Recursive Partitioning Method on Competing Risk Outcomes

Wei Xu[1,2], Jiahua Che[1,3] and Qin Kong[1,3]

[1]Department of Biostatistics, Princess Margaret Cancer Centre, Toronto, ON, Canada. [2]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. [3]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada.

Supplementary Issue: Integrative Analysis of Cancer Genomic Data

ABSTRACT: In some cancer clinical studies, researchers have interests to explore the risk factors associated with competing risk outcomes such as recurrence-free survival. We develop a novel recursive partitioning framework on competing risk data for both prognostic and predictive model constructions. We define specific splitting rules, pruning algorithm, and final tree selection algorithm for the competing risk tree models. This methodology is quite flexible that it can corporate both semiparametric method using Cox proportional hazards model and parametric competing risk model. Both prognostic and predictive tree models are developed to adjust for potential confounding factors. Extensive simulations show that our methods have well-controlled type I error and robust power performance. Finally, we apply both Cox proportional hazards model and flexible parametric model for prognostic tree development on a retrospective clinical study on oropharyngeal cancer patients.

KEYWORDS: competing risk outcomes, survival tree model, recursive partitioning algorithm, Cox proportional hazards model, parametric competing risk model, prognostic and predictive effect, clinical cancer outcomes

SUPPLEMENT: Integrative Analysis of Cancer Genomic Data

CITATION: Xu et al. Recursive Partitioning Method on Competing Risk Outcomes. *Cancer Informatics* 2016:15(S2) 9–16 doi: 10.4137/CIN.S39364.

TYPE: Methodology

RECEIVED: March 15, 2016. **RESUBMITTED:** June 22, 2016. **ACCEPTED FOR PUBLICATION:** July 03, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 560 words, excluding any confidential comments to the academic editor.

FUNDING: Authors disclose no external funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: wxu@uhnres.utoronto.ca

Introduction

In clinical research, prognostic and predictive models play important roles. Prognostic model is constructed to predict patients' clinical outcomes, given their demographic, clinical, and genetic characteristics. It is important to define interpretable prognostic classification rules for understanding the nature of patients' performance. On the other hand, predictive model is developed for treatment decision-making, such as optimization of treatment strategy for patients with specific characteristics.

In cancer research, researchers usually deal with the time-to-event clinical endpoints, such as standard censored survival outcomes. Typical example is the overall survival. Sometimes the research interests are on more complicated type of data, such as recurrence-free survival, which involves competing risk events (ie, death without tumor recurrence). A competing risk event can preclude the event of interest from occurring. For example, we could not observe tumor recurrence if a patient died without recurrence. There is loss of information if we ignore this type of event. Thus, it is important to pay attention to these complex time-to-event data as a way of understanding *real-world* cancer outcomes.

First introduced by Morgan and Sonquist,[1] the tree-based method became widely used due to the work by Breiman et al.[2], who developed the Classification and Regression Tree algorithm. An attractive feature of the tree-based model is the connection between partitioning a covariate space and a binary decision-making process. The methodology was further developed for both continuous and categorical outcomes, adjusting for the effect of confounders.[3] For standard censored survival data, we have developed a general framework to create prognostic and predictive survival trees for time-to-event outcomes based on recursive partitioning algorithm.[4] This framework also allows for adjusting for possible clinical confounders, which are not of direct interest. The tree-based methods are appealing as the flexibility to detect direct or interactive effects on survival outcomes and select covariates in the presence of high-dimensional data. Simulation results show the well performance of the method and robustness to large-dimensional covariate spaces for time-to-event data.[4] However, currently, little work has been applied on competing risk outcomes for prognostic and predictive model development.

To model competing risk events, the most commonly used method is based on the Cox proportional hazards models. While some argue that in large epidemiological studies, the assumption of proportional hazards is sometime problematic. A flexible parametric model[5] and a parametric mixture model[6] are developed to model the cause-specific hazard function to corporate time-dependent hazards. Another method is

developed to model the cumulative incidence function directly based on pseudovalues.[7]

Several decision tree models were developed for survival data,[8,9] but only few were found for competing risk outcomes. One work on competing risk survival data generates prognostic survival tree only,[10] and all the input covariates are potential splitting variables. It cannot adjust for potential confounders. As we know, there is no existing method that is applicable for predictive tree construction that can deal with treatment interactive effect on competing risk outcomes.

The primary aim of this article is to extend our current survival tree framework to competing risk data for both prognostic and predictive tree constructions. We focus on the semiparametric method using Cox proportional hazards algorithm and use likelihood ratio test (LRT) for the splitting rule. This novel tree framework is quite flexible that it can also corporate parametric competing risk model and overcome the restriction of proportional hazards assumption.

In this article, first we introduce the basic structures of a recursive partitioning algorithm of competing risk decision trees. In addition, we define the splitting rules, pruning algorithm, and methodology to choose the final tree structure on the competing risk outcomes. Extensive simulations are conducted to evaluate the performance of this innovative methodology framework for both prognostic and predictive trees. To deal with time-dependent hazards, we develop a method to combine the flexible parametric model[5] and the semiparametric survival tree framework. Finally, we apply both Cox proportional hazards model and flexible parametric model for prognostic tree construction on an oropharyngeal cancer study.

Method

Algorithm overview. The tree-based method using a recursive partitioning procedure consists of a splitting rule, a pruning algorithm, and an approach to select the final tree structure. The splitting rule is used to partition covariate space into subgroups representing patient prognosis or prediction.

The partition is represented as a tree T, with terminal nodes $\tilde{T}$ corresponding to the partition of the covariate space into $|\tilde{T}|$ subsets. It is applied recursively until there are very few patients in each group or a prespecified number of groups are created. For competing risk outcomes, different algorithms can be applied for splitting rule constructions, such as LRT, log-rank test, and Gray's test.[11] We developed a likelihood ratio-based test and used as the splitting rule. This also implies the flexibility of the tree framework. It can solve various problems with different likelihood constructions. The large tree created by the splitting rule usually has the problem of overfitting and performs poorly out of samples. Thus, pruning is necessary to search and find the optimal subtree structure. The final subtree structure is then selected using a resampling algorithm.

Competing risk method based on Cox model. *Splitting rule.* Competing risks model are developed when an event can be caused by multiple reasons and interest lies in modeling one particular cause.[12,13] There are two key measures in competing risks analysis, cause-specific hazard, and hazard of the subdistribution, which differ in the risk sets by definition. Several regression approaches are applicable to estimate these two quantities.[14] In this article, we focused on modeling cause-specific hazard. For simplicity, in this article, we only consider making binary splits on binary splitting variables.

To partition a node h, we need to find the split s such that some measure of the improvement $G(s, h)$ with or without this split is maximized.

$$G\left(s^*, h\right) = \max_{s \in S_h} G\left(s, h\right)$$

where S_h is the set of all the possible splits that can be made at node h. s^* represents the best split for node h with the maximum measure of improvement.

Data setup for competing risks model is $(y_i, \mathbf{x}_i, \delta_i)$. For observation i, $\mathbf{x}_i$ is the covariate vector and y_i is the time-to-event outcome. $\delta_I = 0$ if censored, $\delta_i = 1$ if an event is of interest, and $\delta_i = 2, 3, \ldots$ indicating other events that are competing risks. Compared to standard censored survival data, the major difference is on the definition of censoring indicator.

We can also define the cause-specific hazard rate for cause j,

$$\lambda_j(t,x) = \lim_{dt \to 0} \frac{P(t < T < t + dt, J = j \,|\, T > t, x)}{dt}$$

Then, the cumulative hazard for cause j is

$$H_j(t,x) = \int_0^t \lambda_j(u,x)du.$$
$$f_j(t,x) = \lambda_j(t,x)S(t,x)$$

holds for each cause j.

The likelihood involving a specific type of failure is exactly the same as the likelihood obtained by treating all other types of failures as censored observations. Thus, the full likelihood is the product of likelihood of each specific cause failure. Here, we assume

$$\lambda_j(t|x) = \lambda_{j0}(t)e^{\beta_j' x}$$

The partial likelihood is

$$L(\beta) = \prod_{j=1}^{m} \prod_{i=1}^{k_j} \frac{e^{\beta_j' x_{ji(j)}}}{\sum_{k \in R(t_{ji})} e^{\beta_j' x_{jk}}}$$

where k_j is the number of distinct times of event due to cause j, t_{ji} denotes the ith such time, $R(t_{ji})$ is the risk set at time

t_{ji}, and $i(j)$ is the index of the event that happened at t_{ji}. We maximize log-likelihood $l(\beta)$ to obtain the MLE of parameters β_j's.

Consider two nested models, $m_0 : \log\lambda = \beta_0' x_0$ and $m_1 : \log\lambda = \beta_0' x_0 + \beta_1' x_1$. The LRT statistic corresponding to the hypothesis test

$$H_0 : \beta_1 = 0$$

is

$$-2(l_{m_0}(\hat{\beta}_{m_0}) - l_{m_1}(\hat{\beta}_{m_1})) \sim \mathcal{X}^2_{rank(x1)}$$

Here, l_{m0} and l_{m1} are the log-likelihood of the two nested models. In order to adapt the competing risk model to the survival tree structure, the constructed LRT can be regarded as $G(T)$, the goodness of split of tree T.

To specify, let vector $\boldsymbol{x}^0$ be the covariates that needs to be adjusted for confounding effects, $\boldsymbol{x}^1$ the true splitting covariates, and x^t the treatment. We define two splitting rules that are used to create adjusted prognostic (marginal) trees and adjusted predictive (interactive) trees, respectively.

For prognostic tree, the null model is $\log\lambda_j = \beta_j^{0'} x_j^0$ and the alternative is $\log\lambda_j = \beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1$. j represents a specific cause of the event that is of interest. While for competing risks, we assume $\log\lambda_l = \beta_l^{0'} x_l^0$. The LRT statistic corresponding to $H_0 : \beta_j^1 = 0$ is defined as the split complexity $G_m(s, h)$.

$$G_m(s,h) = -2\left(l_{m_0} - l_{m_1}\right)$$

$$l_{m_0} = \sum_{l=1}^{m}\sum_{i=1}^{k_l}(\beta_l^{0'} x_l^0 - \log \sum_{k\in R(t_{li})} e^{\beta_l^{0'} x_{lk}^0})$$

$$l_{m_1} = [\sum_{i=1}^{k_j}(\beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 - \log \sum_{k\in R(t_{ji})} e^{\beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1})] + \sum_{l=1, l\neq j}^{m}\sum_{i=1}^{k_l}(\beta_l^{0'} x_l^0 - \log \sum_{k\in R(t_{li})} e^{\beta_l^{0'} x_{lk}^0})$$

For predictive tree, the null model is $\log\lambda_j = \beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 + \beta_j^{t'} x_j^t$ and the alternative is $\log\lambda_j = \beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 + \beta_j^{t'} x_j^t + \beta_j^{i'} x_j^t * x_j^1$. For competing risk events, we assume $\log\lambda_l = \beta_l^{0'} x_l^0$. The LRT statistic corresponding to $H_0 : \beta_j^i = 0$ is defined as the split complexity $G_i(s, h)$.

$$G_i(s,h) = -2\left(l_{m_0} - l_{m_1}\right)$$

$$l_{m_0} = \sum_{l=1}^{m}\sum_{i=1}^{k_l}(\beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 + \beta_j^{t'} x_j^t - \log \sum_{k\in R(t_{li})} e^{\beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 + \beta_j^{t'} x_j^t})$$

$$l_{m_1} = [\sum_{i=1}^{k_j}(\beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 + \beta_j^{t'} x_j^t + \beta_j^{i'} x_j^t * x_j^1 - \log \sum_{k\in R(t_{ji})} e^{\beta_j^{0'} x_j^0 + \beta_j^{1'} x_j^1 + \beta_j^{t'} x_j^t + \beta_j^{i'} x_j^t * x_j^1})] + \sum_{l=1, l\neq j}^{m}\sum_{i=1}^{k_l}(\beta_l^{0'} x_l^0 - \log \sum_{k\in R(t_{li})} e^{\beta_l^{0'} x_{lk}^0})$$

Pruning algorithm. The split complexity $G_\alpha(T)$ is defined as

$$G_\alpha(T) = G(T) - \alpha|S|$$

where $S = T - \tilde{\tilde{T}}$ is the set of internal nodes of tree T; $|S|$ is the cardinality of S; $\alpha \geq 0$ is the complexity parameter; and $G(T)$, the goodness of split of tree T, is the sum of the splitting statistics over the full tree.

$$G(T) = \sum_{h\in S} G(h)$$

We can interpret $G(T)$ as how well the prognostic or predictive tree structure fit in the data. LRT, log-rank test, or other standardized statistical distance measures can be used for such measurement. Here, we use the LRT statistics. From the above definition, if α is small, the penalty on the tree size is small and the tree with large tree size has large split complexity. On the contrary, if α gets larger, the final tree size will be smaller. This is a trade-off between the tree size and goodness of split of the tree structure.

The idea of the pruning algorithm is to cut the branches that have the weakest link to the tree. Statistically, it is the improvement of the overall split complexity. Performing this cut once at a time, a nested sequence of subtrees $T_m \prec \ldots \prec T_k \prec \ldots \prec T_0$ are obtained where T_m is the root node and corresponding complexity parameters $\infty > \alpha_m > \ldots > \alpha_k > \ldots > \alpha_1 > \alpha_0 = 0$.[15]

Selection algorithm of final tree. The previous section yields a sequence of optimally pruned subtrees. In this section, we aim to select a final tree structure for decision-making. Since the tree structures are determined by maximizing LRT statistics, the split complexity would be larger than expected with the same training sample.

An effective method to deal with this issue is to randomly split the data into two sets, namely, a training set and a test set, and repeat the process multiple times. To implement this, bootstrap method is applicable. We first grow and prune a tree with the training set and then force test set data into the sequence of the pruned trees. The split complexity $G(h)$ can be calculated for each internal node h using the test sample. The best pruned subtree is chosen with the maximum split complexity. For simplicity, we recommend using the penalty $\alpha = 4$ since it is similar to the 0.05 significance level for a single split.[15]

Flexible parametric model. In cancer epidemiological studies, sometimes the proportional hazards assumption is not relevant for each specific cause. This leads us to explore a more general method to overcome this issue. The use of parametric model may have some advantages. Efron[16] and Oakes[17] showed that, under certain circumstances, parametric models result in more efficient parameter estimation than Cox's model. With decreasing sample sizes, parametric models may have better performance in terms of efficiency. When empirical information is sufficient, parametric models can provide some insight into the shape of the baseline hazard. In addition, it does not have the restriction of proportional hazards assumption, thus is easier to deal with time-dependent effects. It gives an estimation of the baseline hazard, and the visualization of the hazard function is much easier.[18] Royston and Parmar[18] also proposed an extension of Weibull and log-logistic model using cubic splines to smooth the baseline log cumulative hazard. Hinchliffe and Lambert[5] further applied this flexible parametric model to competing risk framework. In this article, we incorporate Hinchliffe and Lambert model into our survival tree framework and applied on a real clinical study.

Flexible parametric model. The basic idea of flexible parametric model starts from the Weibull distribution,[5,19,20]

$$S(t) = \exp\left(-\mu t^{\gamma}\right).$$

If we take a complementary log-log transformation, we have

$$\ln H(t) = \ln \mu + \gamma \ln t$$

In this setting, ln $H(t)$ is a linear function of ln t, it can be generalized to use restricted cubic spline for the estimation. Introducing linear combination of covariates to estimate ln μ, we have the following,

$$\ln H(t) = x^T \beta + s(\ln t | \gamma, n_0)$$

$s(\ln t|\gamma, n_0) = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_{N-1} z_{N-1}$ is the restricted cubic spline function, which is a function of ln t. $\gamma_0, \gamma_1, \ldots, \gamma_{N-1}$ are parameters and n_0 is a vector of N knots. z_i's are functions of ln t and the knots. Then, maximize the likelihood to obtain the estimation of parameters γ_i and β.

With the estimation of ln $H(t)$, it is easy to calculate the following,

$$S(t) = \exp\left(-H(t)\right)$$
$$\lambda(t) = \frac{ds(\ln t|\gamma, n_0)}{dt} \exp \ln H(t)$$

Likelihood construction. Based on flexible parametric survival tree model, we construct a likelihood function to be used as a splitting rule within the survival tree framework to apply to the competing risk data. The likelihood function can be written as the product of k distinct likelihood for each failure cause. Thus,

$$L = \prod_{j=1}^{k} L_j = \prod_{j=1}^{k} \prod_{i=1}^{n} \lambda_j(y_i)^{\delta_{ij}} S_j(y_i)$$

i represents the index of subject and j the index of failure cause. $\lambda_j(t_i)^{\delta_{ij}}$ and $S_j(t_i)$ can be calculated for each specific failure cause according to the previous section. Hence, the likelihood function is estimable.

Time-dependent hazards. The flexible parametric model can be adapted to time-dependent hazards by adding the interaction terms between covariates and the restricted cubic spline function.[5] Suppose there are D time-dependent effects, then

$$\ln H(t) = x^T \beta + s(\ln t | \gamma, n_0) + \sum_{j=1}^{D} s(\ln t | \alpha_j, n_j) x_j$$

At each splitting time, to select a best splitting covariate, we compare the improvement of the following two models,

$$m_0 : \ln H(t) = s(\ln t | \gamma, n_0),$$

and

$$m_1 : \ln H(t) = x^T \beta + s(\ln t | \gamma, n_0) + s(\ln t | \alpha, n) x_k,$$

x_k represents a potential time-dependent effect.

By doing this, no matter this potential effect is time dependent or not, there will be an improvement on the test statistics and our algorithm will capture this effect if it is truly related to the outcome.

Simulation settings. Extensive simulations are conducted for the competing risk tree-based model performance evaluation.

Simulation parameters. For the simulation parameters, first, we choose different sets of the sample size and the number of covariates for potential splitting. We assume that all the splitting variables are binary and generated from Bernoulli distributions with parameter p, $x \sim B(n, p)$. The true covariate x_1 among the splitting variables associated with the outcome is set for different hazard ratios. A continuous confounding variable x_c is generated from normal distribution $N(0, 1)$ and its hazard ratio is set to be 0.5. All the simulations are run adjusting for this confounder. For splitting algorithm, maximum number of splits and minimum number of events are set to 10 and 20, respectively, in advance. The parameter settings for each model are shown in Table 1.

Simulate time to event. There are two types of time to event. One is time to event that is of interest t_1 and the other is time to competing risk t_2. Both of them are simulated to

Table 1. Parameter settings for different tree models.

TREE TYPE	SAMPLE SIZE N	SPLITTING COVARIATES	*P*	HAZARD RATIO
Null tree	500	10,50,100	0.6	1
Prognostic (Marginal)	500	10,20,50,100	0.6	2.0,2.5,3.0
	500	10	0.1,0.2,0.3	2.5
Predictive (Interactive)	1000	10,20	0.6	2.0,2.5,3.0

follow exponential distribution with different prespecified hazard ratios. These hazard ratios are determined by the type of trees and the true related variables.

For example, the log hazard ratio for event of interest assumes to be $\log h_1 = x_c \log 0.5 + x_1 \log \beta$, β is the hazard ratio for x_1. The log hazard ratio for competing risk is $\log h_2 = x_c \log 0.5$. t_1 is generated from $T_1 \sim \text{Exp}(h_1)$ and t_2 is generated from $T_2 \sim \text{Exp}(h_2)$.

Simulate censoring time. Censoring times c are generated from uniform distribution on $(0, \sigma^* T_{\max})$. $T_{\max}$ is the maximum of t_1 in the previous step. σ can be chosen to control the censoring rate. In our simulation, σ is set to 0.8.

Simulated response time. Then, we take the minimum of t_1, t_2, and c as the response time for each individual. The status of each individual is defined according to the minimum time.

After generating the data from above, we then use our tree model to fit the data.

Results

Simulation results. The simulation results for both prognostic and predictive tree models are presented for this study.

Prognostic tree. Table 2 shows the model performance under the null model for prognostic tree for type I error evaluation. A total of 500 samples are simulated for each replication, and 1000 replicates have been conducted. The chance of selecting the wrong tree slightly increases as the number of splitting covariates becomes larger, but the model still performs well with only a 1.2%–4.3% chance of having wrong tree.

Tables 3–5 show the power performances under alternative model with different hazard ratios. Increasing the splitting covariates will reduce the chance of selecting the right tree. The model performs better when the effect size is larger. When the hazard ratio is 3.0, the probability of selecting the right tree remains high as the number of potential splits increases. Even for 100 potential splits, the probability of identifying the true tree as an optimal subtree remains 89% and the chance of selecting the correct tree is 66%.

Table 6 shows that the model has better performance when the true splitting variable is more balanced. With proportion of the splitting variable (p) to be 0.3 and 0.6, 91% of chances will the true tree be identified as an optimal subtree and 87% and 91% chances of selecting the correct tree, respectively. For highly unbalanced variable with $P = 0.1$, the chance of selecting correct tree is only 59%.

Predictive tree. For predictive tree, Table 7 shows the performance under null hypothesis and Table 8 focuses on the effect of number of splitting covariates and effect size under alternative. Similar patterns can be found as in prognostic

Table 2. Tree performance under null hypothesis for prognostic tree.

SAMPLE SIZE N	*P*	SPLITTING COVARIATES	TYPE I ERROR
500	0.6	10	0.012
500	0.6	50	0.036
500	0.6	100	0.043

Table 3. Tree performance under alternative model for prognostic tree HR = 2.0.

N	SPLITTING COVARIATES	HAZARD RATIO	PROPORTION OF TREES HAVE X_1 AS THE FIRST SPLIT	PROPORTION OF TREES ONLY HAVE X_1 AS THE FIRST SPLIT (CORRECT TREES)
500	10	2.0	0.63	0.57
	20	2.0	0.43	0.34
	50	2.0	0.38	0.17
	100	2.0	0.29	0.02

Table 4. Tree performance under alternative model for prognostic tree HR = 2.5.

N	SPLITTING COVARIATES	HAZARD RATIO	PROPORTION OF TREES HAVE X_1 AS THE FIRST SPLIT	PROPORTION OF TREES ONLY HAVE X_1 AS THE FIRST SPLIT (CORRECT TREES)
500	10	2.5	0.91	0.87
	20	2.5	0.83	0.74
	50	2.5	0.68	0.40
	100	2.5	0.51	0.15

Table 5. Tree performance under alternative model for prognostic tree HR = 3.0.

N	SPLITTING COVARIATES	HAZARD RATIO	PROPORTION OF TREES HAVE X_1 AS THE FIRST SPLIT	PROPORTION OF TREES ONLY HAVE X_1 AS THE FIRST SPLIT (CORRECT TREES)
500	10	3.0	1.00	0.99
	20	3.0	0.98	0.95
	50	3.0	0.94	0.85
	100	3.0	0.89	0.66

Table 6. Tree performance under alternative hypothesis for prognostic tree.

N	P	SPLITTING COVARIATES	HAZARD RATIO	PROPORTION OF TREES HAVE X_1 AS THE FIRST SPLIT	PROPORTION OF TREES ONLY HAVE X_1 AS THE FIRST SPLIT (CORRECT TREES)
500	0.1	10	2.5	0.59	0.49
	0.2	10	2.5	0.80	0.77
	0.3	10	2.5	0.91	0.87
	0.4	10	2.5	0.91	0.91

Table 7. Tree performance under null hypothesis for predictive tree.

SAMPLE SIZE N	P	SPLITTING COVARIATES	TYPE I ERROR
500	0.6	10	0.014
500	0.6	50	0.051
500	0.6	100	0.077

tree results. The model performance is better when there are smaller number of split variables and stronger effect size.

When effect size is greater than 2.5, there is over 91% chance of predictive tree model selecting the true tree with 20 potential split variables. If there are only 10 potential split variables, the chance will be increased to over 96%.

Application to clinical data. We applied the competing risk tree model on a Human papillomavirus positive (HPV+) oropharyngeal cancer (OPC) study with 573 patients.[21] The endpoint of interest is recurrence-free survival with death without recurrence defined as competing risk. Potential splitting covariates are age, gender, smoking pack-year, alcoholism, T stage, and N stage. For this retrospective data, prognostic tree model is constructed with treatment been adjusted for confounding effect. Appling both Cox-based model and flexible parametric model[22] to this cancer clinical data, we obtain the following tree structures for outcome prognoses, see Figures 1 and 2.

Table 8. Tree performance under alternative for predictive (interactive) tree.

N	P	SPLITTING COVARIATES	HAZARD RATIO	PROPORTION OF TREES HAVE X_1 AS THE FIRST SPLIT	PROPORTION OF TREES ONLY HAVE X_1 AS THE FIRST SPLIT (CORRECT TREES)
1000	0.4	10	2.0	0.83	0.77
	0.4	10	2.5	1.00	0.96
	0.4	10	3.0	1.00	0.97
	0.4	20	2.0	0.69	0.54
	0.4	20	2.5	0.99	0.91
	0.4	20	3.0	1.00	0.96

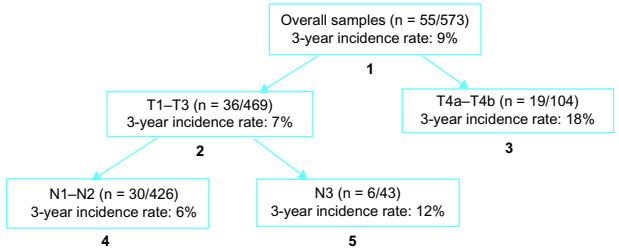

Figure 1. Tree structure using Cox-based method. For each subgroup, *n* indicates the number of event/sample size, and incidence rate represents three-year cumulative incidence rate of recurrence-free survival. Splitting covariate is indicated within each node. The number under each node identifies each subgroup.

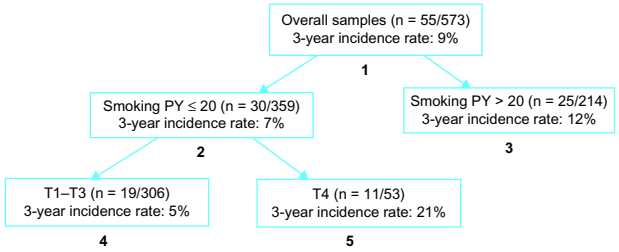

Figure 2. Tree structure using flexible parametric model. For each subgroup, *n* indicates the number of event/sample size, and incidence rate represents three-year cumulative incidence rate of recurrence-free survival. Splitting covariate is indicated within each node. The number under each node identifies each subgroup.

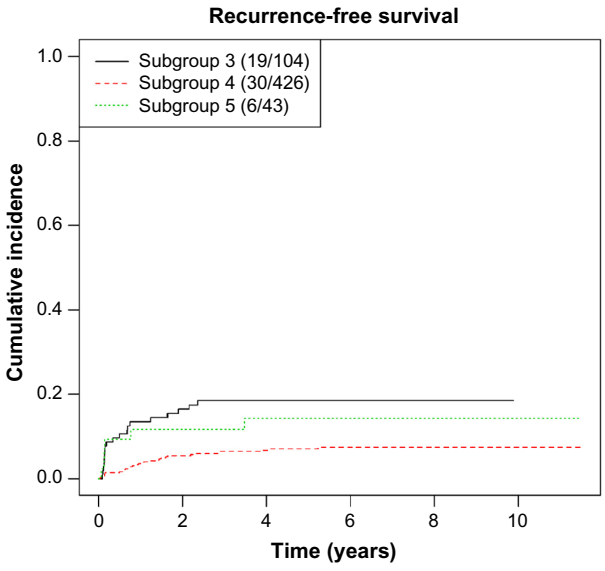

Figure 3. Cumulative incidence curves for each subgroup with Cox-based method.

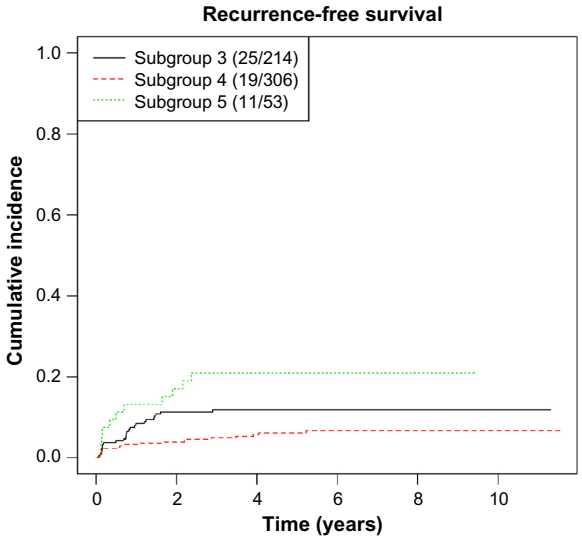

Figure 4. Cumulative incidence curves for each subgroup with flexible parametric model.

Both methods provide final tree structures after splitting, pruning, and final tree selection. Since the two methods are based on different modeling assumptions, they perform differently on the choice of splitting variables and tree structures selection. However, both of them capture the important factors that are associated with the outcomes. They both provide a good separation of the patients into prognostic groups, see Figures 3 and 4. We also apply multivariable analysis on the splitting covariates. Table 9 shows that tumor size (T stage) and nodal status (N stage) are the important prognostic factors to be significantly related to the competing risk outcome, and smoking is moderately significant. These factors are chosen as splitting covariates in both trees.

Table 9. Multivariate analysis results for HPV+ OPC data with 573 patients.

COVARIATE	HR (95%CI)	GLOBAL *P*-VALUE
Tx Regimen		**<0.001**
CRT	reference	
RT alone	2.9 (1.54,5.46)	
Age		0.57
<70	reference	
≥70	0.8 (0.36,1.74)	
T		**<0.001**
T1/T2/T3	reference	
T4ab	2.8 (1.52,5.13)	
N		**0.0018**
N0/N1/N2	reference	
T4ab	3.33 (1.57,7.06)	
Smoking PY		0.83
≤20	reference	
>20	1.07 (0.6,1.88)	

Conclusion and Discussion

In this article, we develop a novel survival tree framework on competing risk outcomes. This innovative method can deal with both prognostic and predictive models, which is important for cancer clinical research. This method fills the gap of current tree model development on clinical time-to-event outcomes. We define specific splitting rules, pruning algorithm, and final tree selection algorithm for this competing risk tree model. Both prognostic and predictive tree models are developed to adjust for potential confounding factors.

Extensive simulations show that the performance of our methods is well controlled under the null hypothesis. This performance is quite robust with a large number of potential splitting variables, which is important for many cancer pharmacogenomics research studies with high-dimensional biomarker space. Moreover, we have shown that the interaction survival tree can perform well with the large number of genetic factors often found in personalized medicine research. Once a tree is created and subgroups are identified, summary statistics such as hazard ratios of treatment, Kaplan–Meier curves, and median survival times for each group can be presented to clinicians. The clinicians can make treatment decision based on the predictive tree results.

Simulations have shown that the power of selecting the right tree structure under the alternative hypothesis is usually high. For predictive tree, to have adequate power, there should be a sufficiently large number of events, interactive effect between the split and treatment, and the balance of the potential splits. In addition, adjusting for clinical confounders in the splitting rule seems to have statistical benefits on most of the cancer clinical studies with potential outcome-related clinical factors such as age, tumor stage, and smoking status.

Our methodology is quite flexible that it can corporate both semiparametric method using Cox proportional hazards model and parametric competing risk model. There are several advantages. First, it can deal with cancer clinical studies that the proportional hazards assumption is not relevant for each specific cause. Furthermore, using parametric models sometimes results in more efficient parameter estimation than Cox's model.[16,17] In addition, when empirical information is sufficient, parametric models can provide some insight into the shape of the baseline hazard. And it is easier to deal with time-dependent effects. However, our application on the real clinical data shows that, since the two methods are based on different modeling assumptions, their performance can be slightly different on the choice of splitting variables and the tree structures selection. Hence, data exploration and model assumption assessment are critical. We suggest conducting proportional hazards assumption test on the data before applying the competing risk tree methods.

For predictive tree, the current method is applicable to randomized clinical trial data in which the treatment assignment is independent of other risk factors. However, for a large number of cancer retrospective studies, the treatment decisions are based on the characteristics of demographic or clinical factors such as age, physical condition, tumor size, stage, performance score, and metastasis. Further extensions of the predictive survival tree model are needed to deal with this challenge for personalized medicine development.

Author Contributions

Conceived and designed the experiments: WX. Analyzed the data: WX, JC, QK. Wrote the first draft of the manuscript: WX, JC. Contributed to the writing of the manuscript: WX, JC. Agree with manuscript results and conclusions: WX, JC, QK. Jointly developed the structure and arguments for the paper: WX, JC, QK. Made critical revisions and approved final version: WX, JC. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. *J Am Stat Assoc*. 1963;58(302):415–34.
2. Breiman L, Friedman J, Stone CJ, et al. *Classification and Regression Trees*. Boca Raton, FL: CRC press; 1984.
3. Chen J, Yu K, Hsing A, et al. A partially linear tree-based regression model for assessing complex joint gene–gene and gene–environment effects [J]. *Genet Epidemiol*. 2007;31(3):238–51.
4. Xu W, Del Bel R, Bairati I, et al. Adjusted survival tree models for genetic association: prognostic and predictive effects. *Austin Biom Biostat*. 2015;2(4):1–8.
5. Hinchliffe SR, Lambert PC. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Med Res Methodol*. 2013;13(1):13.
6. Lau B, Cole SR, Gange SJ. Parametric mixture models to evaluate and summarize hazard ratios in the presence of competing risks with time-dependent hazards and delayed entry. *Stat Med*. 2011;30(6):654–65.
7. Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*. 2005; 61(1):223–9.
8. Therneau TM, Atkinson EJ. *An Introduction to Recursive Partitioning Using the RPART Routines*. Available at: http://www.mayo.edu/hsr/techrpt/61.pdf.
9. Hothorn T, Hornik K, Strobl C, et al. *Party: A Laboratory for Recursive Partitioning*. 2010. Available at: https://cran.r-project.org/web/packages/party/vignettes/party.pdf.
10. Ibrahim NA, Abdul Kudus A, Daud I, et al. Decision tree for competing risks survival probability in breast cancer study. *Int J Biol Med Sci*. 2008;3(1):25–9.
11. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics*. 2014;15(4):757–73.
12. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons; 2011.
13. Pintilie M. *Competing Risks: A Practical Perspective*. Hoboken, NJ: John Wiley & Sons; 2006.
14. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol*. 2009;170(2):244–56.
15. Leblanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc*. 1993;88(422):457–67.
16. Efron B. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc*. 1977;72:557–65.
17. Oakes D. The asymptotic information in censored survival data. *Biometrika*. 1977;64:441–8.
18. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21(15): 2175–97.
19. Royston P. Flexible parametric alternatives to the Cox model, and more. *Stata J*. 2001;1(1):1–28.
20. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis[J]. *Stata J*. 2009;9(2):265.
21. Huang SH, Xu W, Waldron J, et al. Refining UICC TNM stage and prognostic groups for non-metastatic HPV-related oropharyngeal carcinomas. *J Clin Oncol*. March 10, 2015;33(8):836–45.
22. Jackson CH. *Flexsurv: A Platform for Parametric Survival Modelling in R*. Available at: https://cran.r-project.org/web/packages/flexsurv/vignettes/flexsurv.pdf.