Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Decision tree for modeling survival data with competing risks

Kazeem Adesina Dauda [a], *Biswabrata Pradhan* [b], *B. Uma Shankar* [c,*], *Sushmita Mitra* [c]

[a] *Department of Statistics and Mathematical Sciences, Kwara State University, Malete, Nigeria*
[b] *SQC & OR Unit, Indian Statistical Institute, Kolkata, India*
[c] *MIU, Indian Statistical Institute, Kolkata, India*

ARTICLE INFO

Article history:
Received 15 January 2019
Received in revised form
16 April 2019
Accepted 1 May 2019
Available online 4 June 2019

Keywords:
Cross validation
Conditional incident function
Impurity measure
Log-odds residual
Marginal distribution function
Survival Classification and
Regression Tree (SCART)

ABSTRACT

This work considers decision tree for modeling survival data with competing risks. A Survival Classification and Regression Tree (SCART) technique is proposed for analysing survival data by modifying classification and regression tree (CART) algorithm to handle censored data for both regression and classification problems. Different performance measures for regression and classification tree are proposed. Model validation is done by two different cross-validation methods. Two real life data sets are analyzed for illustration. It is found that the proposed method improve upon the existing classical method for analysis of survival data with competing risks.

1. Introduction

Survival analysis with competing risks data is very common in the field of medical studies and now gaining more attention and interest in many research areas [1–3]. For instance, a physician may be interested in time to death of a particular disease, but the patient can die from another competing disease(s) that may be potentially dependent or independent of disease of interest [4]. The most common classical methods for analyzing survival data with and without competing risks are the Cox proportional hazard (Cox-PH) method [5], cumulative incidence function (CIF) techniques [6,7] and Fine and Gray method [8] among others. Fine and Gray proposed a proportional hazards model, aiming to model the CIF with covariates by treating the CIF curve as a subdistribution function. For further investigation on classical competing risks analysis, see [8], Chen et al. [9], Webin and Yu [10], Haller et al.

* *Corresponding author at*: MIU, Indian Statistical Institute, Kolkata, India.
E-mail address: uma@isical.ac.in (B. Uma Shankar).

https://doi.org/10.1016/j.bbe.2019.05.001

List of symbols used

Symbol	Symbol description
MF	marginal distribution function
CIF	conditional incidence function
RF	log-odds residuals
T	failure time
C	failure mode
k	number of causes
$\lambda_j(t)$	cause-specific hazard function for cause j
S(t)	survival function
$\Lambda_j(t)$	cumulative hazard function for cause j
$F_j(t)$	cumulative incidence function for cause j
Δ	censoring random variable
$\hat{\Lambda}_j(t)$	Nelson-Aalen estimator of $\Lambda_j(t)$
$\lambda_{0j}(t)$	base line hazard function for cause j
N_h	number of cases in node h
ϕ	goodness of split
L	split to the left (left daughter node or left child node)
R	split to the right (right daughter node or right child node)
SS	sum of squared deviations
SS(h)	sum of squared deviations of the node h
SS(L)	sum of squared deviations of the left child node
SS(R)	sum of squared deviations of the right child node
SS(P)	sum of squared deviations of the parent node
ϕ_{LS}	goodness of least-squares split
ϕ_{LAD}	goodness of least absolute deviation split
ϕ_{LS}(MF)	goodness of least-squares split for MF
ϕ_{LS}(CIF)	goodness of least-squares split for CIF
ϕ_{LS}(RF)	goodness of least-squares split for RF

[11], Gerd et al. [12], Kundu and Pradhan [13], and Bhattacharya et al. [14].

The decision tree has the advantage that the results can be easily understood and explained. The motivation of tree-based survival analysis came into practice as one may be interested in identifying the groups effect based on the prognostic factors. However the classical methods lack the capability of partitioning the subjects into subgroups based on the prognostic factors in the form of tree structure (which is readily understandable to practitioners). Also it may be noted that the classical methods are based on stringent model assumptions [15].

There have been a number of works on tree-based survival analysis by using different splitting criteria. Gordon and Olshen [16] introduced CART technique for censored data. The data mining methods such as multivariate adaptive regression splines (MARS) [17,18], and Bayesian adaptive regression splines [19] are also introduced in the literature. A tree-based survival analysis that utilizes log-rank test statistic as splitting criteria is developed by [20]. Similar works are considered by LeBlanc and Crowley [21] by introducing an efficient pruning algorithm for within-node separation. Keles and Segal [22] developed a residual-based tree-structure via martingale residual and its splitting rule was developed accordingly. In recent time, many authors such as [23] developed a multivariate survival tree using robust Wald-test statistic as splitting rule and [24] also developed a survival forest based on dependent censoring.

All these mentioned research works considered survival analysis without competing risk data except Ishwaran et al. [25–27]. However, [25] modified the random forest algorithm for survival data by constructing trees via maximization of between-node heterogeneity (log-rank score). Moreover, [28] proposed tree-structure survival analysis with competing risks which is based on between-node heterogeneity and defined deviance statistic derived from the likelihood ratio test as the impurity function. Similarly, [29] developed a classification tree in competing risks based on within-node homogeneity survival data using Martingale residual as splitting rule by modifying CART algorithm. One can see related works on survival analysis with competing risks [3,30–41]. However, none of these considered the simultaneous effect of two or more competing events, including the conditional inference tree is based on the conditional probability of an unbiased recursive partitioning of tree [42].

Recently, Bellot et al. [43] considered analysis of competing risk data by multitask learning technique with the help of resampling and boosting. Lee et al. [30] introduced the deep neural network called Deephit into competing risks using the estimated CIF as a common response. This strictly refers to unsupervised learning, which mainly works with the covariates.

The aim of this work is to propose different mechanisms for building tree-based survival analysis with competing risks data. We propose a decision-tree by modifying CART algorithm to handle censored data for both regression and classification trees. The impurity measures are provided based on MF, CIF and RF. The resulting MF, CIF and RF estimates are further categorized into low and high risks using their median and results are then used as response in classification tree.

The contribution of this paper can be described as follows:

- An algorithm SCART is proposed which incorporates the concept of survival analysis with competing risks into the framework of CART.
- The proposed SCART is implemented by using impurity measures based on marginal cumulative function (MF), conditional incidence function (CIF) and log-odds residuals (RF).
- The SCART for survival analysis with competing risks provide an understandable and clear interpretation of survival data.
- Splitting rules for SCART are proposed based on MF, CIF and RF.
- The methods provide a solution to simultaneous inference on survival analysis with competing risks, without any assumption on the distribution of the data and sub-distribution hazard assumptions as the case may be.
- The tree-based technique provides a trade-off between the bias and variance of the model and prevent both over-fitting and under-fitting.

The major advantage of MF, CIF and RF is that, unlike other existing methods they allow incorporation of competing risks effect simultaneously. However, the limitation of this study is its inability to work with survival analysis without competing risk.

The rest of the paper is organized as follows. Competing risks model and related quantities are discussed in Section 2. Tree-based survival analysis for competing-risks data is considered in Section 3. Different performance measures for regression and classification tree are proposed in Section 4. Model validation by different cross-validation methods is discussed in Section 5. Classical model assessment method is discussed in Section 6. Two real life data sets are analyzed for illustration in Section 7. Finally, conclusions appear in Section 8.

2. The model

Here we consider competing risks model and related quantities which will be used for construction of decision tree. Suppose T is the failure time and C is the mode of failure, where C takes on values in the set $\{1, \ldots, k\}$. For each individual we observe the pair (T, C). The joint distribution of T and C is specified by the cause-specific hazard function given by:

$$\lambda_j(t) = \lim_{dt \to 0} \frac{Pr(t \le T < t + dt, C = j / T \ge t)}{dt}, \quad j = 1, \ldots, k. \tag{1}$$

The survival function of T is given by $S(t) = Pr(T \ge t) = \exp[-\sum_j^k \Lambda_j(t)]$, where $\Lambda_j(t) = \int_0^t \lambda_j(u)du$ is the cumulative hazard function for cause j, for $j = 1, \ldots, k$. The marginal cumulative distribution function of T is given by:

$$F(t) = 1 - S(t) = \sum_{j=1}^{k} F_j(t), \tag{2}$$

where $F_j(t) = Pr(T \le t, C = j) = \int_0^t S(u)d\Lambda_j(u)$ is the cumulative incidence function for cause j.

Next, we provide conditional incidence function (CIF) which is defined (Pepe and Mori [6]) as the probability of observing an event by time t given that a subject did not experience a competing cause, and is given by:

$$\begin{aligned} CIF_j &= Pr(T \le t, C = j / \text{no other events occurred by time } t) \\ &= Pr(T \le t, C = j / (T \le t, C \ne j)^c) \\ &= \frac{Pr(T \le t, C = j)}{1 - Pr(T \le t, C \ne j)} \\ &= \frac{F_j(t)}{1 - \sum_{\forall\ k \ne j} F_k(t)} \end{aligned} \tag{3}$$

The log-odds residual(RF) for the MF and is given by:

$$RF = \log\left(\frac{1 - F(t)}{F(t)}\right). \tag{4}$$

We consider the censored data on n individuals. Let Δ be the censoring random variable. We observe the data (t_i, δ_i, C_i), where $t_i = \min(T_i, \Delta_i)$ and $\delta_i = I(T_i \le \Delta_i)$. If T_i is censored at t_i, then the cause of failure is unknown. So the observed data for individual i consist of either $(T_i = t_i, C_i)$ or $T_i > t_i$. The Nelson-Aalen estimator of $\Lambda_j(t)$ is given by $\hat{\Lambda}_j(t) = \sum_{i:t_i \le t} \frac{\delta_{ij}}{n_i}$, where $\delta_{ij} = I(C_i = j, \delta_i = 1)$, n_i is the number of individuals alive and uncensored just prior to time t_i. An estimate of $S(t)$ is given by:

$$\hat{S}(t) = \exp\left[-\sum_{j=1}^{k} \hat{\Lambda}_j(t)\right]$$

The cumulative incidence function $F_j(t)$ is estimated by:

$$\hat{F}_j(t) = \int_0^t \hat{S}(u)d\hat{\Lambda}_j(u) = \hat{F}_j(t) = \sum_{i:t_i \le t} \hat{S}(t_i)\frac{\delta_{ij}}{n_i}.$$

Then RF is estimated by

$$\widehat{RF} = \log\left(\frac{1 - \hat{F}(t)}{\hat{F}(t)}\right). \tag{5}$$

Next, we consider regression model for the competing risks data. The survival probability under the effect of covariate Z is given by $S(t|Z) = P[T \ge t|Z] = \exp(-\Lambda(t|Z))$, where $\Lambda(t|Z) = \int_0^t \lambda(u|Z)du = \sum_{j=1}^k \int_0^t \lambda_j(u|Z)du$. We consider Cox proportional hazard model for $\lambda_j(t|Z)$ as follows:

$$\lambda_j(t|Z) = \lambda_{0j}(t)\exp(\beta_j' Z), \quad j = 1, \ldots, k.$$

where $\lambda_{0j}(t)$ is the base line hazard function for mode j.

The parameters $\beta_{01}, \ldots, \beta_{0k}$ are estimated by the partial likelihood method and $\Lambda_{0j}(t)$ is estimated as generalized Nelson-Aalen estimator (see Lawless [44], for details) by using the estimates $\hat{\beta}_{01}, \ldots, \hat{\beta}_{0k}$. Then $S(t|Z)$ for $Z = z$ is estimated by:

$$\hat{S}(t|z) = \exp[-\hat{\Lambda}(t|z)], \tag{6}$$

where $\hat{\Lambda}(t|z) = \sum_{j=1}^k \hat{\Lambda}_{0j}(t|z)$.

3. Regression tree for competing risks

Here we consider construction of regression and classification tree for survival data with competing risks. Learning decision tree for classification or regression instances involves moving from the root node to the leaf nodes. It recursively and repeatedly partition the covariates until they contain few and homogeneous response variable like MF, CIF and RF via impurity functions. The impurity functions are described below.

The common splitting rules or impurity functions for CART are least-squares (LS) and least absolute deviation (LAD) functions. In this work, we consider only the LS impurity measure. Note that the mechanism for both the functions are same. When dealing with the continuous outcome (y) in CART, the regression tree uses sum of squared deviations around node average as a measure of impurity. The sum of squared deviations is defined in Eq. (7):

$$SS(h) = \sum_{i \in h} (y_i - \bar{y}_h)^2 \tag{7}$$

where $\bar{y}_h = \frac{\sum_{i \in h} y_i}{N_h}$ and N_h is the number of cases in node h. Thus, the corresponding goodness of least-squares split method ϕ_{LS} for partitioning parent node P into left L and right R child nodes are given by:

$$\phi_{LS} = SS(P) - SS(L) - SS(R). \tag{8}$$

The quantity in Eq. (7) require high computational burden in the part where we need all the possible split on a variable. Each potential sub-tree need to be evaluated, by going from Eq. (7)

to (8) recursively. It should be noted that Eq. (7) can be rewritten in the form of quantity in Eq. (9).:

$$SS(h) = \sum_{i \in h} y_i^2 - N_h \bar{y}_h^2. \tag{9}$$

Thus, $SS(P) = \sum_{i \in P} y_i^2 - N_P \bar{y}_P^2$, $SS(L) = \sum_{i \in L} y_i^2 - N_L \bar{y}_L^2$ and $SS(R) = \sum_{i \in R} y_i^2 - N_R \bar{y}_R^2$. All these quantities are substituted into Eq. (8) and this yield

$$\phi_{LS} = N_L \bar{y}_L^2 + N_R \bar{y}_R^2 - N_P \bar{y}_P^2. \tag{10}$$

By replacing $\bar{y}_p$ with $\frac{[N_L \bar{y}_L + N_R \bar{y}_R]}{N_P}$ in (10), we have

$$\phi_{LS} = \frac{N_L N_R (\bar{y}_L - \bar{y}_R)^2}{N_P}. \tag{11}$$

In order to extend the CART structure to survival analysis with competing risks, MF, CIF and RF are treated as outcome variables. Therefore, the impurity measures based on MF, CIF and RF respectively, are given as follows.:

$$\phi_{LS}(MF) = \frac{N_L N_R (\bar{F}(t)_L - \bar{F}(t)_R)^2}{N_P} \tag{12}$$

$$\phi_{LS}(CIF) = \frac{N_L N_R (\overline{CIF}_L - \overline{CIF}_R)^2}{N_P} \tag{13}$$

$$\phi_{LS}(RF) = \frac{N_L N_R (\overline{RF}(t)_L - \overline{RF}(t)_R)^2}{N_P}, \tag{14}$$

where $\bar{F}(t)_L$, $\overline{CIF}_L$ and $\overline{RF}(t)_L$ are the means of estimated MF, CIF and RF, respectively, in left node of the tree, and $\bar{F}(t)_R$, $\overline{CIF}_R$ and $\overline{RF}(t)_R$ are the means of estimated MF, CIF and RF, respectively, in right node of the tree.

4. Procedure of SCART algorithm

The **Input 1** in the algorithm contain a set of observed survival data $(T_i, \delta_i, C_i, Z_i)$ where $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})'$ is a vector of covariates, $i = 1, 2, \ldots, n$. Then, the corresponding values of MF, CIF and RF in Eqs. (2)–(4) respectively, are computed with regards to whether the focus is on regression or classification (**Input 2**) and this gives a new data set with MF, CIF and RF. The new data set will be passed on to lines 1–22 to build the SCART model (**output**).

Learning decision tree for classification or regression instances involves moving from the root node top down to the leaf nodes. In particular, it recursively and repeatedly partitions the covariates (features) until they contain few homogeneous response variables.

The impurity functions provided in Eqs. (12)–(14) are used for partitioning. The partitioning is done based on the nature of the covariate. If the covariate is categorical, then the splitting is done on the classes of the covariate. For continuous or ordinal variables, the cut-point values are chosen via the median of the covariate or some other acceptable criteria established in the literature. The final node (terminal) of the tree contains some homogenous information of the response variable and gives the details of the effects of a covariate on the response. In case of multiple covariates, the best classifier among covariates is the one with the minimum impurity value. The first most useful covariate will be put under root node and split into left and right child node depending on the number of classes and the stopping criteria (i.e. node with certain number of observations). Subsequently, other covariates will then be introduced into the tree based on their usefulness with regards to the impurity measures, this will continue until all the covariates are used up.

The procedure of decision tree learning is described in Algorithm 1. In line 4, of the algorithm, the SCART is designed to select the most useful covariate with respect to the impurity measures. In this case, the covariates that contribute maximally to the impurity measures would be selected at each iteration. The proposed procedure is presented as flowchart in Fig. 1. After growing the tree, pruning [20] is adopted to remove the redundant trees that can cause overfitting of the fitted model. The basic idea behind pruning is that it involves performing a complete induction on the data, splitting the nodes until they cannot be split any further. A pruning algorithm then tests the performance of the tree with and without its lower nodes, and those nodes are removed if the test set error improves. The proposed Algorithm 1 provides a Survival Classification And Regression Tree (SCART) for analysis of survival data with competing risks.

Algorithm 1. Survival Classification And Regression Tree (SCART)

Input 1: A set of training survival data. Let z_{is} be the value of the covariate Z_s for i^{th} individual in the training set, for $s = 1, 2, \ldots, p$.
Input 2: Compute the values of MF or CIF or RF and formulate a new data frame that contains the marginal CIF in the training set
Output: A Survival Classification and Regression Tree by High risk and Low risk groups with marginal cumulative distribution function

1 **if** *the stopping criterion is satisfied* **then**
2 create a leaf that corresponds to all remaining training instances ;
3 **else**
4 choose the best covariate Z_s;
5 label the current node with Z_s;
6 **for** *each value* z_{is} *of the covariate* Z_s **do**
7 label the outgoing edge with value Z_s;
8 recursively build a sub-tree based on a subset of training instances that meet the condition $Z_s = z_s$;
9 **end**
10 **Return Tree** (T)
11 **end**
12 **Tree Pruning**
13 **while** T *is grown* **do**
14 Select a node t in T such that pruning it maximally improve some evaluation criteria ;
15 **for** $T \neq \phi$ **do**
16 $T = Pruned(T, t)$;
17 **end**
18 **repeat**
19 step 15-17
20 **until** $T = \phi$;
21 **Return Pruned Tree** (T)
22 **end**

5. Performance measures

In this section, various performance indices for regression and classification trees are considered.

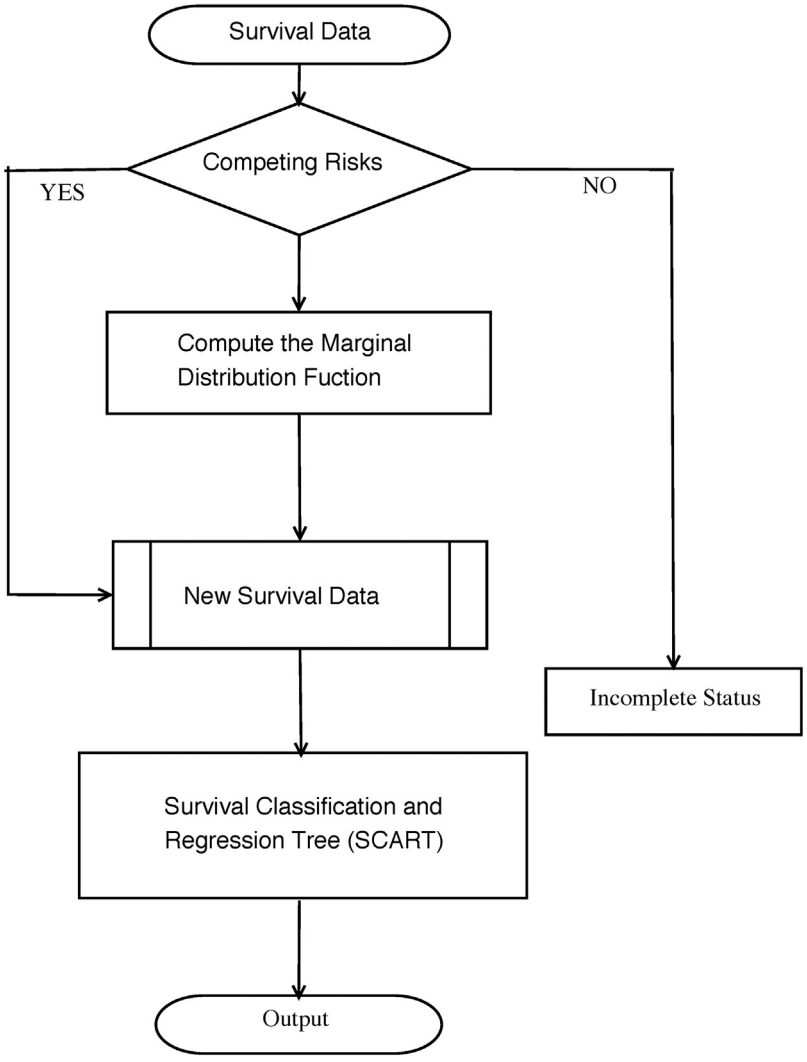

Fig. 1 – Proposed flowchart.

5.1. *Regression tree performance measure*

The common measures for comparing performance of two regression models are mean square error (MSE) and root mean square errors (RMSE) and they are given by:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

and

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}},$$

where $y_1, \ldots, y_n$ are the observed values and $\hat{y}_1, \ldots, \hat{y}_n$ are the corresponding predicted values.

In survival analysis setup, [45] pointed out that a straightforward approach can be used to predict time-to-event $\hat{\tau}$ of a patient with covariate $Z = z$ and this may be the mean or median survival time of the patients. Moreover, $\hat{\tau}$ may also be a quantity suggested by the clinician. Then, the mean square error (MSE) and root mean square errors (RMSE) are given by:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(t_i - \hat{\tau})^2}{n}$$

and

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - \hat{\tau})^2}{n}}.$$

5.2. *Classification tree performance measure*

Here we propose performance measure for classification tree by discretizing the response variable (MF) into high risk and low risk of surviving from any kind of diseases. The procedure for splitting the MF is given by:

$$\text{MF} = \begin{cases} 1 \text{ (High risk)}, & \text{if MF} \geq \text{median (MF)} \\ 0 \text{ (Low risk)}, & \text{if MF} < \text{median (MF)}. \end{cases}$$

We apply Algorithm 1, for construction of classification and regression tree, but the proposed impurity measures need to be adjusted in order to account for a categorical response. The modified Gini Index technique originally proposed by [46] is used to deal with the problem of impurity measures. This modification is done by introducing some model performance measures such as Leave-One-Out and Monte Carlo Cross Validations [47–49]. We consider random forest to circumvent the problem of imbalance data, if arises, during tree splitting. The tree-based technique provides a trade-off between the bias and variance of the model and prevent both over-fitting and under-fitting. In the end, the performance of our fitted models is tested using classification error rate and ROC curve. The details of all these indices are given in the subsequent sections. Ultimately, based on the available data, we measure the Prediction Accuracy Rate (PAR) and Prediction Error Rate (PER) and they are given by:

$$\text{Prediction Accuracy Rate (PAR)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Prediction Error Rate (PER)} = \frac{FP + FN}{TP + TN + FP + FN},$$

where TP = number of true positive, FP = number of false positive, TN = number of true negative and FN = number of false negative.

6. Model validation

The validation is commonly used to assess and enhance the model's performance in classification problem [50]. We consider model performance by cross validation methods. Leave-One-Out Cross Validation (LOOCV) and Monte Carlo Cross Validation (MCCV) techniques are considered in this work.

Leave-One-Out Cross Validation (LOOCV) is one of the validation methods that involves leaving out one data point while building the model by the rest of the data and the model is tested on the basis of left out data point. The procedure is repeated until all the data points are independently tested [47,51,52]. The accuracy of each point can be tested using the area under the curve (AUC) and prediction accuracy. In the end, the best model is chosen based on one of these indices.

Monte Carlo Cross Validation (MCCV) method is another cross-validation technique which commonly used in chemometrics and it has been proofing to perform better than LOOCV in some conditions [53,52]. It was first introduced by [54]. The procedure of MCCV is similar to LOOCV, but different in the sense that the MCCV leaves out a notable part of the dataset (not a point like LOOCV) called training (calibration) set (e.g. 80%) and the second part (e.g. 20%) called test (validation) set at a time during the model building. This process is repeated until the pre-specified number of iterations is reached. At each iteration, the value of the performance measures is recorded and at the end the best model is chosen based on these value.

7. Classical model assessment

For the classical model, Brier score (BS) [55] is used for model prediction accuracy together with ROC curve. Various methodology of these criteria assessment have been described in literatures [56–59] and their explanation are given in the subsequent sections.

7.1. Brier score (BS)

Brier score (BS) [55] is defined as the square distance between the observed status $Y_i = I(T_i > t_i)$, and the predicted survival probability $\hat{S}(t_i|z)$ defined in Eq. (6). [56,60–64] make use of BS as a way of identifying model prediction accuracy in survival analysis. The BS is mathematically denoted as:

$$\text{BS} = (Y(C = status) - \hat{S}(t_i|z))^2,$$

where

$$Y(C = status) = \begin{cases} 1 : \text{if } T \le \Delta, & \text{that is, } T \text{ is uncensored} \\ 0 : \text{if } T > \Delta, & \text{that is, } T \text{ is censored.} \end{cases}$$

Then the BS is defined by:

$$BS = \frac{\sum_{i=1}^{n} (Y_i(C = status) - \hat{S}(t_i|z))^2}{n}.$$

7.2. ROC curve

The ROC curve is the set of functions that map from false positive rate (FPR) to the true positive rate (TPR) [56,65,66]. However, when the outcomes are binary, the accuracy of its prediction depends on correct classification rate called sensitivity and specificity. The correct definition of these classification rates under survival analysis based on the predicted survival probabilities $\hat{S}(t_i|z)$ of the marker (covariate) with respect to the cut-point $\xi \in [0, 1]$ can be given as:

$$\text{Sensitivity} = Pr(\hat{S}(t|z) \ge \xi | Y(C = 1))$$

$$\text{Specificity} = Pr(\hat{S}(t|z) < \xi | Y(C = 0)).$$

Thus, the corresponding FPR and TPR can be derived using the quantities above as follows:

$$\begin{aligned} \text{TPR} = \text{Sensitivity} &= Pr(\hat{S}(t|z) \ge \xi | Y(C = 1)) \\ &= \frac{Pr(\hat{S}(t|z) \ge \xi, Y(C = 1))}{Pr(Y(C = 1))} \end{aligned}$$

$$\begin{aligned} \text{FPR} = 1 - \text{Specificity} &= 1 - Pr(\hat{S}(t|z) < \xi | Y(C = 0)) \\ &= Pr(\hat{S}(t|z) \ge \xi | Y(C = 0)) \end{aligned}$$

This gives

$$\text{FPR} = \frac{Pr(\hat{S}(t|z) \ge \xi, Y(C = 0))}{Pr(Y(C = 0))}$$

where ξ is the all possible threshold (cut-point), and ROC is monotonically increasing function in [0, 1]. Note that $Pr(Y(C = 1))$ is the probability that an individual experience the event of interest $Y(C = 1)$. Since only the event and censoring can be consider at a time, then

$$Pr(Y(C = 1)) + Pr(Y(C = 0)) = 1$$

Note that $Pr(\hat{S}(t|z) \geq \xi, Y(C=1))$ is the predicted probability of an individual at a certain cut-point that experience the event, while $Pr(\hat{S}(t|z) \geq \xi, Y(C=0))$ is the predicted probability of an individual at a certain cut-point that does not experienced the event (censored). For simplicity, we follow the counting process pointed out in [56] as:

$$Pr(\hat{S}(t|z) \geq \xi, Y(C=status)) = \sum_{i=1}^{n} I[\hat{S}(t_i|z) \geq \xi, Y_i(C=status)]$$

and

$$Pr(Y(C=status)) = \sum_{i=1}^{n} I[Y_i(C=status)].$$

Thus, the corresponding estimates of TPR and FPR can be given as:

$$\widehat{TPR} = \frac{\sum_{i=1}^{n} I[\hat{S}(t_i|z) \geq \xi, Y_i(C=1)]}{\sum_{i=1}^{n} I[Y_i(C=1)]}$$

$$\widehat{TPR} = \frac{\sum_{i=1}^{n} I[\hat{S}(t_i|z) \geq \xi, Y_i(C=0)]}{\sum_{i=1}^{n} I[Y_i(C=0)]}.$$

8. Data analysis

In this section, we consider two real life data for illustration.

8.1. *Analysis of primary biliary cirrhosis (PBC) data*

The Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver transplant was conducted between 1974 and 1984 [67]. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo-controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis. The patients age (years), albumin: serum albumin (g/dl), alkaline phosphotase (U/l) (alk.phos), ascites (presence of ascites), ast: aspartate aminotransferase (U/ml), bili: serum bilirubin (mg/dl), chol: serum cholesterol (mg/dl), copper: urine copper (ug/day), edema: (0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy), hepato (presence of hepatomegaly or enlarged liver), platelet (platelet count), protime: standardized blood clotting time, sex, spiders: blood vessel malformations in the skin, stage: histologic stage of disease (needs biopsy), and trig: triglycerides (mg/dl) were considered as a major factors that influence patience survival from the liver transplant. Moreover, the survival time (in days) and status of the patients with two competing risks transplant (1) and death (2), and censored (0) are all present in the data.

A careful data cleaning is done on the data to remove some noise data and the row with most missing observation and these processes reduce the data to 312 observations. The summary of the data is given in Table 1.

Results from Table 1 reveal how patients were randomized based on the treatment they received together with the status of their outcome. It may be noted that some patient receives D-penicillamine and yet they experience death before the transplant, this amount to sixty-five (65) of them, and sixty (60) patient were given no treatment (placebo) and yet experience death before the transplantation. Moreover, the number of patients that finally got the organ with respect to whether they receive both D-penicillamine and placebo are 10 and 9, respectively.

Table 1 – Summary of the characteristics of 312 liver patients with respect to the number and types of events (outcomes) as well as the treatment received.

Treatment	Event (Outcome/Status)			
	Censored (0)	Transplant (1)	Death (2)	Total
D-penicillamine	83	10	65	**158**
Placebo	85	9	60	**154**
Total	**168**	**19**	**125**	**312**

In order to identify and classify the patients into their respective prognostic risk, the tree-structured survival analysis are employed based on the three proposed methodologies as explained in the previous section and some other existing methods. Firstly, we consider the use of MF on the PBC data, this account for joint behaviors of the patient that experienced both the events and we have to build a tree in that respect. This shows that the joint behaviors of the patient that experience both events, as the waiting time for transplantation increases, the risk of not surviving the disease also increases since those who experience the death (125) are far more than those that experience the transplantation (19). Next, we classify the patients based on CIF. Finally, we build tree via the log-odds residuals (RF). Implementing standard CART into survival analysis has been problematic due to censoring outcome, but the proposed log-odds residuals (RF) approach using SCART can handle this difficulty posed by the censoring. Here, the log-odds residual is adopted as the outcome in the standard CART and this also combines the events as in MF.

8.1.1. *Regression tree performance measure*

The performance of the regression tree techniques is assessed using mean square error as shown in Table 2. It is observed that CIF perform better than any other technique irrespective of the impurity.

8.1.2. *Classification tree performance measure*

Furthermore, we investigate the performance of the proposed classification tree techniques using Prediction Error Rate and area under the curve (AUC). The results are presented

Table 2 – Performance measures on the various proposed methods via regression tree on PBC data.

Method	Mean square error	Root mean square error
MF	0.0156	0.1250
CIF	**0.0026**	**0.0514**
RF	0.5492	0.7411
Classical methods	0.8539	0.9241

Table 3 – Performance measures on the various proposed methods via classification tree on PBC data.

Method	Prediction accuracy	Area under curve
MF	0.8438	**0.9042**
CIF	0.7188	0.7857
RF	0.8125	**0.9042**
Classical methods	**0.8600**	0.7394
Random survival forest (cause 2)	0.6296	0.6988

in Table 3 and Fig. 2. It is observed that the MF perform better than any other techniques irrespective of the impurity.

8.1.3. Performance measure of both LOOCV and MCCV

Finally, we investigate the performance of the proposed classification tree techniques by LOOCV and MCCV. The results are presented in Table 4 and we observe that the results of LOOCV and MCCV are similar irrespective of the impurity.

8.2. Results on Hodgkin's disease (HD)

The second data set is on Hodgkin's disease available in [68]. The data contain six (6) covariates of the patients and this include age, sex, trtgiven: treatment given (RT = radiation, CMT = chemotherapy and radiation), medwidsi: mediastinum involvement (N = no, S = small, L = large), extranod: extra-nodal disease (M = extra-nodal disease, N = nodal disease), and clinstg: clinical stage (1 = stage I, 2 = stage II). The survival time is given in years and the status with two competing risks (relapse (1) and death (2)) and zero (0) if censored. The total number of patients treated was 865 between 1968 and 1986 from the Princess Margaret Hospital.

The performance results of the proposed methods and other classical method based on regression and classification trees of the data are provided in Tables 5 and 6. The ROC of the classification tree are presented in Fig. 3 for further clarification. The various results under regression tree revealed that

Table 4 – Comparing performance of MCCV and LOOCV on PBC.

Method	% PER	% PAR
MF (MCCV)	29	71
CIF (MCCV)	32	68
RF (MCCV)	23	77
MF (LOOCV)	25	75
CIF (LOOCV)	32	68
RF (LOOCV)	26	74

PER = Prediction Error Rate and PAR = Prediction Accuracy Rate.

Table 5 – Performance measures on the various proposed methods via regression tree on Hodgkin's disease (HD).

Method	Mean square error	Root mean square error
MF	0.0232	0.1523
CIF	**0.0130**	**0.1152**
RF	0.8432	0.9182
Classical method	0.9034	0.9505

Table 6 – Performance measures on the various proposed methods via classification tree on Hodgkin's disease (HD).

Method	Prediction accuracy	Area under curve
MF	**0.7955**	**0.8015**
CIF	**0.7955**	**0.8015**
RF	0.7727	**0.8179**
Classical method	0.7010	0.6715
Random survival forest (cause 2)	0.5949	0.5522

the CIF perform better than any other method irrespective of the two measures of indices. The MF and CIF under classification tree performed equally and they are better than any other method in terms of prediction accuracy and area under the curve (AUC).

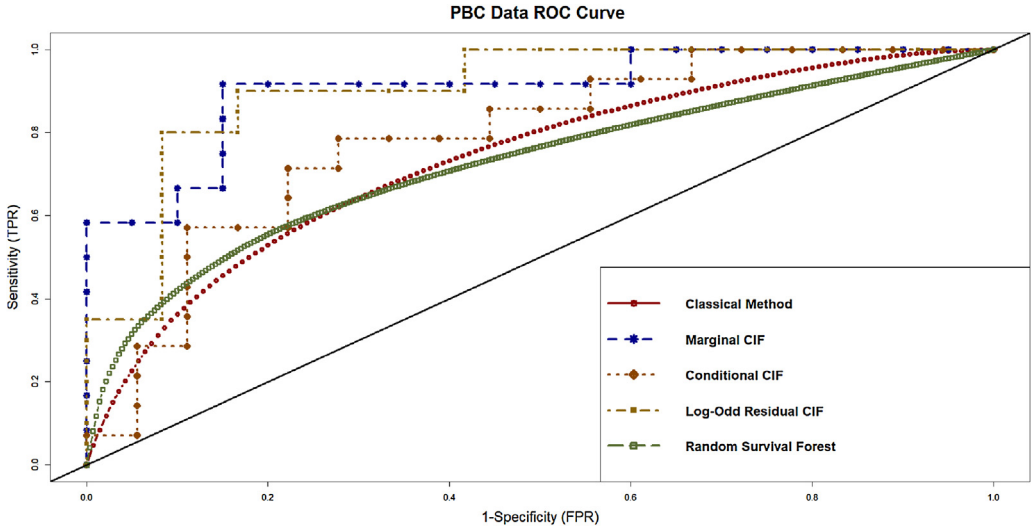

Fig. 2 – ROC values of various methods based on PBC data.

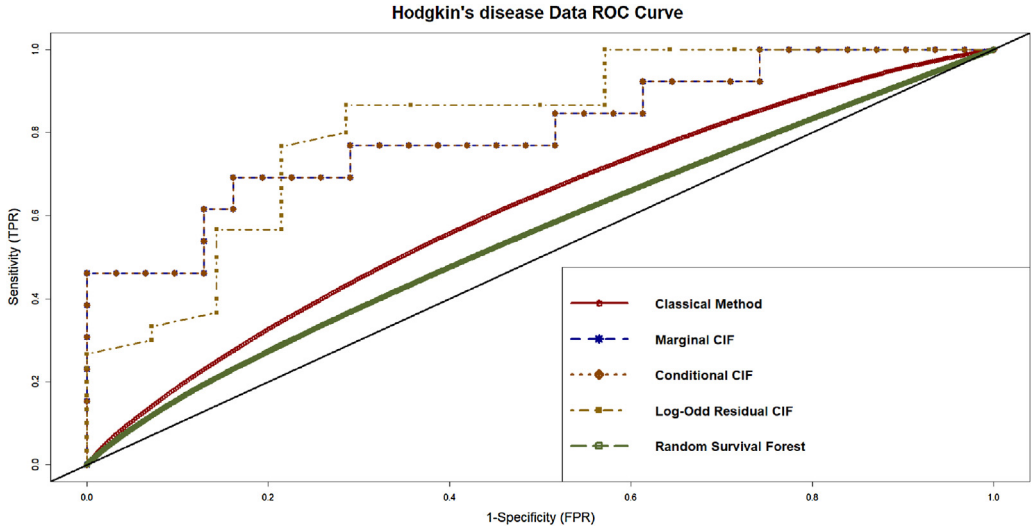

Fig. 3 – ROC values of various methods based on HD data.

8.3. *Further analysis*

We analyze the data to identify group of patients with maximum risk when experiencing the joint events simultaneously. The marginal probability of experiencing both the events based on the tree-structured is obtained from their prognostic risk factors. Eight terminal nodes are observed from the tree-structure (Fig. 4). The description of the eight terminal nodes are given below:

1. `Node 4: bili >= 2.15 and stage >= 3.5 and edema >= 0.25`
2. `Node 5: bili >= 2.15 and stage >=3.5 and edema < 0.25`
3. `Node 6: bili>= 2.15 and stage < 3.5`
4. `Node 9: bili < 2.15 and albumin < 3.535 and albumin < 3.535`
5. `Node 10: bili < 2.15 and albumin < 3.535 and albumin >= 3.535`

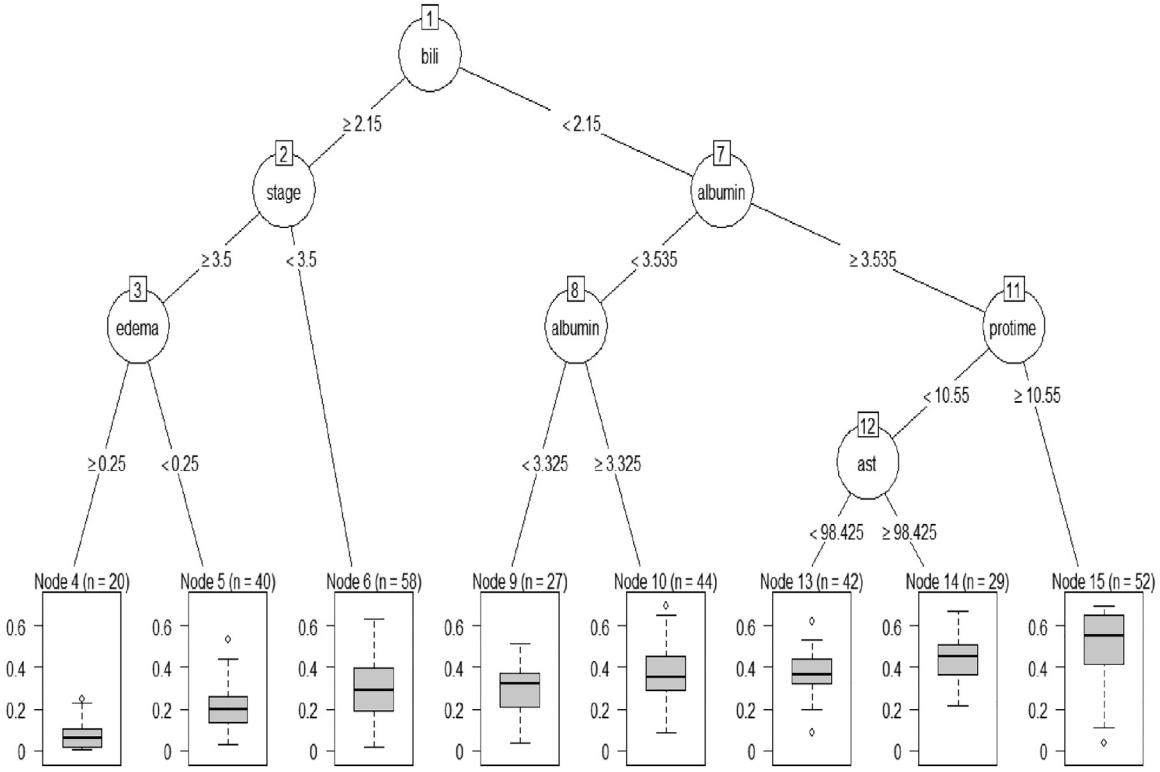

Fig. 4 – Tree-structured of death and transplant events.

Table 7 – Summary of the terminal nodes results of the liver transplant.

Terminal node	Sample size		Total	Median	Risk
	Censored	Joint events		Survival time	
4	1	19	20	299	II
5	8	32	40	1077	III
6	17	41	58	2288	IV
9	15	12	27	2847	V
10	32	12	44	–	I
13	37	5	42	–	I
14	23	6	29	4079	**VI**
15	35	17	52	–	I

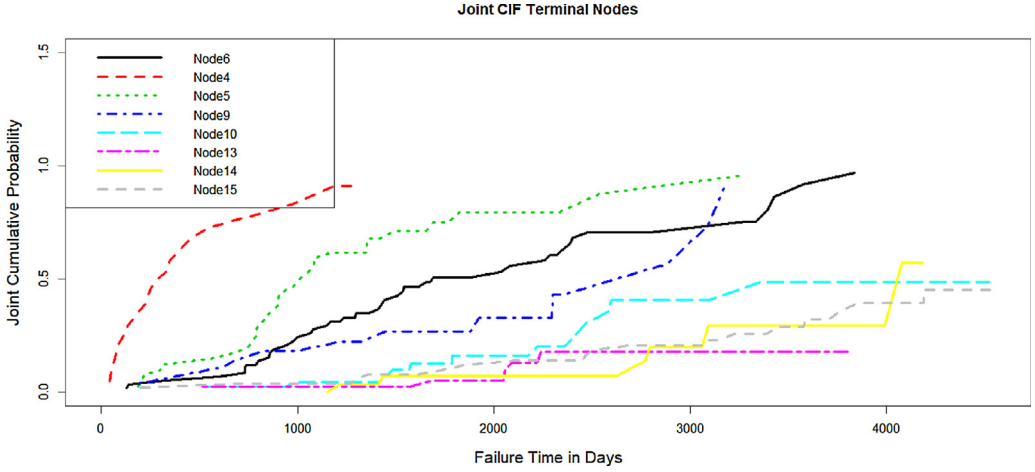

Fig. 5 – MF of death and transplant events for each terminal node.

6. Node 13: bili < 2.15 and albumin >= 3.535 and protime < 10.55 and ast < 98.425
7. Node 14: bili < 2.15 and albumin >= 3.535 and protine < 10.55 and ast >= 98.425
8. Node 15: bili < 2.15 and albumin >= 3.535 and protine >=10.55

This information is further summarized in Table 7, based on the median survival time of each terminal nodes of the tree and while Fig. 5, plots the MF of each terminal node. The results review that, group of nodes 15, 13 and 10 have the highest risk of dying before the transplant, follows by nodes 4, 5, 6 and 9. Finally, the group with the lowest risks are in node 14, and this suggests that the combination of low serum bilirubin (bili), high serum albumin (albumin), low in blood clotting time (protime) and low aspartate aminotransferase (ast) levels places patients at the low risk of death while on waiting list of liver transplant. Present work can be summarised using the concordance index [69] in Table 8. It has been found that the concordance index of the proposed method is higher with MF & RF for PBC data and RF in Hodgkin's disease data set. The major advantage of our study is that, unlike existing methods, we propose survival analysis with competing risk by considering the failure event simultaneously.

Table 8 – Concordance index

Techniques	Concordance Index (%)	
	PBC data	Hodgkin's disease data
MF	**90.4**	80.2
CIF	80.0	80.2
RF	**90.4**	**81.8**
Cause specific hazard	80.3	58.0
Random survival forest	89.0	67.5
Fine and gray	86.9	57.8
Conditional inference tree (CIT)	72.6	80.4
CoxBoost	85.4	73.5

9. Conclusions

In this study, we have developed splitting rules based on marginal cumulative function, conditional incidence function and log-odds residuals. The proposed SCART provide an understandable and clear interpretation of survival data with competing risks. Moreover, the proposed methods seem to provide a solution to simultaneous inference on survival

analysis with competing risk, without any stringent assumption on the distribution of the data and sub-distribution hazard assumptions as the case may be.

The three methods (MF, CIF and RF) and some other existing methods (Classical method and Random survival forest) are compared based on the two real-life data sets. The results of MSE and RMSE for the regression tree-structured shows that the proposed conditional incidence function perform better than the other techniques for both PBC and HD data. However, these results are strictly based on the two real-life data sets. Predictions accuracy and area under the curve are used to assess the performance of the methods based on the classification tree, and the results revealed that the marginal cumulative incidence function and conditional incidence function perform equally well on the HD data. Interestingly, the classical methods perform better than other methods for the PBC data with prediction accuracy, while the proposed methods are better with regard to AUC. This finding corroborates with the finding of [26]. They also pointed that the linear relation between covariate and the response variable may be another reason for the better performance of the classical method.

Acknowledgement

Dr. Dauda would like to thank the Department of Science and Technology (DST) and Ministry of External Affairs (MEA), Government of India (GoI) for awarding him C.V. Raman Doctoral Fellowship for African Researchers programme under the India-Africa Forum Summit, through the Federation of Indian Chambers of Commerce & Industry (FICCI).

REFERENCES

[1] Crowder M. Multivariate survival analysis and competing risks. Boca Raton, London New York: Chapman & Hall/CRC Press; 2012.
[2] Crowder M. Classical competing risks. 1st ed. Boca Raton, London New York: Chapman & Hall/CRC Press; 2001.
[3] Zhang Z. Survival analysis in the presence of competing risks. Ann Transl Med 2017;5(3):47 [pages 9].
[4] Gang L, Qing Y. Joint inference for competing risks survival data. J Am Stat Assoc 2016;111(515):1289–300.
[5] Cox D. Regression models and life-tables. J R Stat Soc Ser B (Methodol) 1972;34(2):187–220.
[6] Pepe MS, Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? Stat Med 1993;12(8):737–51.
[7] Bajorunaite R, Klein JP. Two-sample tests of the equality of two cumulative incidence functions. Comput Stat Data Anal 2007;51(9):4269–81.
[8] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999;94 (446):496–509.
[9] Chen BE, Kramer JL, Greene MH, Rosenberg PS. Competing risks analysis of correlated failure time data. Biometrics 2008;64(1):172–9.
[10] Webin L, Yu L. Analysis of competing risks data with missing cause of failure under additive hazards model. Stat Sin 2008;18:219–34.
[11] Haller B, Schmidt G, Ulm K. Applying competing risks regression models: an overview. Lifetime Data Anal 2013;19 (1):33–58.
[12] Gerds TA, Scheike TH, Andersen PK. Absolute risk regression for competing risks: interpretation, link functions, and prediction. Stat Med 2012;31(29):3921–30.
[13] Kundu D, Pradhan B. Bayesian analysis of progressively censored competing risks data. Sankhya B 2011;73(2): 276–96.
[14] Bhattacharya S, Pradhan B, Kundu D. Analysis of hybrid censored competing risks data. Statistics 2014;48(5): 1138–54.
[15] Zhang Z, Cortese G, Combescure C, Marshall R, Lee M, Lim HJ, et al. Overview of model validation for survival regression model with competing risks using melanoma study data. Ann Transl Med 2018;6(16):325 [pages 9].
[16] Gordon L, Olshen R. Tree-structured survival analysis. Cancer Treat Rep 1985;69(10):1065–9.
[17] Dauda KA, Yahya WB, Banjoko AW. Survival analysis with multivariate adaptive regression splines using cox-snell residual. Ann Comput Sci Ser 2015;13(2):25–41.
[18] Kriner M. Survival analysis with multivariate adaptive regression splines, Ph.D. thesis. Ludwig Maximilian University of Munich; 2007.
[19] Sparapani RA, Logan BR, McCulloch RE, Laud PW. Nonparametric survival analysis using Bayesian additive regression trees (bart). Stat Med 2016;35(16):2741–53.
[20] Segal MR. Regression trees for censored data. Biometrics 1988;44(1):35–47.
[21] LeBlanc M, Crowley J. Survival trees by goodness of split. J Am Stat Assoc 1993;88(422):457–67.
[22] Keles S, Segal MR. Residual-based tree-structured survival analysis. Stat Med 2002;21(2):313–26.
[23] Feng G, Amita KM, Shande C. Developing multivariate survival trees with a proportional hazards structure. J Data Sci 2006;4:343–56.
[24] Moradian H, Larocque D, Bellavance F. Survival forests for data with dependent censoring. Stat Methods Med Res 2019;28(2):445–61.
[25] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2(3):841–60.
[26] Ishwaran H, Gerds T, Kogalur U, Moore R, Gange SLBM. Random survival forests for competing risks. Biostatistics 2014;15(4):757–73.
[27] Ishwaran H, Kogalur UB. Random forests for survival, regression, and classification (RF-SRC), R package version 2.5.1; 2017, https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf.
[28] Ibrahim NA, Kudus A, Daud I, AbuBakar MR. Decision tree for competing risks survival probability in breast cancer study. Int J Biol Life Sci 2008;3(1):337–44.
[29] Fiona CM. Classification trees for survival data with competing risks, Ph.D. thesis. Department of Biostatistics, University of Pittsburgh; 2008.
[30] Lee C, Zame WR, Yoon J, van der Schaar M. Deephit: a deep learning approach to survival analysis with competing risks. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). 2018. pp. 2314–21.
[31] Bellot A, Schaar M. Tree-based Bayesian mixture model for competing risks. International Conference on Artificial Intelligence and Statistics. 2018. pp. 910–8.
[32] Wu J, Tan Y, Chen Z, Zhao M. Decision based on big data research for non-small cell lung cancer in medical artificial system in developing country. Comput Methods Prog Biomed 2018;159:87–101.
[33] Han S, Andrei A-C, Tsui K-W. Multiple imputation for competing risks survival data via pseudo-observations. CSAM (Commun Stat Appl Methods) 2018;25(4):385–96.

[34] Scrutinio D, Guida P, Passantino A, Ammirati E, Oliva F, Lagioia R, et al. Application of competing risks analysis improved prognostic assessment of patients with decompensated chronic heart failure and reduced left ventricular ejection fraction. J Clin Epidemiol 2018;103:31–9.
[35] Kretowska M. Piecewise-linear criterion functions in oblique survival tree induction. Artif Intell Med 2017;75:32–9.
[36] Linden A, Yarnold PR. Modeling time-to-event (survival) data using classification tree analysis. J Eval Clin Pract 2017;23(6):1299–308.
[37] Huebner M, Wolkewitz M, Enriquez-Sarano M, Schumacher M. Competing risks need to be considered in survival analysis models for cardiovascular outcomes. J Thorac Cardiovasc Surg 2017;153(6):1427–31.
[38] Hsu JY, Roy JA, Xie D, Yang W, Shou H, Anderson AH, et al. Statistical methods for cohort studies of ckd: survival analysis in the setting of competing risks. Clin J Am Soc Nephrol 2017;12(7):1181–9.
[39] Alaa AM, van der Schaar M. Deep multi-task Gaussian processes for survival analysis with competing risks. Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc.; 2017. p. 2326–34.
[40] Xu W, Che J, Kong Q. Recursive partitioning method on competing risk outcomes. Cancer Inform 2016;15:CIN–S39364.
[41] Scrutinio D, Guida P, Ammirati E, Oliva F, Frigerio M. Long-term prognostic implications of the adhf/nt-probnp risk score in patients admitted with advanced heart failure. J Heart Lung Transplant 2016;35(10):1264–7.
[42] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 2006;15(3):651–74.
[43] Bellot A, van der Schaar M. Multitask boosting for survival analysis with competing risks. Advances in Neural Information Processing Systems. 2018. pp. 1397–406.
[44] Lawless JF. Statistical models and methods for lifetime data. 2nd ed. New York: Wiley; 2003.
[45] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999;18(17–18):2529–45.
[46] Breiman L, Friedman J, Olshen R, Stone CS. Classification and regression trees. Boca Raton, London, New York: Chapman & Hall/CRC Press; 1984.
[47] Haddad K, Rahman A, Zaman MA, Shrestha S. Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. J Hydrol 2013;482:119–28.
[48] Haque MM, Hadda K, Rahman A, Hossain M, Hagere D, Kabria G. Long term demand forecasting: use of monte carlo cross validation for the best model selection. 20th International Congress on Modelling and Simulation; 2013. pp. 2311–7.
[49] Konovalov DA, Sim N, Deconinck E, Vander HY, Coomans D. Statistical confidence for variable selection in QSAR models via Monte Carlo cross-validation. J Chem Inf Model 2008;48(2):370–83.
[50] Mosteller F, Tukey JW. Data analysis and regression: a second course in statistics. Reading, MA: Addison-Wesley; 1977.
[51] Efron B, Tibshirani R. Improvements on cross-validation: the 0.632+ bootstrap method. J Am Stat Assoc 1997;92(438):548–60.
[52] Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemom 2004;18(2):112–20.
[53] Marter H, Martern M. Multivariate analysis of quality: an introduction. Chichester: John Wiley & Sons Ltd.; 2001.
[54] Picard RR, Cook RD. Cross-validation of regression models. J Am Stat Assoc 1984;79(387):575–83.
[55] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.
[56] Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J 2008;50(4):457–79.
[57] Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biom J 2006;48(6):1029–40.
[58] Gerds TA, Schumacher M. Efron-type measures of prediction error for survival analysis. Biometrics 2007;63(4):1283–7.
[59] Uno H, Cai T, Tian L, Wei L. Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc 2007;102(478):527–37.
[60] Choi I, Wells BJ, Yu C, Kattan MW. An empirical approach to model selection through validation for censored survival data. J Biomed Inform 2011;44(4):595–606.
[61] Porzelius C, Schumacher M, Binder H. The benefit of data-based model complexity selection via prediction error curves in time-to-event data. Comput Stat 2011;26(2):293–302.
[62] Porzelius C, Binder H. peperr:parallelised estimation of prediction error, R package version 1; 2010;1–7, https://cran.r-project.org/web/packages/peperr/peperr.pdf.
[63] Porzelius C, Schumacher M, Binder H. A general, prediction error-based criterion for selecting model complexity for high-dimensional survival models. Stat Med 2010;29(7–8):830–8.
[64] Porzelius C, Binder H, Schumacher M. Parallelized prediction error estimation for evaluation of high-dimensional models. Bioinformatics 2009;25(6):827–9.
[65] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56(2):337–44.
[66] Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61(1):92–105.
[67] Fleming TR, Harrington DP. Counting processes and survival analysis. New York: Wiley; 1991.
[68] Pintilie M. Competing risks: a practical perspective. West Sussex: John Wiley and Sons; 2006.
[69] Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med 2013;32(13):2173–84.