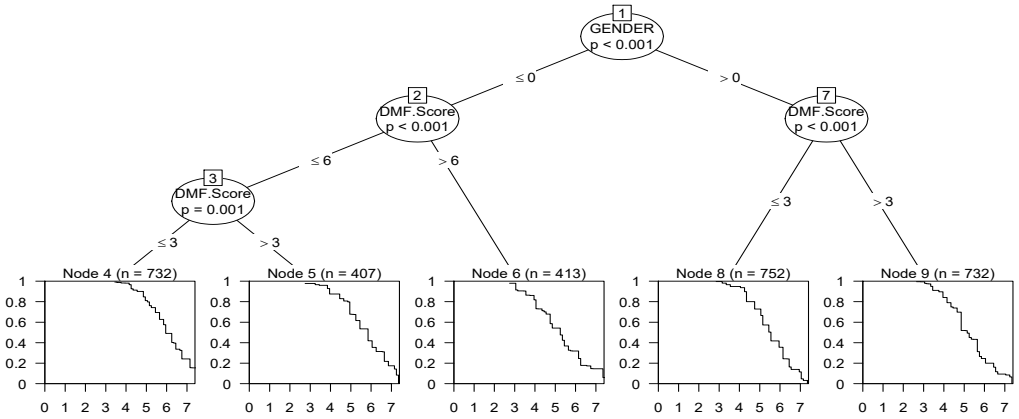
- geographical factor(the province of residence);
- gender;
- use of fluoride;
- type of education system;
- starting age of brushing the teeth;
- total number of deciduous teeth extracted due to orthodontic reasons denoted by BAD;
- total number of decayed,filled or missing deciduous teeth due to caries denoted by DMF.Score.

More details about the data set can be seen in [23, 34].

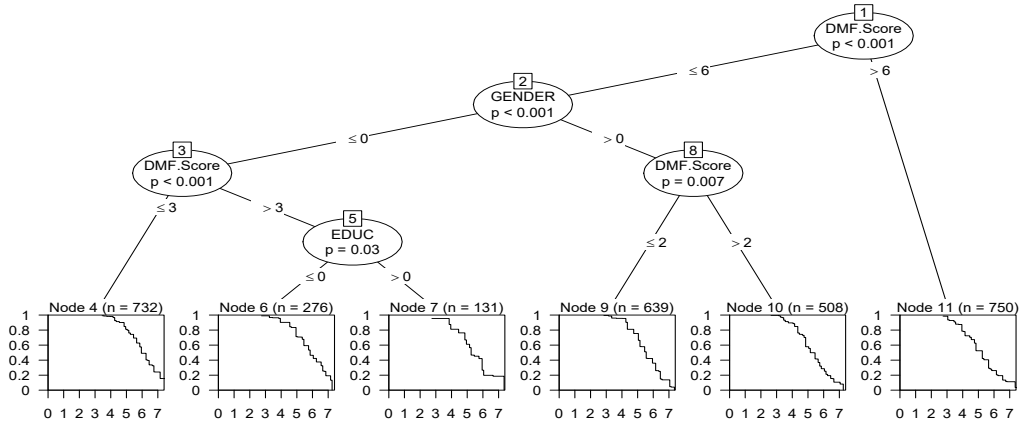
Figures 7–8 show the instances are classified into several categories by tree models, and the important covariates are selected for the survival time. We conclude that both gender and DMF.Score are more important than other covariates. $ICS_1^{0.05,0.05}$ show that gender is more important than DMF.Score, which is consist with *ICtree*. Whereas the gender is critical factor only when DMF.Score is less than or equal to 6 in the view of $ICS_1^{0.69,0.03}$ and $ICS_2^{0.69,0.03}$.

*ICWtree* grows with fewer splitting on DMF.Score in the right branch and terminal nodes are also fewer. As far as $ICG^{0.01,0.01}tree$ is concerned, DMF.Score=4 is the splitting rule of root node which is quite different from the others.

Based on the survival curves in terminal nodes in $ICS_1^{0.69,0.03}tree$ in Figure 7, we find the time to emergence of permanent tooth tends to be earlier for girls (value of Gender = 1) than boys (value of Gender = 0), and permanent tooth of the children with more total number of decayed or missing deciduous teeth due to caries (DMF.Score > 6) emerge earlier than others. This results are also applicable for the *ICWtree* and $ICG^{0.01,0.01}tree$ in Figure 8.

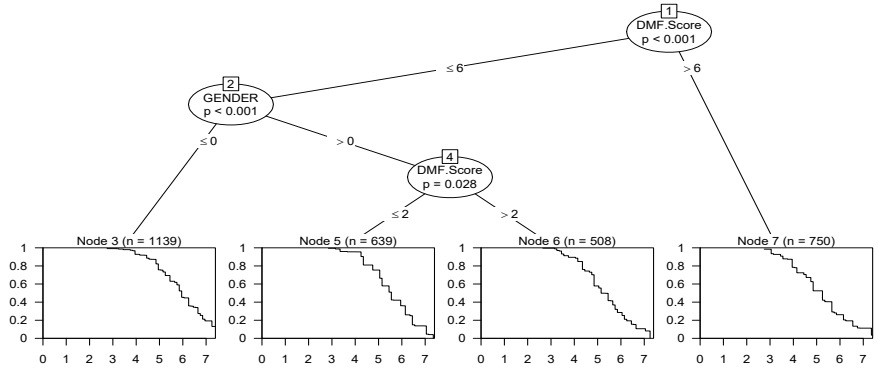



(a) $ICS_1^{0.05,0.05}tree$

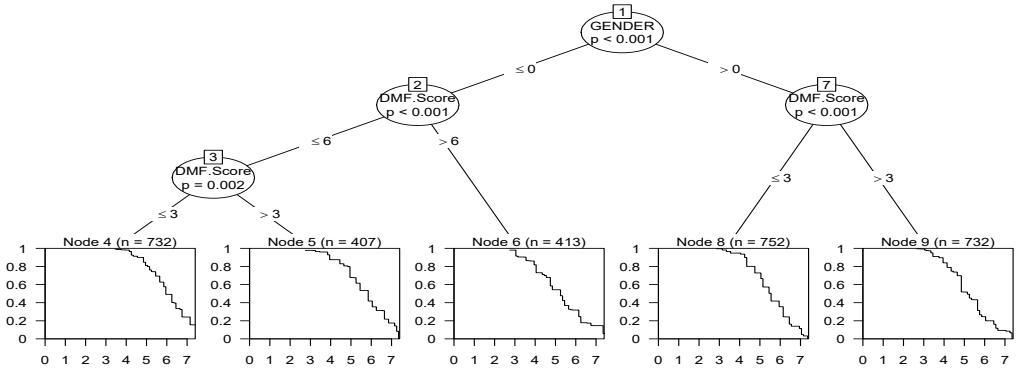



(b) $ICS_1^{0.69,0.03}tree$ and $ICS_2^{0.69,0.03}tree$

**Figure 7.** Interval-censored trees for the emergence of time in years of permanent first upper right premolar 14.




(a) *ICWtree*




(b) $ICG^{0.01,0.01}tree$

**Figure 8.** Interval-censored trees for the emergence of time in years of permanent first upper right premolar 14.

The time to emergence of permanent tooth is considered as survival time, and we predict them by tree methods. Here we show the results of permanent second premolar (tooth 15–45 in European dental notation), for the results of other teeth are similar. The sample size of data set N=3036. We split them into two parts by 2:1 as training set $A$ and test set $D$ respectively. Algorithm 2 is applied to the training set $A$ for hyper-parameter tuning with the measurement $\phi(\cdot) = d_{mean}$.

$$d_{mean} = \frac{\sum_{j=1}^{N_{out}} d_j}{N_{out}}$$

where $d_j$ denotes the distance between the predicted time $\tilde{T}_j$ for the $j$th observation and the end point of the observed interval $(L_j, R_j]$ which is closer to $\tilde{T}_j$ when $\tilde{T}_j$ falls out of $(L_j, R_j]$, $N_{out}$ is the number of the predicted time of permanent teeth emergence which fall outside of the corresponding observed interval [36]. In a word, $d_{mean}$ represents the average absolute distance away from the observed intervals when the predicted time fall outside of observed intervals. The smaller $d_{mean}$ is, the better the predictive performance is.

The optimal hyper-parameter values are around (0.5,0.03) and (0.05,0.21) for $ICS_1^{\rho,\gamma}tree$, (0.5,0.03) and (0.69,0.21) for $ICS_2^{\rho,\gamma}tree$, which are quite different from the values in simulations. A possible explanation is the right censoring rate is very high (over 58%), and larger than the one discussed in previous simulations. In this case (0.05,0.21) are selected, this is in line with the guidance that (0.05,0.21) is the recommended value when right censoring rate is over 40% in Subsection 3.1.4.

5-fold cross validation is executed in test set $D$ to evaluate the predictive performance of tree models with hyper-parameters. The outcome are reported in Table 8. The predictive accuracy is improved by $ICS_1^{\rho,\gamma}tree$. Even when the hyper-parameters is not always the optimal, our tree methods are also flexibly adaptive to data and have a good performance in prediction.

**Table 8.** Evaluation on permanent 2nd premolar data in Signal $Tandmobiel^{®}$ study.

| Models | *Hyperparameter* | 15 (58%) | 25 (57%) | 35 (59%) | 45 (59%) |
|---|---|---|---|---|---|
| | $(\rho, \gamma)$ | $\bar{d}_{mean}$ | $\bar{d}_{mean}$ | $\bar{d}_{mean}$ | $\bar{d}_{mean}$ |
| *ICtree* | - | 0.6696 | 0.6985 | 0.6797 | 0.6803 |
| *ICPH* | - | 0.6982 | 453.35* | 0.7154 | 0.7069 |
| $ICS_1^{\rho,\gamma}tree$ | (0.43,0.01) | **0.6673** | **0.6798** | **0.6772** | **0.6724** |
| | (0.51,0.03) | **0.6673** | **0.6806** | **0.6772** | **0.6593** |
| | (0.7,0.02) | **0.6689** | **0.6773** | **0.6760** | **0.6615** |
| | (0.03,0.09) | 0.6712 | 0.6958 | **0.6786** | 0.6898 |
| | (0.05,0.21) | **0.6673** | **0.6932** | **0.6785** | 0.6806 |
| *ICWtree* | - | 0.7066 | **0.6628** | **0.6410** | **0.6442** |
| $ICG^{\rho,\gamma}tree$ | (0.03,0.03) | **0.6675** | **0.6793** | 0.6797 | **0.6581** |
| | (0.27,0.01) | 0.6696 | **0.6760** | **0.6789** | **0.6581** |
| | (0.01,0.01) | 0.6696 | **0.6769** | **0.6797** | **0.6581** |
| $ICS_2^{\rho,\gamma}tree$ | (0.51,0.03) | 0.6699 | **0.6819** | **0.6771** | **0.6593** |
| | (0.71,0.02) | 0.6718 | **0.6735** | **0.6760** | **0.6615** |
| | (0.69,0.21) | 0.6731 | **0.6846** | **0.6789** | **0.6769** |

The percentage in parentheses denotes the right-censored rate. The bolded values are the smaller values of $\bar{d}_{mean}$ from our tree methods than the one from *ICtree*.

The performance of *ICWtree* and $ICG^{\rho,\lambda}tree$ are also better *ICtree* for the real data, but are still inferior to $ICS_1^{\rho,\gamma}tree$ in a whole. What's more, $\bar{d}_{mean}$ from the various trees are smaller than the one from *ICPH* introduced in Subsection 3.2.3. An unexpected result for *ICPH* (we emphasis it with *) is obtained. A possible explanation is as follows: if the $j$th interval is a finite interval which means the

observation value falling in the interval is not right-censoring, while one of the $j$th predicted value of the $ICPH$ is far from the endpoints of $j$th interval, but falls to right censoring area, then the predicted time is not in the $j$th interval, so the distance between them is very large. Of course that is a terribly bad prediction for $ICPH$.

## 5. Conclusions

The new test statistics of GLRT are constructed and applied to CIF, then $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$ with hyper-parameter for interval censored survival data are proposed. In simulation, hyper-parameter tuning is explored, the predictive power of $ICS_1^{\rho,\gamma}tree$ is improved by hyper-parameter tuning.

The optimal hyper-parameter values depend on the distribution and the right censoring rate of data. However, we already find some recommended values for real data according to the right censoring rate: (0.05,0.05) is considered as default hyper-parameter value for lower right censoring rate ($\leq 40\%$), (0.51,0.03) is for higher right censoring rate ($> 40\%$). The results of simulation show that $ICS_1^{\rho,\gamma}tree$ has an advantages in predictive accuracy with the recommended hyper-parameter values.

Although in some cases the performance of $ICWtree$ in predictive accuracy beat that of the other tree methods in tree-structured data without right censoring occurring, $ICS_1^{\rho,\gamma}tree$ is still more stable. Those tree algorithms in this paper are provided with R code in github web *.

The ensemble method has powerful predictive performance, it would be interesting to construct the ensemble method by utilizing our new tree methods with hyper-parameter, which will be our future research work.

## Acknowledgements

Data collection of the Signal $Tandmobiel^{®}$ data was supported by Unilever, Belgium. The Signal-$Tandmobiel^{®}$ project comprises the following partners: Dominique Declerck (Department of Oral Health Sciences, KU Leuven), Luc Martens (Dental School, Gent Universiteit), Jackie Vanobbergen (Oral Health Promothion and Prevention, Flemish Dental Association and Dental School, Gent Universiteit),Peter Bottenberg (Dental School, Vrije Universiteit Brussel), Emmanuel Lesaffre (L-Biostat, KU Leuven), and Karel Hoppenbrouwers (Youth Health Department, KU Leuven; Flemish Association for Youth Health Care). The final version of the article benefited from a discussion with prof. Dominique Declerck to which the authors express a sincere thanks.

Authors would like to thank for funding this work through National Natural Science Foundation of China (Grants 11671054), Education Key Laboratory of Symbolic Computation and Knowledge Engineering of Minstry of Education Open Fund Projects (Grants 93K172021K10), Science and Technology Major Project of Changchun, Jilin Province (Grants 20210301038GX).

## Conflict of interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work. No conflict of interest exits in the submission of this manuscript.

*https://github.com/chen-jia-123/EXICtree

## References


1. D. Chen, J. Sun, K. E. Peace, *Interval-censored time-to-event data:Methods and Applications*, 1 Eds., Florida: Chapman and Hall/CRC, 2013. https://doi.org/10.13140/2.1.3493.2169
2. D. R. Cox, Regression models and life-tables, *J. R. Sta. Soc. B,* **34** (1972), 187–220. http://dx.doi.org/10.1111/j.2517-6161.1972.tb00899.x
3. D. M. Finkelstein, A proportional hazards model for interval-censored failure time data, *Biometrics,* **42** (1986), 845–854. http://dx.doi.org/10.2307/2530698
4. J. Sun, *The statistical analysis of interval-censored failure time data*, 1 Eds., New York: Springer Press, 2006. http://dx.doi.org/10.1007/0-387-37119-2
5. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, *Biometrics,* **40** (1984), 358. http://dx.doi.org/10.2307/2530946
6. I. Bou-Hamad, D. Larocque, H. Ben-Ameur, A review of survival trees, *Stat. Surv.,* **5** (2011), 44–71. http://dx.doi.org/10.1214/09-SS047
7. L. Gordon, R. A. Olshen, Tree-structured Survival Analysis, *Cancer Treat. Rep.* **69** (1985), 1065–1069. https://pubmed.ncbi.nlm.nih.gov/4042086/
8. M. R. Segal, Regression trees for censored data, *Biometrics,* **44** (1988), 35–47. http://www.jstor.org/stable/2531894
9. A. Ciampi, S. A. Hogg, S. McKinney, J. Thiffault, RECPAM: A computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics.I. Methods and program features, *Comput. Meth. Prog. Bio.,* **26** (1988), 239–256. http://dx.doi.org/10.1016/0169-2607(88)90004-1
10. A. Ciampi, S. A. Hogg, S. McKinney, J. Thiffault, RECPAM: A computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics.II. Applications to data on small cell carcinoma of the lung (SCCL), *Comput. Meth. Prog. Bio.,* **30** (1989), 283–296. http://dx.doi.org/10.1016/0169-2607(89)90099-0
11. G. V. Kass, An exploratory technique for investigating large quantities of categorical data, *Appl. Stat.,* **29** (1980), 119–127. http://dx.doi.org/10.2307/2986296
12. T. Hothorn, K. Hornik, A. Zeileis, Unbiased recursive partitioning: A conditional inference framework, *J. Comput. Graph. Stat.,* **15** (2006), 651–674. https://doi.org/10.1198/106186006X133933
13. H. Strasser, C. Weber, On the asymptotic theory of permutation statistics, *Math. Methods Stat.,* **8** (1999), 220–250. http://epub.wu.ac.at/102/1/document.pdf
14. W. Fu, J. S. Simonoff, Survival trees for left-truncated and right-censored data with application to time-varying covariate data, *Biostatistics,* **18** (2017), 352–369. http://dx.doi.org/10.1093/biostatistics/kxw047
15. L. Breiman, Random forests, *Mach. Learn.,* **45** (2001), 5–32. http://dx.doi.org/10.1023/A:1010933404324
16. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, Random survival forests, *Ann. Appl. Stat.,* **2** (2008), 841–860. http://dx.doi.org/10.1214/08-AOAS169

17. J. A. Steingrimsson, L. Diao, R. L. Strawderman, Censoring unbiased regression trees and ensembles, *J. Am. Stat. Assoc.,* **114** (2019), 370–383. http://dx.doi.org/10.1080/01621459.2017.140777

18. Y. M. Yin, S. J. Anderson, Tree-structured modeling for interval-censored survival data, *Joint Statistical Meetings*, (2002), 3877–3882. https://www.researchgate.net/publication/265027875

19. W. Fu, J. S. Simonoff, Survival trees for interval-censored survival data, *Stat. Med.,* **36** (2017), 4831–4842. http://dx.doi.org/10.1002/sim.7450

20. W. Pan, Rank invariant tests with left truncated and interval censored data, *J. Stat. Comput. Sim.,* **61** (1998), 163–174. http://dx.doi.org/10.1080/00949659808811907

21. H. Y. Cho, N. P. Jewell, M. R. Kosorok, Interval censored recursive forests, *J. Comput. Graph. Stat.,* (2021), in press. https://doi.org/10.1080/10618600.2021.1987253

22. J. G. Sun, Q. Zhao, X. Q. Zhao, Generalized log-rank tests for interval-censored failure time data, *Scand. J. Stats.,* **32** (2005), 49–57. http://dx.doi.org/https://doi.org/10.1002/bimj.200710419

23. J. Vanobbergen, L. Martens, E. Lesaffre, D. Declerck, The Signal-Tandmobiel project a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results, *Eur. J. Paediatr. Dent.,* **2** (2000), 87–96. http://hdl.handle.net/1854/LU-127864

24. C. Anderson-Bergman, An efficient implementation of the EMICM algorithm for the interval censored NPMLE, *J. Comput. Graph. Stat.,* **26** (2017), 463–467. http://dx.doi.org/10.1080/10618600.2016.1208616

25. C. Anderson-Bergman, icenReg: Regression models for interval censored data. Version 2.0.15. , (2020). https://cran.r-project.org/web/packages/icenReg/index.html.

26. Y. Benoist, P. Foulon, F. Labourie, On Convergence of convex minorant algorithms for distribution estimation with interval-censored data, *J. Comput. Graph. Stat.,* **1** (1992), 129–140. http://dx.doi.org/10.1080/10618600.1992.10477009

27. G. Gomez, R. O. Pique, A new class of rank tests for interval-censored data, *Harvard University Biostatistics Working Paper Series,* (2008), unpublished work. http://biostats.bepress.com/harvardbiostat/paper93

28. P. Wei, A comparison of some two-sample tests with interval censored data, *J. Nonparametr. Stat.,* **12** (1999), 133–146. https://doi.org/10.1080/10485259908832801

29. D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, *J. Cheminformatics,* **6** (2014), 1–15. http://www.jcheminf.com/content/6/1/10

30. T. R. Tsai, S. H. Wu, Y. Shen, Model selection methods for reliability assessment based on interval-censored field failure samples, *Int. J. Reliab. Qual. Sa.*, **27** (2020), 1–19. http://dx.doi.org/10.1142/S0218539320500187

31. E. Graf, C. Schmoor, W. Sauerbrei, M. Schumacher, Assessment and comparison of prognostic classification schemes for survival data, *Stat. Med.,* **18** (1999), 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18¡2529::AID-SIM274¿3.0.CO;2-5

32. S. Tsouprou, *Measures of discrimination and predictive accuracy for interval censored survival data*, *MA. D. thesis,* University Leiden, 2015. https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/mastertsouprou.pdf

33. T. Hothorn, B. Lausen, A. Benner, M. Radespiel-Troger, Bagging survival trees, *Stat.Med.,* **23** (2004), 77–91. https://doi.org/10.1002/sim.1593

34. A. Komárek, *bayesSurv: Bayesian survival regression with flexible error and random effects distributions*, R package version3.3, 2020. https://cran.r-project.org/web/packages/bayesSurv/index.html

35. E. Lesaffre, A. Komárek, D. Declerck, An overview of methods for interval-censored data with an emphasis on applications in dentistry, *Stat. Methods Med. Res.,* **14** (2005), 539–552. https://doi.org/10.1191/0962280205sm417oa

36. W. C. Yao, H. Frydman, J. S. Simonoff, An ensemble method for interval-censored time-to-event data, *Biostatistics,* **22** (2021), 198–213. http://dx.doi.org/10.1093/biostatistics/kxz025

## Appendix

**Table 9.** The optimal hyper-parameter values $(\tilde{\rho}, \tilde{\gamma})$ for classification under Weibull1 distribution.

| Type of Trees | 0% | 20% | 40% |
|---|---|---|---|
| | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ |
| $ICS_1^{\rho,\gamma}tree$ | (0.57,0.33) | (0.53,0.37) | (0.09,0.13) |
| $ICS_2^{\rho,\gamma}tree$ | (0.21,0.01)<br>- | (0.21,0.05)<br>- | (0.25,0.21)<br>(0.29,0.21) |

0%,20% and 40% denote the right censoring rate of tree-structured data.

**Table 10.** The optimal hyper-parameter values $(\tilde{\rho}, \tilde{\gamma})$ for classification under Weibull2 distribution.

| Type of Trees | 0% | 20% | 40% |
|---|---|---|---|
| | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ |
| $ICS_1^{\rho,\gamma}tree$ | (0.09,0.09) | (0.09,0.01) | (0.41,0.01) |
| $ICS_2^{\rho,\gamma}tree$ | (0.61,0.45) | (0.41,0.17) | (0.21,0.01) |

0%,20% and 40% denote the right censoring rate of tree-structured data.

**Table 11.** The optimal hyper-parameter values $(\tilde{\rho}, \tilde{\gamma})$ for classification under Log-Normal distribution.

| Type of Trees | 0% | 20% | 40% |
|---|---|---|---|
| | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ | $(\tilde{\rho}, \tilde{\gamma})$ |
| $ICS_1^{\rho,\gamma}tree$ | (0.17,0.01) | (0.05,0.03) | (0.09,0.09) |
| | - | - | (0.07,0.03) |
| $ICS_2^{\rho,\gamma}tree$ | (0.41,0.01) | (0.21,0.01) | (0.29,0.37) |

0%, 20% and 40% denote the right censoring rate of tree-structured data.