Asanao Shimokawa*, Yohei Kawasaki and Etsuo Miyaoka

Comparison of Splitting Methods on Survival Tree

Abstract: We compare splitting methods for constructing survival trees that are used as a model of survival time based on covariates. A number of splitting criteria on the classification and regression tree (CART) have been proposed by various authors, and we compare nine criteria through simulations. Comparative studies have been restricted to criteria that suppose the survival model for each terminal node in the final tree as a non-parametric model. As the main results, the criteria using the exponential log-likelihood loss, log-rank test statistics, the deviance residual under the proportional hazard model, or square error of martingale residual are recommended when it appears that the data have constant hazard with the passage of time. On the other hand, when the data are thought to have decreasing hazard with passage of time, the criterion using the two-sample test statistic, or square error of deviance residual would be optimal. Moreover, when the data are thought to have increasing hazard with the passage of time, the criterion using the exponential log-likelihood loss, or impurity that combines observed times and the proportion of censored observations would be the best. We also present the results of an actual medical research to show the utility of survival trees.

Keywords: survival tree, CART, recursive partitioning

DOI 10.1515/ijb-2014-0029

1 Introduction

In the field of medical research, analysis of time-to-event data is an important subject. The estimation of a survival function using time-to-event data cannot be considered a simple regression problem owing to the presence of censored data. Censored data does not have the correct interval length between the start point (e.g. detection date of illness or surgery date) and end point (e.g. date of death or date of recurrence) as a response variable. In this paper, we deal with right censored cases because they are frequently encountered in medical data analyses. In order to handle a regression problem that includes censored data based on covariates, the Cox proportional hazard (PD) model [1] has been most widely used. In addition to the simpleness of inference, this semiparametric model has an advantage in that it can easily understand the covariate effects. However, this model requires PD assumptions, and certain assumptions about the relationship between covariates and response variables. Moreover, when this model includes many covariates, interpretation is difficult. In this paper, we deal with survival trees, which involve constructing a tree-structured model based on covariates. Because the proposed method uses a hierarchical structure, the relationship between covariates and hazards can be determined easily. Moreover, it is easy to predict the survival function for a new patient based on the estimated model. For example, men older than 10 years with a tumor of size greater than 10 mm have a high risk of mortality.

The classification and regression tree (CART), proposed by Breiman et al. [2], is used extensively for constructing tree structures. We only deal with binary tree structures, which dichotomize sample data recursively. The CART is composed by three steps: splitting, pruning, and selection. Learning samples are recursively dichotomized in the splitting step, and thereby, a maximum size tree is constructed. Criterions

***Corresponding author: Asanao Shimokawa,** Department of Mathematical information Science, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan, E-mail: a.shimokawa0226@gmail.com
Yohei Kawasaki, Etsuo Miyaoka, Department of Mathematics, Tokyo University of Science, Tokyo, Japan, E-mail: ykawasaki@ma.kagu.tus.ac.jp, miyaoka@rs.kagu.tus.ac.jp

for splitting have been proposed by various authors, and these criterions of two types: one that minimizes the risk within the node, and the other that maximizes the degree of separation between nodes. The maximum size tree obtained by the splitting step suffers from an overfitting problem. To handle this problem, a set of nested subtrees is produced from the maximum size tree in the pruning step. The cost-complexity measure, proposed by Breiman et al. [2], or the split-complexity measure, proposed by Leblanc and Crowley [3], is used in the pruning step. Each subtree obtained by this step is considered as a candidate of the final survival regression model. In the selection step, the optimal size tree is selected.

In this paper, we compare the nine splitting criteria that suppose the survival model for each terminal node in the final tree as a non-parametric model. That is, the survival functions constructed from each terminal node of the final tree model are estimated by the Kaplan–Meier method. Specifically, the criterions compared are as follows. Gordon and Olshen [4] used the L^1-Wasserstein distance between Kaplan–Meier survival curves associated with two child nodes as a criterion of separation. Davis and Anderson [5] used the exponential log-likelihood loss (EL) as the criterion, where the split that minimizes the loss among the possible splits is selected. The criteria using two-sample test statistics have been studied by some authors [3],[6],[7]. These statistics have various forms based on the choice of the weights. We use the log-rank (LR), generalized Wilcoxon (GW) and Tarone–Ware (TW) test statistics in this paper. Leblanc and Crowley [8] used the node deviance measure under the PD model, the full likelihood of which is approximated by replacing the cumulative baseline hazard function by the Nelson–Aalen estimator. Zhang [9] proposed a criterion that combines two impurity measures, one for the observed times and one for the proportion of censored observations. Finally, we use the martingale residuals (MR) from a null Cox model, proposed by Therneau et al. [10]. Keles and Segal [11] constructed a survival tree based on the square error of these residuals. In addition to the MRs, we compared the criterion based on the deviance residuals (DR) from the null Cox model.

A comparative research of splitting methods on survival trees was performed by Radespiel-Tröger et al. [12]. They compared seven splitting criteria. Some of these criteria use pruning while the others do not. In Radespiel-Tröger et al. [13], six splitting methods were compared using Gallstone Clearance data. We restrict the splitting methods that are considered in this study to those using the pruning algorithm and compare the criteria in more situations than those done in previous studies using simulations. As one of the simulation-based evaluation methods, we use the integrated Brier score [14], which was used by Radespiel-Tröger et al. Bou-Hamad et al. [15] have provided a good review that includes the method for constructing the survival tree and for applying the model.

Toward the end of this paper, we construct a survival tree using the bone marrow transplantation data for leukemia patients as an example of the application of the survival tree. This data consists of 137 samples and 54 of these are censored. We use 10 covariates for each patient.

The remainder of this paper is organized as follows. In Section 2, we describe the notations and constructing method of survival trees, splitting methods are compared. Simulation methods, validation methods and the results are shown in Section 3. We analyze the actual data using survival trees in Section 4, and the conclusion is presented in Section 5.

2 Methods

2.1 Notation and survival function

We denote the true survival time as Y and the true censoring time as C. Then, the observation time is given by $X = \min(Y, C)$. $\delta = I(Y \leq C)$ represents the censoring indicator, which is 1 if the observation is an event and is 0 if the observation is censored. Let $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)$ denote a p-dimensional covariate vector. In this study, we only deal with the continuous covariate; however, more generally, it is possible to include the categorical variable. An observed learning sample will be represented by $\mathscr{L} = \{(x_i, \delta_i, z_i); i = 1, 2, \cdots, N\}$.

Let t denote a node in the tree T. We denote a split of the node t as s_t and denote a set of terminal nodes of T that represents the nodes at the bottom layer of the tree structure as $\tilde{T}$. Further, the set of internal nodes of T, which represents the nodes other than the bottom layer of the tree structure, is denoted by S.

In this paper, we consider a non-parametric estimate of the survival function at each terminal node of T. That is, the survival function of each terminal node is estimated by the Kaplan–Meier method. If the learning samples in the node t are represented as $\mathscr{L}_{t|} = \{(x_i, \delta_i, z_i); i = 1, 2, \cdots, N_t\}$, then the Kaplan–Meier survival function of t is given by

$$\hat{S}_t(x) = \begin{cases} 1 & (x < y_{(1)}) \\ \prod\limits_{i(x_i \leq x)} \left(1 - \frac{d_i}{n_i}\right) & (y_{(1)} \leq x) \end{cases}, \tag{1}$$

where $y_{(1)}$ represents the earliest event occurrence time in $\mathscr{L}_t$. d_i and n_i represent the number of events and risk at time $x_i (i = 1, 2, \cdots, N_t)$, respectively. We will be able to estimate the conditional survival function $S(x|z_{\text{new}})$ for a new patient with covariate $Z = z_{\text{new}}$ by precomposing the tree structure and the Kaplan–Meier survival function for each terminal node using $\mathscr{L}$.

2.2 Construction of survival tree

The CART algorithm for constructing the tree structure using $\mathscr{L}$ learning samples consists of three steps: splitting, pruning, and selection. First, all the learning samples are divided into two groups according to the covariates. The selection method of the covariate and the threshold for this separation is described later. A maximum size tree, T_0, is constructed by repeated divisions of data in each node. Tree T_0 is the optimal tree model of the learning samples, but it is not always optimal for a new patient (i.e. overfitting occurs). Therefore, we need to yield an optimal model through pruning and selection.

In the pruning step, a set of subtrees $T_1, T_2, \ldots, T_M$ can be constructed by using the cost-complexity measure or the split-complexity measure, and we use the latter measure in this study. The split-complexity measure is constructed from the degree of separation between nodes in the tree and the complexity of the tree:

$$G_\alpha(T) = \sum_{t \in S} G(t) - \alpha|S|,$$

where $G(t)$ is a measure of separation between child nodes of t, and we use the standardized log-rank test statistics according to Leblanc and Crowley [3]. $|S|$ is the number of internal nodes of T, which is a complexity measure of the tree. $G_\alpha(T)$ returns a high value when the degree of separation in each internal node of T is high and the model is simple. The optimal subtree T_j for an arbitrary α is determined by using these measures. If the value of α is 0, then the optimal subtree is T_0. If α is ∞, on the other hand, then the subtree of root node T_M only (i.e. a model that is not considered a tree structure) is selected as the optimal subtree. The set of optimal subtrees is given by gradually increasing α from 0. It can be assured that these subtrees have nesting structures on account of this algorithm, that is, T_j is the subtree of T_{j-1} ($j = 1, 2, \ldots, M$).

Finally in the selection step, an optimal subtree is selected from the set of subtrees created through the pruning process. We use the bootstrap bias correction method in this step as described in Leblanc and Crowley [3]. The number of bootstrap samples is set to 50, and a fixed penalty term of $\alpha_c = 2$ for each internal node is used.

2.3 Splitting methods

Splitting methods compared in this study are as follows:

(i) The criterion using the L^1-Wasserstein distance between Kaplan–Meier survival curves (WD)

Let $F_L(x)$ and $F_R(x)$ be the improper functions that are obtained from the Kaplan–Meier survival curves $S_L(x)$ and $S_R(x)$ associated with two child nodes, where

$$F_L(x) = 1 - S_L(x) \text{ and } F_R(x) = 1 - S_R(x).$$

Let m_L and m_R be the limit values of $F_L(x)$ and $F_R(x)$, and let m be $\min(m_L, m_R)$. Then, the L^1-Wasserstein distance between the Kaplan–Meier survival curves obtained by two child nodes can be written as

$$G(t) = \int_0^m |F_L^{-1}(u) - F_R^{-1}(u)| du$$

where

$$F_L^{-1}(u) = \min_x F_L(x) \geq u \text{ and } F_R^{-1}(u) = \min_x F_R(x) \geq u.$$

This criterion function represents the area between the Kaplan–Meier survival functions of the child nodes, which is obtained by splitting. The split that maximizes $G(t)$ is selected.

(ii) The criterion using the EL

Let the hazard of a node be

$$\lambda(x|t) = \lambda_t,$$

where λ_t represents a constant parameter. Then the maximum log-likelihood estimator of λ_t is given by

$$\hat{\lambda}_t = \frac{\sum_{i \in \mathscr{L}_t} \delta_i}{\sum_{i \in \mathscr{L}_t} x_i}.$$

The EL of the node t is given by

$$\begin{aligned} R(t) &= -\log L(\hat{\lambda}_t) \\ &= \sum_{i \in \mathscr{L}_t} \delta_i - \sum_{i \in \mathscr{L}_t} \delta_i \log(\hat{\lambda}_t). \end{aligned}$$

The split that minimizes this EL is selected as the optimal.

(iii) The criterion using the two-sample test statistics

The two-sample test statistics for child nodes have the following form:

$$G(t) = \frac{\sum_{i \in \mathscr{L}_t} w_i [d_{Li} - \mathrm{E}(D_{Li})]}{\sqrt{\sum_{i \in \mathscr{L}_t} w_i^2 \mathrm{Var}(D_{Li})}}, \tag{2}$$

where d_{Li} is the number of events in the left child node at x_i, and w_i are constants that are used to weight the respective statistic. $\mathrm{E}(D_{Li})$ and $\mathrm{Var}(D_{Li})$ are represented as follows:

$$\mathrm{E}(D_{Li}) = n_{Li} \frac{d_i}{n_i},$$

$$\mathrm{Var}(D_{Li}) = \frac{n_{Li}}{n_i} \left(1 - \frac{n_{Li}}{n_i}\right) \left(\frac{n_i - d_i}{n_i - 1}\right) d_i,$$

where n_{Li} is the number of risks at x_i in the left child node. With appropriate choices of weights w_i, statistic (2) becomes many test statistics. We used the LR, GW, and TW test statistics. That is, letting $w_i = 1$ leads to the LR test statistic, $w_i = n_i$ gives the GW test statistic, and $w_i = \sqrt{n_i}$ gives the TW test statistic. The split that maximizes the statistic is selected.

(iv) The criterion using the DR under the PD model
Let the hazard of a node be

$$\lambda(x|t) = \lambda_0(x)\theta_t,$$

where $\lambda_0(x)$ is the baseline hazard and θ_t is a nonnegative parameter. Then, the full likelihood for sample data given tree T is

$$L = \prod_{t \in \tilde{T}} \prod_{i \in \mathscr{L}_t} (\lambda_0(x_i)\theta_t)^{\delta_i} \exp(-\Lambda_0(x_i)\theta_t), \tag{3}$$

where $\Lambda_0(x)$ is the baseline cumulative hazard function. The Nelson–Aalen estimate of $\Lambda_0(x)$ is

$$\hat{\Lambda}_0(x) = \sum_{i(x_i \leq x)} \frac{d_i}{n_i}. \tag{4}$$

The one-step estimate of θ_t is

$$\hat{\theta}_t = \frac{\sum\limits_{i \in \mathscr{L}_t} \delta_i}{\sum\limits_{i \in \mathscr{L}_t} \hat{\Lambda}_0(x_i)}.$$

By using these estimates, the full likelihood DR is obtained as follows:

$$R(t) = \sum_{i \in \mathscr{L}_t} 2 \left[\delta_i \log \left(\frac{\delta_i}{\hat{\Lambda}_0(x_i)\hat{\theta}_t} \right) - (\delta_i - \hat{\Lambda}_0(x_i)\hat{\theta}_t) \right]$$

The split that minimizes the deviance is selected.

(v) The criterion using the impurity that combines observed times and the proportion of censored observations (NI)
The impurity criterion proposed by Zhang [9] is

$$R(t) = w_1 i_x(t) + w_2 i_\delta(t)$$

where w_1 and w_2 are prespecified weights, and $i_x(t)$ and $i_\delta(t)$ are the impurities of node t for the observed time and censoring, respectively. In this paper, $i_x(t)$ is defined as follows:

$$i_x(t) = \frac{\sum\limits_{i \in \mathscr{L}_t} (x_i - \bar{x}(t))^2}{n_t}$$

where $\bar{x}(t)$ is the average of the observation times in node t, and n_t is the number of samples in the node t. The impurity $i_\delta(t)$ is defined as

$$i_\delta(t) = -p_t \log(p_t) - (1 - p_t) \log(1 - p_t),$$

where p_t is the proportion of censoring in node t,

$$p_t = \frac{\sum\limits_{i \in \mathscr{L}_t} (1 - \delta_i)}{n_t}$$

We compared three pairs of weights $(w_1 = 1, w_2 = 0)$, $(w_1 = 1, w_2 = 1)$ and $(w_1 = 3, w_2 = 1)$ in the simulation. The split that minimizes the impurity criterion is selected.

(vi) The criterion using the square error of residuals from a null Cox model
The sum of the square error of residuals in a node is

$$R(t) = \sum_{i \in \mathscr{L}_t} (r_i - \bar{r}(t))^2$$

where r_i is the residual of observation i, and $\bar{r}(t)$ is the average of the residuals in node t. We use the MR M_i and DR D_i from a null Cox model such as r_i. These residuals are represented as follows:

(1) MR

The MR is given by

$$M_i = \delta_i - \hat{\Lambda}_{0p}(x_i),$$

where $\hat{\Lambda}_{0p}(x)$ is the Nelson–Aalen estimator of the parent node t_p, which is given by eq. (4).

(2) DR

The DR is

$$D_i = \operatorname{sgn}(M_i)[-2\{M_i + \delta_i \log(\delta_i - M_i)\}]^{\frac{1}{2}},$$

where M_i is the MR of observation i.

The split that minimizes the deviation of residuals is selected.

We make comparative studies for all these 9 splitting methods (there are 11 patterns if the three pairs of weights in the NI method are included) by using simulation studies. The details of the simulation methods are described in the next section.

3 Simulations

3.1 Validation methods

The integrated Brier score for censored observations proposed by Graf et al. [14] is used to compare the split methods. This score is calculated using the test sample $\mathscr{L}_{\text{test}} = \{(x_i, \delta_i, z_i); i = 1, 2, \cdots, N_{\text{test}}\}$, which is an independent sample drawn from the same simulated population. The integrated Brier score for survival function $\hat{S}(x|z)$ that is modeled by the tree structure T is defined by

$$\text{IBS}_T = \frac{1}{\max(x_i)} \int_0^{\max(x_i)} \text{BS}_T(x) dx,$$

where $\text{BS}_T(x)$ is the Brier score of T. The Brier score is interpreted as the mean square error between the inferred survival function $\hat{S}(x|z)$ and the test data that are weighted such that the loss of information due to censoring is compensated:

$$\begin{aligned}\text{BS}_T(x) = \frac{1}{N_{\text{test}}} \sum_{i \in \mathscr{L}_{test}} \{&(0 - \hat{S}(x|z_i))^2 I(x_i \le x, \delta_i = 1)(1/\hat{G}(x_i)) \\ &+ (1 - \hat{S}(x|z_i))^2 I(x_i > x)(1/\hat{G}(x))\},\end{aligned}$$

where $\hat{G}(x)$ is the Kaplan–Meier estimate of the censoring distribution of C, that is, the Kaplan–Meier estimate based on $(x_i, 1 - \delta_i),\ \ i = 1, 2, \ldots, N_{\text{test}}$. Let IBS_{T_M} be the integrated Brier score evaluated from the T_M that only has a root node. Then, the measure of the explained residual variation is given by

$$R^2 = 1 - \frac{\text{IBS}_T}{\text{IBS}_{T_M}}.$$

We evaluate the tree T obtained by each splitting method using IBS_T and R^2.

3.2 Setting and tree structure

The true tree structure used in the simulation is given by Figure 1. This structure is constructed based on the research of Radespiel-Tröger et al. [13]. The circles in the figure represent internal nodes. The covariates

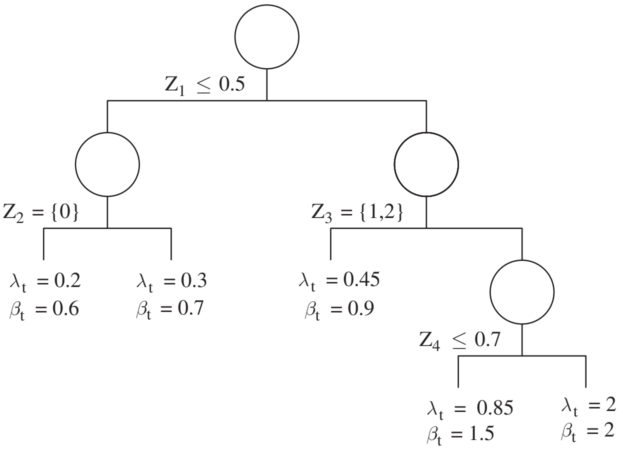
$Z_1 \leq 0.5$
$Z_2 = \{0\}$
$Z_3 = \{1,2\}$
$Z_4 \leq 0.7$
$\lambda_t = 0.2$, $\beta_t = 0.6$
$\lambda_t = 0.3$, $\beta_t = 0.7$
$\lambda_t = 0.45$, $\beta_t = 0.9$
$\lambda_t = 0.85$, $\beta_t = 1.5$
$\lambda_t = 2$, $\beta_t = 2$

Figure 1: True tree structure used in simulations.

used in this simulation are $Z_1, Z_2, \ldots, Z_{10}$; Z_1 and $Z_4 - Z_6$ are continuous with uniform distribution on the interval $[0,1]$, Z_2 and $Z_7 - Z_8$ are binary values of $\{0,1\}$, and Z_3 and $Z_9 - Z_{10}$ are categorical values with five levels $\{1,2,3,4,5\}$. The variables Z_1, Z_2, Z_3, Z_4 are used in the true tree structure and the other variables are nuisances. The true thresholds for splitting of the root node is determined as 0.5.

We suppose four patterns of the survival model. First, we suppose the exponential model, the survival function of which is given by

$$S(y;\lambda_t) = P(Y > y;\lambda_t) = \exp(-\lambda_t y),$$

where the parameter λ_t represents the constant hazard of node t. The value of λ_t is defined as follows:

$$\lambda_t = \begin{cases} 0.2 & (Z_1 \leq 0.5 \cap Z_2 = \{0\}) \\ 0.3 & (Z_1 \leq 0.5 \cap Z_2 = \{1\}) \\ 0.45 & (Z_1 > 0.5 \cap Z_3 = \{1,2\}) \\ 0.85 & (Z_1 > 0.5 \cap Z_3 = \{3,4,5\} \cap Z_4 \leq 0.7) \\ 2 & (Z_1 > 0.5 \cap Z_3 = \{3,4,5\} \cap Z_4 > 0.7) \end{cases}$$

Second, we suppose the Weibull model, the survival function of which is given by

$$S(y;\alpha,\beta_t) = P(Y > y;\alpha,\beta_t) = \exp(-\beta_t y^{\alpha})$$

where α and β_t are the shape parameter and scale parameter, respectively. The value of α is set to 0.5 in this study, where the model has decreasing hazard with the passage of time. The values of β_t are set as follows:

$$\beta_t = \begin{cases} 0.6 & (Z_1 \leq 0.5 \cap Z_2 = \{0\}) \\ 0.7 & (Z_1 \leq 0.5 \cap Z_2 = \{1\}) \\ 0.9 & (Z_1 > 0.5 \cap Z_3 = \{1,2\}) \\ 1.5 & (Z_1 > 0.5 \cap Z_3 = \{3,4,5\} \cap Z_4 \leq 0.7) \\ 2 & (Z_1 > 0.5 \cap Z_3 = \{3,4,5\} \cap Z_4 > 0.7) \end{cases}$$

Third, we suppose the Weibull model which has the $\alpha = 1.5$, where the model has increasing hazard with the passage of time. The same values of β_t in the second model are set to this model.

Fourth, we suppose the bathtub-shaped hazard model [16], the survival function of which is given by

$$S(y;a_t,b,c) = P(Y > y;a_t,b,c) = \frac{\exp\left(-\frac{1}{2}a_t y^2\right)}{(1+cy)^{\frac{b}{c}}},$$

where $b = 1$, $c = 5$, and a_t is set to the same values of β_t in the second model.

Through the all models, the covariate Z_1 is considered easy to detect from the difference of hazards in the tree model. On the other hand, the covariate Z_2 is considered difficult to detect. The censoring rates used are 0% and approximately $25\%, 50\%$, and 75% using uniform random numbers. The number of learning samples N, and test samples, N_{test}, are set to 250. We set 10 minimum number of events in nodes as the stop condition of splitting in this study. Simulations are repeated 300 times in every data group.

3.3 Results

The results of the simulations are shown in Tables 1–4. Table 1 lists the average values of the integrated Brier scores, the explained residual validations, and the number of terminal nodes on each splitting method obtained through simulations when the exponential model is used as the true model. The results of simulations when the Weibull survival model, which has decreasing and increasing hazard with the passage of time, are used as the true model are listed in Tables 2 and 3, respectively. The results of simulations when the bathtub-shaped hazard model is used as the true model are shown in Table 4.

It turns out that the EL, LR, GW, TW, PD, NI(1,1), MR, and DR methods show good results about all patterns of the censoring rate from Table 1. On the other hand, the WD and NI(1,0) methods show the not good results. The MR and DR methods, which use the square error of residuals show similar good results as the EL and PD methods when the censoring rate is high. The NI, MR, and DR methods have a tendency to underestimate the size of tree even when the censoring rate is 0%. When the censoring rate is low, the NI method without the weight for impurity of the censoring probability ($w_2 = 0$) shows better results than the results of the NI method with the weight. If the censoring rate is high, on the other hand, the NI method with the weight shows better results.

The best method of detecting the covariate Z_2, which is difficult to detect, is the LR method, that can be found about 20% when the censoring rate is 0%. When the censoring rate becomes about 25% and 50%, the LR and PD methods show the best Z_2 detection result that can be found about 11%. If the censoring rate is over 75%, the Z_2 is difficult to detect.

Table 1: The results of the simulations on splitting methods using the exponential survival model in the true tree structure.

	Censor rate 0%			Censor rate 25%			Censor rate 50%			Censor rate 75%		
	$\text{IBS}_{T_M} = 7.2$			$\text{IBS}_{T_M} = 14.6$			$\text{IBS}_{T_M} = 21.0$			$\text{IBS}_{T_M} = 17.2$		
	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$
WD	6.7	7.2	4.12	13.4	8.2	3.76	19.4	7.8	3.54	16.9	1.9	2.63
EL	6.6	8.5	3.84	13.2	9.2	3.78	18.9	10.4	3.72	16.2	5.9	2.79
LR	6.6	8.4	3.99	13.3	9.0	3.90	18.9	10.1	3.57	16.3	5.6	2.73
GW	6.6	8.5	3.58	13.2	9.4	3.45	18.9	10.4	3.30	16.3	5.3	2.65
TW	6.6	8.4	3.84	13.2	9.2	3.61	18.9	10.2	3.52	16.3	5.6	2.69
PD	6.6	8.5	3.88	13.2	9.2	3.73	18.9	10.3	3.69	16.2	6.0	2.75
NI(1,0)	6.6	8.5	3.38	13.2	9.5	3.26	19.0	9.7	2.89	16.7	3.2	2.39
NI(1,1)	6.6	8.5	3.38	13.2	9.3	3.38	18.8	10.8	3.19	16.2	6.3	2.62
NI(3,1)	6.6	8.5	3.38	13.2	9.2	3.36	18.8	10.4	3.17	16.2	6.0	2.73
MR	6.6	8.7	3.25	13.2	9.2	3.37	18.8	10.6	3.52	16.2	6.0	2.74
DR	6.6	8.7	3.42	13.2	9.4	3.43	18.9	10.4	3.53	16.2	5.8	2.76

Note: IBS_{T_M}: the integrated Brier score evaluated from the root node ×100, IBS_T: the integrated brier score evaluated from the selected tree ×100, R^2: the explained residual variation ×100, $\lvert\tilde{T}\rvert$: the number of terminal nodes about selected tree, WD: the criterion using the L^1-Wasserstein distance between Kaplan–Meir survival curves, EL: the criterion using the exponential log-likelihood loss, LR: the criterion using the log-rank test statistic, GW: the criterion using the generalized Wilcoxon test statistic, TW: the criterion using the Tarone-Ware test statistic, PD: the criterion using the deviance residual under the proportional hazard model, $\text{NI}(w_1, w_2)$: the criterion using the impurity which combines observed times and the proportion of censored observations, MR: the criterion using the square error of the martingale residual, DR: the criterion using the square error of the deviance residual.

Table 2: The results of simulations on splitting methods by using the Weibull survival model which has the hazard to decrease with the passage of time in the true tree structure.

	Censor rate 0%			Censor rate 25%			Censor rate 50%			Censor rate 75%		
	$\text{IBS}_{T_M} = 3.8$			$\text{IBS}_{T_M} = 17.0$			$\text{IBS}_{T_M} = 23.4$			$\text{IBS}_{T_M} = 18.1$		
	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$
WD	3.9	−2.1	5.06	16.9	0.3	3.52	23.4	0.0	3.18	18.3	−0.8	2.27
EL	3.7	1.5	3.13	16.5	2.4	3.08	23.0	1.9	2.98	18.2	−0.7	2.42
LR	3.7	1.8	3.27	16.5	2.4	3.15	23.0	1.8	3.03	18.2	−0.6	2.41
GW	3.7	2.1	2.91	16.5	2.8	2.96	23.0	1.9	2.84	18.3	−0.8	2.44
TW	3.7	1.9	3.19	16.5	2.6	3.05	23.0	1.8	2.94	18.2	−0.6	2.39
PD	3.7	1.9	3.15	16.5	2.4	3.12	23.0	1.9	3.01	18.2	−0.7	2.44
NI(1,0)	3.8	0.9	3.08	16.5	2.5	2.85	23.0	1.6	2.65	18.2	−0.4	2.37
NI(1,1)	3.8	0.9	3.08	16.5	2.5	2.80	22.9	2.4	2.72	18.3	−0.8	2.44
NI(3,1)	3.8	0.9	3.08	16.5	2.4	2.84	22.9	2.1	2.87	18.3	−0.8	2.45
MR	3.7	2.0	2.73	16.5	2.5	3.05	23.0	1.8	3.03	18.2	−0.7	2.40
DR	3.7	2.0	2.95	16.5	2.9	2.91	22.9	2.2	2.90	18.2	−0.7	2.42

Note: See note in Table 1.

Table 3: The results of simulations on splitting methods by using the Weibull survival model which has the hazard to increase with the passage of time in the true tree structure.

	Censor rate 0%			Censor rate 25%			Censor rate 50%			Censor rate 75%		
	$\text{IBS}_{T_M} = 9.8$			$\text{IBS}_{T_M} = 12.3$			$\text{IBS}_{T_M} = 17.8$			$\text{IBS}_{T_M} = 16.3$		
	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$	IBS_T	R^2	$\lvert\tilde{T}\rvert$
WD	9.8	0.1	3.93	12.4	−0.1	3.39	17.9	−0.8	3.28	16.5	−1.3	2.44
EL	9.5	2.6	3.09	12.1	2.1	2.97	17.6	1.2	3.02	16.4	−0.7	2.48
LR	9.5	2.3	3.27	12.2	1.5	3.08	17.6	0.9	3.07	16.5	−1.4	2.50
GW	9.5	2.5	2.91	12.1	1.9	2.74	17.6	0.8	2.80	16.4	−0.5	2.42
TW	9.5	2.4	3.19	12.1	1.9	2.87	17.6	1.0	2.91	16.4	−0.8	2.48
PD	9.5	2.5	3.15	12.1	1.8	3.14	17.6	0.8	3.08	16.5	−1.3	2.59
NI(1,0)	9.5	2.4	3.05	12.1	1.9	2.76	17.7	0.5	2.72	16.5	−1.0	2.46
NI(1,1)	9.5	2.4	3.05	12.1	2.1	2.48	17.6	1.2	2.56	16.4	−0.5	2.45
NI(3,1)	9.5	2.4	3.05	12.0	2.4	2.62	17.5	1.7	2.55	16.4	−0.3	2.43
MR	9.5	2.7	2.73	12.1	2.1	2.77	17.6	1.0	2.97	16.5	−1.3	2.56
DR	9.5	2.5	2.95	12.1	2.1	2.86	17.6	1.0	2.95	16.4	−0.5	2.51

Note: See note in Table 1.

As shown in Table 2, the not good results are obtained when the Weibull survival model which has the hazard to decrease with the passage of time is used as the true model. The PD, the methods using two-sample test statistic (GW, TW), and the methods using residuals (MR, DR) show good results when there is no censoring data. When the censoring rate is increased, the methods of DR and NI, which has the weight for impurity of the censoring probability ($w_2 = 1$), show good results. On the other hand, the method that does not construct a tree structure shows better results when the data have high probability of censoring.

It turns out that the not good results are obtained when the bathtub-shaped hazard with the passage of time is used as the true model from Table 4. Even when the censoring is not occurred, the method that does not construct a tree structure shows better results. From the results of additional research (not shown), we can find that, if sample size is increased, the performances of tree-structured model are improved. However, in the case of canonical sample size and if the data is considered to have a bathtub-shaped hazard, we recommend not using the survival tree.

Table 4: The results of simulations on splitting methods by using the bathtub-shaped hazard with the passage of time in the true tree structure.

	Censor rate 0%			**Censor rate 25%**			**Censor rate 50%**			**Censor rate 75%**		
	$\mathbf{IBS}_{T_M} = 11.0$			$\mathbf{IBS}_{T_M} = 13.5$			$\mathbf{IBS}_{T_M} = 19.1$			$\mathbf{IBS}_{T_M} = 17.3$		
	$\mathbf{IBS}_T$	R^2	$\lvert\tilde{T}\rvert$	$\mathbf{IBS}_T$	R^2	$\lvert\tilde{T}\rvert$	$\mathbf{IBS}_T$	R^2	$\lvert\tilde{T}\rvert$	$\mathbf{IBS}_T$	R^2	$\lvert\tilde{T}\rvert$
WD	11.4	−3.7	4.84	14.0	−4.1	4.04	19.9	−4.2	3.34	17.8	−2.4	2.09
EL	11.2	−1.4	3.27	13.8	−2.7	3.40	20.0	−4.9	3.41	18.0	−4.0	2.35
LR	11.2	−1.2	3.29	13.8	−2.3	3.31	19.9	−4.2	3.28	18.0	−3.7	2.23
GW	11.3	−2.8	3.76	14.0	−3.7	3.63	19.9	−4.4	3.28	17.9	−3.0	2.18
TW	11.2	−1.8	3.49	13.9	−3.4	3.63	19.9	−4.5	3.29	18.0	−3.6	2.29
PD	11.2	−1.3	3.37	13.8	−2.4	3.39	19.9	−4.3	3.25	18.0	−4.1	2.32
NI(1,0)	11.2	−1.3	3.29	13.8	−2.6	3.39	19.7	−3.2	3.23	17.7	−2.1	2.27
NI(1,1)	11.2	−1.3	3.29	13.8	−2.2	3.26	19.7	−3.2	2.85	18.0	−3.5	2.25
NI(3,1)	11.2	−1.3	3.29	13.8	−2.5	3.26	19.8	−3.9	3.07	18.0	−3.5	2.28
MR	11.1	−1.0	3.24	13.8	−2.2	3.13	19.9	−4.2	3.23	18.0	−3.9	2.29
DR	11.3	−2.3	3.64	13.9	−3.0	3.39	19.9	−4.4	3.26	18.0	−3.6	2.27

Note: See note in Table 1.

To research more complicated situations, we simulated under the model which the leftmost terminal node and the rightmost terminal node in Figure 1 are replaced. From the results of simulations in several patterns of hazard and splitting methods, we confirmed that the performance of all methods would be decreased while the tendency of results in each method is maintained. The reason of this decrease is considered to be due to the cause of exclusive OR problem. That is, if the true model is linearly inseparable, then the tree becomes prohibitively large and sometimes it causes to lower performance of the obtained model.

Across the whole results, the average tree sizes given in Tables 1–4 are tend to be underestimated than the true tree size. As one reason for this, it is considered that the number of samples for constructing survival trees has not been sufficient in this simulations. To confirm this, we run additional simulations which the number of learning samples is changed to $N = 1000$ under several settings. As the results, in each settings, we confirmed that the size of tree becomes approximately 0.5 larger than the tree which obtained under $N = 250$.

To compare the correctness of the split selection in a node for each methods, the percentages that Z_1 is selected in root nodes, and the medians of thresholds at that time when censoring rate is 50% are shown in Table 5. The methods other than WD, NI(1,0), and NI(3,1) show the high performance when the exponential

Table 5: The results of the percentages that Z_1 is selected at root nodes, and the medians of thresholds at that time when censoring rate is approximately 50%.

	Constant		**Decreasing**		**Increasing**		**Bathtub shape**	
	$\mathbf{Z_1}$(%)	**Med.**	$\mathbf{Z_1}$(%)	**Med.**	$\mathbf{Z_1}$(%)	**Med.**	$\mathbf{Z_1}$(%)	**Med.**
WD	89	0.47	53	0.41	65	0.50	27	0.48
EL	100	0.50	88	0.50	87	0.50	43	0.48
LR	99	0.50	87	0.50	86	0.51	43	0.50
GW	99	0.51	84	0.51	81	0.51	25	0.52
TW	99	0.50	87	0.50	83	0.51	35	0.51
PD	100	0.50	88	0.50	84	0.50	46	0.48
NI(1,0)	95	0.50	75	0.49	69	0.50	29	0.46
NI(1,1)	99	0.50	88	0.50	79	0.50	44	0.49
NI(3,1)	97	0.50	89	0.50	86	0.50	40	0.49
MR	100	0.50	88	0.50	86	0.50	46	0.48
DR	100	0.50	88	0.50	85	0.50	36	0.49

Note: Z_1(%): the percentage that Z_1 is selected at root node, med.: the median of thresholds when Z_1 is selected at root node.

survival model is used in the true tree structure. On the other hand, when the Weibull survival model is used in the true tree the EL, LR, PD, and MR methods show the high performance. Moreover, when the bathtub-shaped hazard model is used in the true tree the percentages that the true covariate is selected in root nodes is lower than 50%.

From these results, we conclude that the EL, LR, PD, and MR methods would be recommended when the data appears to have constant hazard at the time. On the other hand, the DR and the methods using two-sample test statistics would be the best when the data are thought to have decreasing hazard with the passage of time. When the data are thought to have increasing hazard with the passage of time, the methods of EL and NI, which have the weight for impurity of the censoring probability would be the best. Moreover, as mentioned previously, we recommend not using the survival tree when the data is considered to have a bathtub-shaped hazard or the censoring rate is very high.

4 Example

In this section, we show the specific application of the survival tree using the leukemia patients bone marrow transplantation data. These data pertains to the patients for whom transplants were conducted from 1984 to 1989. The 137 patients used in this example were treated at one of four hospitals. The observation time was defined as from the date of transplant surgery to the date of relapse, death, or the last verification of survival. The censoring indicator is defined as 1 if the end of the observation was determined by death or relapse. Censoring was included in the data of 54 of the patients; the number of covariates used in this study, which are listed in Table 6, was 10. These covariates were measured at the time of transplantation. The details of these data are given in Copelan et al. [17]; the data set used in this section is available at the website offered in Klein and Moeschberger [18].

Table 6: Used covariates for survival tree construction in example.

Z_1: Disease group {1 – ALL, 2 – AML low risk, 3 – AML high risk}
Z_2: Patient age in years – 28
Z_3: Donor age in years – 28
Z_4: Patient sex {1 – male, 0 – female}
Z_5: Donor sex {1 – male, 0 – female}
Z_6: Patient CMV status {1 – CMV positive, 0 – CMV negative}
Z_7: Donor CMV status {1 – CMV positive, 0 – CMV negative}
Z_8: Waiting time for transplant in days
Z_9: FAB {1 – FAB grade 4 or 5 and AML, 0 – otherwise}
Z_{10}: MTX used as a graft-versus-host prophylactic {1 – yes, 0 – no}

Note: ALL: acute lymphoblastic leukemia, AML: acute myelocytic leukemia, CMV: cytomegalovirus, FAB: French–American–British classification, MTX: methotrexate.

In Klein and Moeschberger [18], an optimal hazard model was estimated from these data using the Wald test and Akaike information criteria:

$$\begin{aligned}\lambda(x|Z) = \lambda_0(x)\exp(&-1.091\times I(Z_1=2) - 0.404\times I(Z_1=3)\\ &+0.837\times Z_9 + 0.007\times Z_2 + 0.004\times Z_3 + 0.003\times(Z_2\times Z_3))\end{aligned}$$

We compare the survival trees using the all nine methods. We set five minimum number of events in nodes as the stop condition of splitting in this example. In the selection step, the number of bootstrap samples is set to 50, and a fixed penalty term of $\alpha_c = 2$ for each terminal node is used.

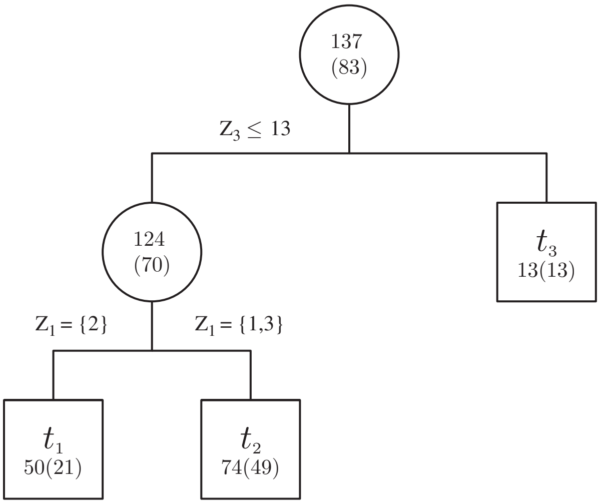

Figure 2: The tree made from the bone marrow transplant patients data by using the EL, LR, TW, PD, MR, and DR methods.

Based on the result of the analysis, the EL, LR, TW, PD, MR, and DR methods have shown exactly the same results. The obtained tree structure is given in Figure 2. The circle and square in the figure represent the internal nodes and terminal nodes, respectively. The values in the shapes represent the number of patients in the node, and the value in the bracket represent the number of events. The Kaplan–Meier survival curves for each terminal node ($t_1 - t_3$) are shown in Figure 3. The tree have three terminal nodes where the node t_1 has the highest risk and t_2 has the lowest risk of death or relapse. As shown in Figure 3, the Kaplan–Meier survival curves are well separated from each other, and we consider that the availability of the survival tree for splitting the data on survival time is shown. The covariates used in the tree structure are Z_3 and Z_1, which are the age of donor and the disease group of the patient, respectively. By using this model, a new patient who transplanted bone marrow from donor of age 41 or younger and the AML low-risk group are considered to be a low risk of death or relapse. On the other hand, a patient who transplanted bone marrow from donor of age 41 or older is considered to be a high risk of death or relapse.

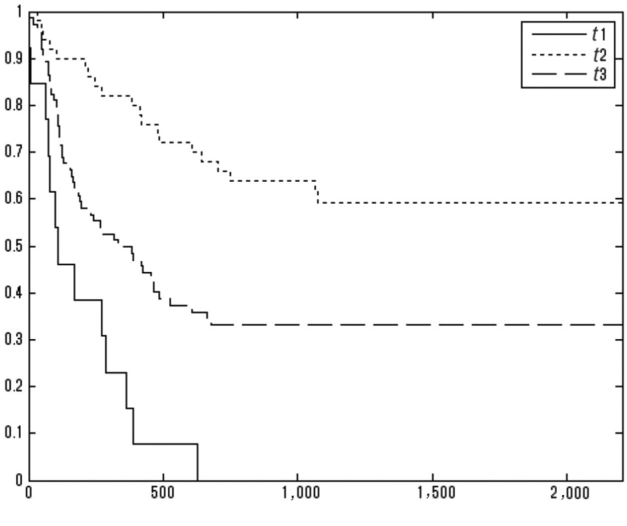

Figure 3: The Kaplan–Meier survival curves for each terminal node of Figure 2 (horizontal axis: days; vertical axis: probability).

The GW method has shown the result that it would be better to not split the left child node of the root node in Figure 2. Therefore, the obtained tree structure has one split and two terminal nodes. On the other hand, the NI method has shown the result that it would be better to not split the data. Moreover, the WD method has shown the completely different result from the other methods. The covariate of waiting time for transplant in days is included in the tree structure as the splitting rule of the root node, and as a result, the obtained tree structure has three splits and four terminal nodes. Although several different tree structures are obtained from each splitting methods, we consider that the tree structure shown in Figure 2 would be best considering the simulation results.

5 Conclusion

Tree-structured model has an advantage in that it is easy to understand the relationship between covariates and hazards. Moreover, there is an additional advantage in that it is easy to predict the survival function for a new patient based on the estimated model. However, it suffers from a disadvantage in that it involves a rapid advancements in long learning time for a large sample data set. However, PC technology in recent years may be able to resolves this disadvantage.

The criteria for splitting in survival tree have been proposed by many researchers. However, extensive research on the comparisons of these criteria has not been conducted. In this paper, we concentrate heavily on the comparison of these criteria using simulations. We have restricted the criteria researched in this study to the methods that suppose the survival model for each terminal node in the final tree as a non-parametric model. The simulation data have been obtained from the assumption of the constant, decreasing, increasing, or bathtub-shaped hazard model with the passage of time. Under these conditions, the performance of the 9 splitting methods (there are 11 patterns if the three pairs of weights in NI method are included) have been compared.

We have concluded the results from the simulations as follows. The EL, LR, PD, and MR methods are recommended when it appears that be the data have constant hazard with the passage of time. On the other hand, when the data are thought to have decreasing hazard with passage of time, the DR and the methods using two-sample test statistics would be the optimal. Moreover, when the data are thought to have increasing hazard with passage of time, the EL and the NI methods would be recommended. Finally, when the data is considered to have a bathtub-shaped hazard or censoring rate is very high we recommend not using a survival tree. From the simulation, it is shown that the splitting method should be selected based on the censoring rate of data and the hazard shape assumed.

As an actual example of the application of survival tree, we have constructed the survival tree of 137 leukemia patients that underwent bone marrow transplantation. The number of covariates used in the study is 10 with a focus on the potential risk factors measured at the time of transplantation. The obtained tree had three terminal nodes. From the results, the survival tree methods are considered to effectively cluster the survival time data.

References

1. Cox DR. Regression models and life-tables. J R Stat Soc 1972;34:187–220.
2. Breiman L, Friedman JH, Olshen RA, Stone C. Classification and regression trees. Belmont, CA: Wadsworth, 1984.
3. Leblanc M, Crowley J. Survival trees by goodness of split. J Am Stat Assoc 1993;88:457–67.
4. Gordon L, Olshen RA. Tree-structured survival analysis. Cancer Treat Rep 1985;69:1065–9.
5. Davis RB, Anderson JR. Exponential survival trees. Stat Med 1989;8:947–61.
6. Ciampi A, Hogg SA, Mckinney S, Thiffault J. RECPAMF: 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 Methods Programs Biomed 1988;26:239–56.
7. Segal MR. Regression trees for censored data. Biometrics 1988;44:35–47.
8. Leblanc M, Crowley J. Relative risk trees for censored survival data. Biometrics 1992;48:411–25.

9. Zhang HP. Splitting criteria in survival trees. In Statistical Modelling. Proceedings of the 10th International Workshop on Statistical Modeling. Innsbruck, Austria: Springer Verlag, 1995: 305–14.
10. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residual for survival models. Biometrika 1990;77:147–60.
11. Keles S, Segal MR. Residual-based tree-structured survival analysis. Stat Med 2002;21:313–26.
12. Radespiel-Tröger M, Rabenstein T, Schneider HT, Lausen B. Comparison of tree-based methods for prognostic stratification of survival data. Artif Intell Med 2003;28:323–41.
13. Radespiel-Tröger M, Gefeller O, Rabenstein T, Hothorn T. Association between split selection instability and predictive error in survival trees. Methods Inf Med 2006;45:548–56.
14. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparisons of prognostic classification schemes for survival data. Stat Med 1999;18:2529–45.
15. Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Stat Surv 2011;5:44–71.
16. Hjorth U. A reliability distribution with increasing, decreasing and bathtub-shaped failure rates. Technometrics 1980;22:99–107.
17. Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, Klein JP, et al. Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with BuCy2. Blood 1991;78:838–43.
18. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data, 2nd ed. New York: Springer, 2003.