Asanao Shimokawa*, Yoshitaka Narita, Soichiro Shibui and Etsuo Miyaoka

Tree Based Method for Aggregate Survival Data Modeling

DOI 10.1515/ijb-2015-0071

Abstract: In many scenarios, a patient in medical research is treated as a statistical unit. However, in some scenarios, we are interested in treating aggregate data as a statistical unit. In such situations, each set of aggregated data is considered to be a concept in a symbolic representation, and each concept has a hyperrectangle or multiple points in the variable space. To construct a tree-structured model from these aggregate survival data, we propose a new approach, where a datum can be included in several terminal nodes in a tree. By constructing a model under this condition, we expect to obtain a more flexible model while retaining the interpretive ease of a hierarchical structure. In this approach, the survival function of concepts that are partially included in a node is constructed using the Kaplan-Meier method, where the number of events and risks at each time point is replaced by the expectation value of the number of individual descriptions of concepts. We present an application of this proposed model using primary brain tumor patient data. As a result, we obtained a new interpretation of the data in comparison to the classical survival tree modeling methods.

Keywords: symbolic data, interval-valued variables, survival tree, CART

1 Introduction

In the field of medical research, the prediction of survival time based on patient information is an important problem. Because of the presence of censored data, this problem cannot be treated as a simple regression problem. One choice for modeling these data is the survival tree, which is constructed in a non-parametric framework. This model has an important advantage in that the relationship between the covariates and hazards can be easily determined from the hierarchical structure. Tree-based methods were introduced by Morgan and Sonquist [1] and developed by Breiman et al. [2] as the Classification and Regression Trees (CART) algorithm, which is extensively used for constructing tree models. The CART algorithm is composed of splitting, pruning, and selection steps. The learning samples are recursively dichotomized in the splitting step and an optimal-sized tree model is constructed in the pruning and selection steps.

Gordon and Olshen [3] were the first to use the tree-structured model for survival data. They constructed a model in which the L^1 -Wasserstein distance between the Kaplan-Meier survival curves of two child nodes is maximized by splitting. As a natural expansion of the fact that the estimation in each node is given by a nonparametric framework, Leblanc and Crowley [4] constructed a model using log-rank test statistics. Whereas the method proposed by Breiman et al. is based on minimizing the impurity measure in a model, Leblanc and Crowley maximized the splitting measure in the model by proposing a new pruning approach known as the split-complexity measure. Along with these approaches, various authors have proposed splitting criteria and methods for constructing a tree-structured model using survival data [5–10]. A comparative study of these splitting criteria has been performed Shimokawa et al. [11].

In these approaches, each patient is treated as a statistical unit. This classical representation is considered obvious in many scenarios because we are usually interested in identifying the variables that can explain the events of patients who has the near values. However, in some scenarios, we are interested in treating the

On Behalf of the Committee of Brain Tumor Registry of Japan Supported by the Japan Neurosurgical Society

***Corresponding author: Asanao Shimokawa**, Department of Mathematics, Tokyo University of Science, Tokyo, Japan, E-mail: shimokawa@rs.tus.ac.jp

Yoshitaka Narita, Department of Neurosurgery and Neuro-Oncology, National Cancer Center Hospital, Tokyo, Japan

Soichiro Shibui, Department of Neurosurgery, Teikyo University School of Medicine, University Hospital Mizonokuchi, Kanagawa, Japan

Etsuo Miyaoka, Department of Mathematics, Tokyo University of Science, Tokyo, Japan

aggregate data as a statistical unit. For example, the goal of one study was to determine whether the type of hospital pathway followed by a patient is predictive of one-year survival Quantin et al. [12]. Because the pathway is considered to be a mediating variable for other prognostic factors, as it is for other factors, this is problematic. They addressed this problem by aggregating the patients who followed a given pathway and constructed a classification tree based on a symbolic analytic method. This approach enabled them to successfully obtain a new interpretation of the data.

In classical analysis, the patient data of p random variables are represented as a single point in p -dimensional space. In contrast, the aggregated data is represented as a p -dimensional hyperrectangle in $\mathbb{R}^p$ for continuous variables or a Cartesian product of p distributions for categorical variables in a symbolic analytic method. In Bertrand and Goupil [13], the basic statistics for interval-valued symbolic variables are given based on the assumption that an individual concept is uniformly distributed over its interval. For a detailed description of symbolic variables, see Billard and Diday [14], Bock and Diday [15], Diday and Noirhomme-Fraiture [16]. In addition, the basic ideas and methods for symbolic data analysis are described in Billard and Diday [17].

As tree-structured modeling methods are based on a symbolic approach, a few methods have been proposed. In Mballo and Diday [18], a classification tree method based on interval-valued explanatory variables was proposed. In their research, the response variable was assumed to be a classical categorical variable, and the Kolmogorov-Smirnov criterion was used for the splitting rule. As introduced earlier, Quantin et al. [12] presented the classification tree that was constructed based on modal multi-valued explanatory variables. In Shimokawa et al. [19], a method was proposed for the construction of a regression tree based on interval-valued response and explanatory variables. In this research, the response variable was treated as a modal binary variable.

Although several symbolic approaches for survival data have been proposed, none of the modeling methods based on survival functions are constructed of symbolic variables of aggregate data. Hence, in this paper, we propose a new approach to model aggregate survival data based on the CART algorithm. In our model, a concept can be included in several terminal nodes in a tree because each concept of aggregate data has a hyperrectangle or multiple points in the covariate space. The survival function, comprising concepts that are partially included in a node, is then constructed based on the Kaplan-Meier method [20], where the number of events and risks at each time point is replaced by the expectation value of the number of individual descriptions of concepts. The distance measure between two survival functions obtained by this method is used as the splitting criterion for a node to construct a binary tree model. The pruning and selection of the tree is performed using the split-complexity measure and bootstrap methods.

The remainder of this paper is organized as follows. In Section 2, we present the motivating example which will analyse by proposed approach in later. In Section 3, we introduce the notation for symbolic data and tree models as well as propose the construction method for survival functions based on concepts that may be partially included in a node. In Section 4, the construction algorithm for the CART-based tree-structured model using aggregate survival data is described. In Section 5, the results of simple simulation studies are described. The results of applying the proposed model to the example data are described in Section 6. Finally, we conclude this paper in Section 7.

2 Motivating example

As an example of motivation to analysis the data by aggregating, we present an application of the proposed method using data of patients with primary brain tumors conducted by the Japan Neurosurgical Society [21, 22]. These data consist of cases that were newly diagnosed during 2001–2004 and the data were collected from 2009 until the end of 2012 by 109 institutions consisting of universities, colleges, and hospitals with boards authorized by the Japan Neurosurgical Society. In our analysis, an event is defined as recurrence or death resulting from any cause. The covariates for this data are listed in Table 1 and the Kaplan-Meier survival curve obtained by these patients is shown in Figure 1. We used the data of 8,231 patients without any missing covariate values.

Table 1: Covariates used for the example survival tree construction.

Z_1 :	Age
Z_2 :	Number of tumors
Z_3 :	Greatest diameter of tumor
Z_4 :	KPS before treatment
Z_5 :	KPS after initial treatment
Z_6 :	Gender {(1) Male or (2) Female}
Z_7 :	Laterality of tumor {(1) right, (2) left, (3) across both sides, (4) on both sides, (5) central neuroaxis, or (6) other}
Z_8 :	Initial treatment {(1) surgery, (2) surgery + RT, (3) surgery + CRT, (4) surgery + CTX, (5) CRT, (6) RT, (7) CTX, (8) observation alone, or (9) other}
Z_9 :	Initial surgery {(1) none, (2) decompressive craniectomy, (3) biopsy, (4) 1–50 % resection, (5) 50–75 % resection, (6) 75–95 % resection, (7) 95–99 % resection, (8) 100 % total resection, (9) extended tumor resection, (10) shunt, (11) drainage, or (12) other}
Z_{10} :	Initial radiotherapy {(1) none, (2) before surgery, (3) intraoperative, (4) after surgery, or (5) at recurrence}
Z_{11} :	Initial chemotherapy {(1) none, or (2) yes}

Note: RT: Radiotherapy, CRT: Chemoradiotherapy, and CTX: Chemotherapy.

Figure 1: Estimated survival function obtained from all patient data (vertical axis: $\hat{S}(x)$, horizontal axis: days).

Data aggregation is based on three criteria: patient neighborhood, institution providing treatment, and tumor histology according to the WHO classification. In the case of the patient neighborhood, we are interested in identifying the variables that could explain the recurrence or death of patients who lived in the same neighborhood. In the case of the second criterion, we are interested in identifying the variables that could explain the recurrence or death of patients who were treated in the same institution. These two criteria for aggregation cannot be considered to be potential predictor variables for death or recurrence at the same level as other variables such as the patient age or number of tumors. However, they can be considered to be intermediate variables between the predictor variables and death or recurrence. That is, variables such as the patient age or degree of severity affect the neighborhood in which they choose to live and the institution the patient chooses for treatment. In turn, the neighborhood and institution affect the decision or timing of treatment, and this could affect death or recurrence. The third criterion is the tumor histology according to the WHO classification. In this case, the class of tumor histology can be considered to affect the death or recurrence of patients. Our goal in this case is to identify the variables that can categorize the class of histology that affects the recurrence or death of patients. The results of applying the proposed model to the data are described in Section 7.

3 Notation, tree model, and survival function

3.1 Survival data aggregation and assumption

Let Y and C be the true survival and censored time, respectively. The observation time is then given by $X = \min(Y, C)$. The censor indicator is denoted by $\delta = I(Y \leq C)$, which is 1 if the observation includes an event and 0 if the observation is censored. Let $\mathbf{Z} = (Z_1, \dots, Z_p)$ be a p -dimensional covariate vector, which for each variable is given as a continuous or categorical variable. An observed learning sample is represented by $\mathcal{L}^c = \{(x_i, \delta_i, z_i), i \in \Omega = \{1, \dots, N\}\}$.

By aggregating the observed learning sample $\mathcal{L}^c$ based on some criterion, we can construct the concepts $\omega_1, \dots, \omega_m$ that have symbolic valued variables. For example, suppose we want to analyze some medical research data, as listed in Table 2. To focus the analysis on hospitals rather than on individual patients, we need to aggregate the observations of the clinical information within the same hospitals. The concept therefore becomes a hospital, and each variable takes the form of a symbolic variable, as shown in Table 3. In the case of Table 3, Hospital 1 takes the interval value of $[30, 60]$ with relative frequency $2/3$ for the event time, the value of 27 with relative frequency $1/3$ for the censoring time, the interval value of $[24, 52]$ for the variable of age, and the relative frequency of the variable Male is $1/3$.

Table 2: Example of patient data.

Sample	Hospital	Observation time	Censoring	Age	Gender
i		X	δ	Z_1	Z_2
Patient 1	Hospital 1	30	1	40	Female
Patient 2	Hospital 2	45	0	28	Female
Patient 3	Hospital 1	27	0	52	Male
Patient 4	Hospital 1	60	1	24	Female
Patient 5	Hospital 3	18	1	33	Male
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 3: Aggregate data by hospital.

Concept	Observation time	age	gender
ω_u	X	Z_1	Z_2
Hospital 1	$\{[30, 60], 2/3; [27, 27], 1/3\}$	$[24, 52]$	$\{\text{Male}, 1/3; \text{Female}, 2/3\}$
Hospital 2	$\{[14, 38], 4/7; [27, 45], 3/7\}$	$[28, 49]$	$\{\text{Male}, 2/7; \text{Female}, 5/7\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

We define the aggregate learning sample as $\mathcal{L} = \{(x_u, z_u); \omega_u \in E = \{\omega_1, \dots, \omega_m\}\}$. Let Ω_u be the set of $i \in \Omega$ that makes up the concept ω_u . The observation time X is then given as a modal interval-valued random variable, and the realization of concept ω_u takes the form

$$X_u = \{[a_{1u}, b_{1u}], p_{1u}; [a_{2u}, b_{2u}], p_{2u}\},$$

Where

$$\begin{aligned} a_{1u} &= \min_{i \in \Omega_u; \delta_i = 1} x_i, & b_{1u} &= \max_{i \in \Omega_u; \delta_i = 1} x_i, \\ a_{2u} &= \min_{i \in \Omega_u; \delta_i = 0} x_i, & a_{2u} &= \max_{i \in \Omega_u; \delta_i = 0} x_i, \end{aligned}$$

and

$$p_{1u} = \frac{\text{num}(i \in \Omega_u; \delta_i = 1)}{\text{num}(i \in \Omega_u)}, \quad p_{2u} = \frac{\text{num}(i \in \Omega_u; \delta_i = 0)}{\text{num}(i \in \Omega_u)}.$$

For example, in the case of Table 3, the concept ω_1 represents Hospital 1, and it consists of the set of patients who were treated in Hospital 1. The minimum and maximum event times of patients that make up ω_1 are $a_{11} = 30$ and $b_{11} = 60$, respectively. Similarly, the minimum and maximum censored times of patients included in ω_1 are $a_{21} = 27$ and $b_{21} = 27$, respectively. The relative frequencies of patients who experienced the event and censoring are $p_{11} = 2/3$ and $p_{21} = 1/3$, respectively.

For the covariates, if Z_j takes a continuous value ($j = 1, \dots, p$), then the realization of Z_j for concept ω_u is given by

$$Z_{ju} = [c_{ju}, d_{ju}],$$

where

$$c_{ju} = \min_{i \in \Omega_u} z_{ji}, \quad d_{ju} = \max_{i \in \Omega_u} z_{ji}.$$

In the case of Table 3, the value of age (Z_1) is continuous, and the minimum and maximum age of patients included in Hospital 1 (ω_1) are $c_{11} = 24$ and $d_{11} = 52$, respectively.

If Z_j takes a categorical value with possible values $F = \{f_{j1}, \dots, f_{jC_j}\}$, then the realization of Z_j for concept ω_u is given by

$$Z_{ju} = \{f_{jc}, p_{jcu}; c = 1, \dots, C_j\},$$

where

$$p_{jcu} = \frac{\text{num}(i \in \Omega_u; z_{ji} = f_{jc})}{\text{num}(i \in \Omega_u)}$$

In the case of Table 3, gender (Z_2) takes a categorical value with possible values $F = \{f_{21} = \text{Male}, f_{22} = \text{Female}\}$. The relative frequencies of male and female patients that are included in Hospital 1 are $p_{211} = 1/3$ and $p_{221} = 2/3$, respectively.

In this study, we assume that X is a discrete random variable and its possible values are $x_1 < x_2 < \dots$. Moreover, it is assumed that within both intervals $[a_{1u}, b_{1u}]$ and $[a_{2u}, b_{2u}]$, each individual description x is uniformly distributed across that interval. It follows that, for each x_k ($k = 1, 2, \dots$),

$$P(X = x_k) = \begin{cases} \frac{1}{b_{lu} - a_{lu} + 1}, & a_{lu} \leq x_k \leq b_{lu}, \\ 0, & \text{otherwise,} \end{cases}$$

for $l = 1, 2$. The observed frequency for x_k is then defined as

$$f_1(x_k) = \sum_{\omega_u \in E} \sum_{x_k \in [a_{lu}, b_{lu}]} \frac{1}{b_{lu} - a_{lu} + 1} p_{lu}. \quad (1)$$

For the continuous variable Z_j , we assume that the individual descriptions z for ω_u are uniformly distributed over the interval Z_{ju} . It follows that

$$P(Z \leq \xi) = \begin{cases} 1, & \xi < c_{ju}, \\ \frac{\xi - c_{ju}}{d_{ju} - c_{ju}}, & c_{ju} \leq \xi \leq d_{ju}, \\ 0, & \xi \geq d_{ju}. \end{cases}$$

For further details regarding symbolic data, see Billard and Diday [14].

3.2 Tree-structured model and survival function

The tree-structured model is a well-established technique whose goal is to split the data into groups with differing survival using covariates. The CART algorithm for constructing the model using a data set

comprises three steps: splitting, pruning, and selection. In the splitting step, all samples are recursively divided and the maximum-size tree model is constructed. In the pruning and selection steps, an optimal-size tree is constructed from the maximum-size tree. We denote a node in tree T as t . Let $\tilde{T}$ be the set of terminal nodes of T that represents the nodes at the bottom layer of the tree. Further, S denotes the set of internal nodes of T , which are the nodes other than those at the bottom layer of the tree.

A split in each node t can be induced by any question of the form “ $Z \in R_t?$,” where $R_t \subset R^p$ and R^p is the p -dimensional covariate space. The most widely used splitting rule consists of a single covariate Z_j . If Z_j is a continuous variable, then the rule is of the form “ $Z_j \leq s?$ ” If Z_j is a categorical variable with possible values $F = \{f_{j1}, \dots, f_{jC_j}\}$, then the rule is of the form “ $Z_j \in R_{tj}?$ ” where $R_{tj} \subset F$. In the classical data case, each sample is represented by a single point in R^p , and therefore it is included in a single terminal node of the tree. In our case, the aggregated data have symbolic variables instead, and each concept is represented as a p -dimensional hyperrectangle in the covariate space R^p , or a Cartesian product of p distributions. Each concept then has the possibility of separation by a split, and therefore, it is natural to construct a model such that a concept can be included in several terminal nodes of a tree.

Let $R_t = R_{t_1} \times \dots \times R_{t_p}$ be the region of a node t in R^p , where

$$R_{tj} = \begin{cases} \{[c_{tj}, d_{tj}]; c_{tj} \leq d_{tj}\}, & Z_j \text{ is a continuous variable,} \\ \{f_{j1}, \dots, f_{jC_j}\}, & Z_j \text{ is a categorical variable,} \end{cases}$$

and $C'_t \leq C_j$. The observed frequency of a concept ω_u included in node t is then given by

$$f_t(u) = \prod_{j=1}^p \left\{ I_j \frac{\|Z_{ju} \cap R_{tj}\|}{\|Z_{ju}\|} + (1 - I_j) \sum_{c \in R_{tj}} p_{jcu} \right\}, \quad (2)$$

where $\|\cdot\|$ is the length of the interval, and I_j is 1 if Z_j is a continuous variable and 0 if Z_j is a categorical variable. For example, in the case of Table 3, let the region of a node t be $R_t = R_{t_1} \times R_{t_2}$, where $R_{t_1} = [0, 30]$ and $R_{t_2} = \{f_{21}' = \text{Female}\}$. Then, the observed frequency of a concept ω_1 included in the node is given by $f_t(1) = (\|30-24\|/\|52-24\|) \times (2/3) = 1/7$.

The total observed frequency included in t is given by

$$f_t = \sum_{\omega_u \in E} f_t(u).$$

The sum of the observed frequency $f_t(u)$ of ω_u included in all terminal nodes $t \in \tilde{T}$ is 1, and the sum of the total observed frequency f_t included in all terminal nodes $t \in \tilde{T}$ is m .

In the analysis of a survival tree based on classical data, the survival function of each node $\hat{S}_t(x_k)$ is constructed using the learning samples included in the node based on the Kaplan-Meier method [20]. Because each learning sample is included in a single terminal node, the amount of information in a sample used for estimating $S_t(x_k)$ is considered to be equal to other samples in the same node. However in our approach, a concept can be included in several terminal nodes in a tree. Hence, we need to define the estimation method for the survival function of each node considering the amount of information for the concept included in that node. Based on the Kaplan-Meier method, we define the estimator of the survival function for node t using the observed frequency of concepts that are included in t as

$$\hat{S}_t(x_k) = \begin{cases} 1, & x_k < a_1(1), \\ \prod_{k'=1}^k \left(1 - \frac{d_t(x_{k'})}{n_t(x_{k'})}\right), & x_k \geq a_1(1), \end{cases} \quad (3)$$

where $a_1(1) = \min_{\omega_u} (a_{1u})$, and $d_t(x_k)$ represents the number of events for x_k included in node t , weighted by the observed frequencies of concepts (1) and (2) as

$$d_t(x_k) = \sum_{\omega_u \in E; x_k \in [a_{1u}, b_{1u}]} f_t(u) \frac{1}{b_{1u} - a_{1u} + 1} p_{1u}. \quad (4)$$

Similarly, $n_t(x_k)$ represents the number of risks for x_k included in node t , weighted by the observed frequencies of concepts as

$$\begin{aligned} n_t(x_k) = & \sum_{\omega_u \in E; x_k \leq b_{1u}} f_t(u) \frac{b_{1u} - \max(a_{1u}, x_k) + 1}{b_{1u} - a_{1u} + 1} p_{1u} \\ & + \sum_{\omega_u \in E; x_k \leq b_{2u}} f_t(u) \frac{b_{2u} - \max(a_{2u}, x_k) + 1}{b_{2u} - a_{2u} + 1} p_{2u}. \end{aligned} \quad (5)$$

The first term on the right-hand side of eq. (5) represents the weighted sum of the event data, and the second term represents the weighted sum of the censored data.

In the case of Table 3 as an example, suppose the observed frequency of the concept ω_1 included in a node t is given by $f_t(1) = 1/7$. Moreover, it is assumed that we want to estimate the survival function for node t at $x_k = 40$ by way of calculation. Then the contributions of the concept ω_1 to the number of events $d_t(40)$ and risks $n_t(40)$ are $1/7 \times 1/(60-30+1) \times 2/3$ and $1/7 \times (60-40+1)/(60-30+1) \times 2/3$ respectively. Since $x_k = 40 > 27 = b_{21}$, ω_1 does not contribute to the second term on the right-hand side of eq. (5).

We can interpret this proposed survival estimator as follows. Because we assume the individual description of variable X for each concept is uniformly distributed across the observed interval, the term containing the fraction multiplied by p_{1u} in eq. (4) represents the probability that the description included in concept ω_u will experience an event at x_k . Similarly, the terms of the fractions multiplied by p_{1u} and p_{2u} in eq. (5) represent the probabilities that the description in ω_u is included in the risk set at x_k . Moreover, the number of observed frequencies $f_t(u)$ in eqs (4) and (5) can be considered to be the probability that the descriptions that are included in concept ω_u will be included in node t . Therefore, $d_t(x_k)$ represents the expected value that the individual descriptions of the learning samples will be included in node t and experience an event at x_k , for which the sum of the number of individual descriptions in each concept is standardized to 1. Similarly, $n_t(x_k)$ represents the expected value that the individual descriptions of the learning samples will be included in node t and the risk set at x_k . Therefore, the proposed survival estimator is constructed by replacing the number of observations in the Kaplan-Meier method with these expected values for the number of individual descriptions. If a data set consists of classical variables only, then this estimator takes the same value as the Kaplan-Meier estimate.

Based on this survival function, we construct a tree-structured model. By constructing a model using this framework, we expect to obtain a more flexible model than one that separates the data completely. At the same time, we expect to retain the interpretive ease of a tree-structured model. That is, if a concept is included in a single terminal node t , then the concept is considered to be sufficiently suitable for that node. On the other hand, if the concept is included in several terminal nodes, the concept is not sufficient to represent survival using a single node. In other words, it is uncertain in which region of the terminal nodes this concept should be included. This uncertainty is reflected by separating the concept using the number of observed frequencies as a criterion of the uncertainty. Therefore, in our approach, the existence of the splitting rule in each node is intended to provide an effective weighting rule for concepts rather than to divide the data completely.

4 Tree construction algorithm

4.1 Splitting method on aggregate data

In the splitting step of the CART algorithm, a learning sample set $\mathcal{L}$ is recursively divided and a maximum-sized tree is constructed. As mentioned in the previous section, each concept in $\mathcal{L}$ can be included in several terminal nodes in our model. Here, we consider the division of a node t into two child nodes t_L and t_R . Let $\mathcal{L}_t = \{(x_u, \mathbf{z}_u, f_t(u)); \omega_u \in E\}$ be the learning samples in t , where $f_t(u)$ is the observed frequency of ω_u included in t and is defined according to eq. (2).

If a continuous variable Z_j is considered for splitting t and the form of the rule is defined as “ $Z_j \leq s$?”, then the total observed frequencies included in the left child node t_L is given by

$$\begin{aligned} f_L &= \sum_{\omega_u \in E} f_L(u) \\ &= \sum_{\omega_u \in E} \frac{\min(s, d_{ju}^*) - \min(s, c_{ju}^*)}{d_{ju}^* - c_{ju}^*} f_t(u), \end{aligned}$$

where $[c_{ju}^*, d_{ju}^*] = [\max(c_{ju}, c_{tj}), \min(d_{ju}, d_{tj})]$ is the interval of Z_{ju} included in R_{tj} . Similarly, the total of the observed frequencies included in the right child node t_R is given by

$$\begin{aligned} f_R &= \sum_{\omega_u \in E} f_R(u) \\ &= \sum_{\omega_u \in E} \frac{\max(s, d_{ju}^*) - \max(s, c_{ju}^*)}{d_{ju}^* - c_{ju}^*} f_t(u). \end{aligned}$$

For example, in the case presented in Table 3, suppose that the age Z_1 and a threshold $s=28$ are considered for splitting a node t . As illustrated in previous examples, suppose the region of t for Z_1 is $R_{t1} = [0, 30]$, and the observed frequency of the concept ω_1 included in the node is $f_t(1) = 1/7$. Then the interval of $Z_{11} = [24, 52]$ included in R_{t1} is $[c_{11}^*, d_{11}^*] = [24, 30]$, and the contribution of the concept ω_1 to the total observed frequencies included in the left child node of t is given by $f_L(1) = \{\min(28, 30) - \min(28, 24)\} / (30 - 24) \times 1/7 = 2/21$. In the same way, the contribution of ω_1 to the right child node is given by $f_R(1) = \{\max(28, 30) - \max(28, 24)\} / (30 - 24) \times 1/7 = 1/21$.

If variable Z_j is a categorical variable and the form of the rule is defined as “ $Z_j \in R_{tj}$?”, where $R_{tj} = \{f_{j1}, \dots, f_{jC_L}\}$ is a subset of R_{tj} , then the total observed frequency included in the left child node t_L is given by

$$\begin{aligned} f_L &= \sum_{\omega_u \in E} f_L(u) \\ &= \sum_{\omega_u \in E} \frac{\sum_{c \in R_{tj}} p_{jcu}}{\sum_{c \in R_{tj}} p_{jcu}} f_t(u). \end{aligned}$$

Similarly, using the relative complement R_{tj}^c of R_{tj} in R_{tj} , the total observed frequency included in the right child node t_R is given by

$$\begin{aligned} f_R &= \sum_{\omega_u \in E} f_R(u) \\ &= \sum_{\omega_u \in E} \frac{\sum_{c \in R_{tj}^c} p_{jcu}}{\sum_{c \in R_{tj}^c} p_{jcu}} f_t(u). \end{aligned}$$

Using these observed frequencies, the survival functions $\hat{S}_L(x_k)$ and $\hat{S}_R(x_k)$ of the two child nodes can be estimated from eq. (3). Searching for the optimal splitting rule for node t that maximizes some improvement criterion $H(t)$ requires us to evaluate each possible splitting rule. In our study, $H(t)$ is defined to be a distance measure between survival functions estimated by two child nodes. In classical analysis, the distance measure for splitting in a survival tree was proposed by Gordon and Olshen [3], who used the L^1 -Wasserstein distance between the Kaplan–Meier survival curves for splitting. We use the same distance measure. Let m_L and m_R be the limits of $\hat{S}_L(x_k)$ and $\hat{S}_R(x_k)$, respectively. We define x_m to be

$$x_m = \min \left\{ \min(x_k; \hat{S}_L(x_k) = m_L), \min(x_k; \hat{S}_R(x_k) = m_R) \right\}.$$

We then define the distance measure between survival functions estimated by two child nodes as

$$H(t) = \sum_{k=1}^m \left| \hat{S}_L(x_k) - \hat{S}_R(x_k) \right|. \quad (6)$$

If Z_j is a categorical variable, then we calculate $H(t)$ for all possible combinations of categories. That is, the number of combinations (or splitting rules for Z_j) to be evaluated is $2^{C_j} - 1$. If Z_j is a continuous variable, we first sort the c_{ju} and d_{ju} that are included in region R_t in ascending order ($\xi(1)$, $\xi(2)$, $\dots$). Next, for each interval between the two points $\xi(l)$ and $\xi(l+1)$, we calculate $H(t)$ for the points, defined as

$$\begin{cases} \frac{\xi(l) + \xi(l+1)}{2}, & \text{there are no concepts in } [\xi(l), \xi(l+1)], \\ \xi(l), \xi(l+1), & \text{there are concepts in } [\xi(l), \xi(l+1)]. \end{cases}$$

For all possible splitting rules, we calculate $H(t)$ and find the split that maximizes it. By repeating this procedure, we construct the maximum-sized tree. The stopping rule for growing the tree is defined using the number of observed frequencies of concepts or events included in each of the terminal nodes.

4.2 Pruning and selection steps

We avoided the problem of over learning by using the pruning and selection steps from the CART algorithm. In the pruning step, the set of nested subtrees is obtained by recursively removing the splits in the maximum-size tree. For this purpose, we used the split-complexity measure [4]:

$$G_\alpha(T) = \sum_{t \in S} G(t) - \alpha|S|, \quad (7)$$

where $G(t)$ is a measure of the separation between t_L and t_R , and $|S|$ is the number of internal nodes of T , which is a complexity measure of the tree. The measure $G_\alpha(T)$ 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 for an arbitrary α is determined using these measures. If the value of α is 0, then the optimal subtree is the maximum-size tree. If α approaches ∞ , on the other hand, then a subtree consisting of the root node only (that is, a model that is not considered to be 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 guaranteed that these subtrees have nesting structures because of this algorithm.

The optimal subtree is selected from the set of pruned subtrees by determining an optimal tree size in the selection step. If a set of large samples is available, we can use the test sample method. That is, we divide the data $\mathcal{D}$ into a learning set $\mathcal{D}_l$ to grow the tree and a test set $\mathcal{D}_t$ to evaluate the performance of the tree. The maximum-size tree and nested subtrees are obtained using $\mathcal{D}_l$ and the data in $\mathcal{D}_t$ are sent down each of the subtrees. The $G(t)$ is then calculated for each internal node $t \in S$ using the data in $\mathcal{D}_t$. The subtree that maximizes $G_{\alpha_c}(T)$ is chosen as the best tree-structured model. The penalty α_c must be pre-determined. However, in general, a large test sample will not be available to us. In this case, we need to use resampling and bias correction methods to select the optimal subtree. In this study, we use the bootstrap algorithm for this purpose. A detailed description of this technique can be found in Leblanc and Crowley [4].

In eq. (7), $G(t)$ is calculated by using the log-rank test statistic in Leblanc and Crowley [4], which enables a theoretically optimal cost α_c to be selected in the selection step. That is, if $G(t)$ is approximately χ_1^2 distributed in the two-sample case when there is no difference between groups, a penalty $\alpha_c = 4$ corresponds roughly to a significance level of 5% for a split. If $\alpha_c = 2$, then the split corresponds to the AIC. The log-rank test can be interpreted as the square of the standardized difference between the observed and expected number of events based on the hypergeometric distribution with parameters $n_t(x)$, $d_t(x)$, and $n_l(x)$. In the classical data case, the observed number of events at each time point is given as a positive integer. However, in our approach, the number of events at each time point is not given as an integer, but as a positive real number instead. Therefore, we need to consider a continuous distribution of the number of events for calculating $G(t)$.

We overcame this problem by using the hypergeometric distribution with parameters $n_t(x_k)$, $n_L(x_k)$, and $d_t(x_k)$, defined by the gamma functions:

$$f(y) = \begin{cases} \frac{1}{\eta} \frac{\alpha(y)\beta(y)}{\gamma}, & \max(0, n_L(x_k) + d_t(x_k) - n_t(x_k) - 1) \leq y \leq \min(d_t(x_k), n_L(x_k)), \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where $\alpha(y)$, $\beta(y)$, and γ are defined as follows:

$$\begin{aligned} \alpha(y) &= \frac{\Gamma(n_L(x_k) + 1)}{\Gamma(y + 1)\Gamma(n_L(x_k) - y + 1)}, \\ \beta(y) &= \frac{\Gamma(n_t(x_k) - n_L(x_k) + 1)}{\Gamma(d_t(x_k) - y + 1)\Gamma(n_t(x_k) - n_L(x_k) - d_t(x_k) + y + 1)}, \\ \gamma &= \frac{\Gamma(n_t(x_k) + 1)}{\Gamma(d_t(x_k) + 1)\Gamma(n_t(x_k) - d_t(x_k) + 1)}. \end{aligned}$$

Here, η is the normalization factor of the distribution:

$$\eta = \int_{\max(0, n_L(x_k) + d_t(x_k) - n_t(x_k) - 1)}^{\min(d_t(x_k), n_L(x_k))} \frac{\alpha(y)\beta(y)}{\gamma} dy$$

As is the case with classical data, if $n_t(x_k)$, $n_L(x_k)$, and $d_t(x_k)$ take only positive integers at all nodes and time points, then the value of η is given by 1, and eq. (8) is equivalent to the hypergeometric distribution with parameters $n_t(x_k)$, $n_L(x_k)$, and $d_t(x_k)$.

Based on this distribution eq. (8), we define the measure of separation between t_L and t_R as

$$G(t) = \frac{\{\sum_k (d_L(x_k) - E_L(x_k))\}^2}{\sum_k \text{Var}_L(x_k)}, \quad (9)$$

where $E_L(x_k)$ and $\text{Var}_L(x_k)$ are the expectation and variance, respectively, of the number of events for t_L at x_k , obtained from eq. (8). In the case of classical data, the value of eq. (9) is equivalent to the value of the ordinary log-rank test statistic. In the case of symbolic data, unfortunately, η , $E_L(x_k)$, and $\text{Var}_L(x_k)$ cannot be derived by analytical calculation. However, we can use a numerical analytical approach to calculate the value of eq. (9). In this study, we used the `int` function of MATLAB (Version R2015a, Mathworks Inc., Natick, MA, USA).

Equation (9) is only changed by the ordinary log-rank test statistics to take a real number between 0 and 1 for each of its composition elements. Therefore, if an element of $G(t)$ takes a value between 0 and 1, it will take a value of between the log-rank test statistics for which the element is 0 and 1. When all elements of $G(t)$ take the value 0 or 1 as in the case of classical data, its value is equivalent to the value of the ordinary log-rank test statistics.

Of course, in the splitting step, we can use $G(t)$ of eq. (9) instead of $H(t)$ of eq. (6). However, because the calculation of $G(t)$ is time consuming, we recommend $H(t)$ for constructing the maximum-size tree and $G(t)$ for pruning and selection of the subtrees.

5 Simulations

5.1 Models and methods

We present simple simulation studies to examine the properties of the proposed approach in several situations. The purpose of these simulations is to compare the proposed symbolic approach with the classical approach in terms of the selection of a splitting rule in a node. The results of the selection performed by our proposed approach are considered to strongly depend on the criterion for aggregation of the data. We studied the influence of the relationship between the variables for aggregation of the data and construction of a tree by assuming several situations about that relationship and comparing the results.

The other purpose was to study the instability of the result obtained by using the proposed approach. It is well known that the selection of the splitting variable and splitting rule strongly depends on the distribution of observations in the learning sample [23]. Because of this, the instability of the tree-structured model results from small changes in the learning samples. Therefore, we consider the selected splitting rule in the root node to have the most influence on the result of the model.

We used the survival data generated from several models and divided them according to the ordinary classical approach and proposed symbolic approach. The variable Z' is generated from a discrete uniform distribution with $\{1, 2, \dots, m\}$. This variable is used for data aggregation in the proposed approach, and affects the generation of covariates. Based on Z' , two continuous covariates Z_1, Z_3 and two binary covariates Z_2, Z_4 are generated from six patterns.

In the first pattern, the covariates are generated as not being related to Z' :

$$C - \text{Model 1} : \begin{aligned} Z_1, Z_3 &\sim \text{Uni}[0, 1], \\ Z_2, Z_4 &\sim \text{Ber}(0.5), \end{aligned}$$

where $\text{Uni}[\alpha, \beta]$ and $\text{Ber}(p)$ are uniform distribution with parameters α and β , and the Bernoulli distribution with parameter p , respectively.

In the second pattern, Z_1 is generated to have a relationship in which the value is “clearly” divided by Z' :

$$C - \text{Model 2} : \begin{aligned} Z_1 &\sim I(Z' \leq m/2) \times \text{Uni}[0, 0.5] + I(Z' > m/2) \times \text{Uni}[0.5, 1], \\ Z_3 &\sim \text{Uni}[0, 1], \\ Z_2, Z_4 &\sim \text{Ber}(0.5), \end{aligned}$$

where $I(\cdot)$ denotes the indicator function, which is equal to 1 if “ $\cdot$ ” holds true and 0 otherwise. In the third pattern, on the other hand, Z_1 is generated in a similar manner to the second pattern but the value is “vaguely” divided by Z' :

$$C - \text{Model 3} : \begin{aligned} Z_1 &\sim I(Z' \leq m/2) \times \text{Uni}[0, 0.7] + I(Z' > m/2) \times \text{Uni}[0.3, 1], \\ Z_3 &\sim \text{Uni}[0, 1], \\ Z_2, Z_4 &\sim \text{Ber}(0.5). \end{aligned}$$

In both the second and third patterns, Z_2 – Z_4 are randomly generated in the same way as in the first pattern.

In the fourth pattern, Z_2 is generated to have a relationship in which the ratio of 0 and 1 is divided by Z' :

$$C - \text{Model 4} : \begin{aligned} Z_1, Z_3 &\sim \text{Uni}[0, 1], \\ Z_2 &\sim I(Z' \leq m/2) \times \text{Ber}(0.2) + I(Z' > m/2) \times \text{Ber}(0.8), \\ Z_4 &\sim \text{Ber}(0.5). \end{aligned}$$

In this case, the other covariates Z_1, Z_3 , and Z_4 are randomly generated in the same way as in the first pattern.

In the fifth pattern, Z_1 and Z_2 are generated as in the second and fourth patterns, respectively. Z_3 and Z_4 are randomly generated as in the first pattern.

$$C - \text{Model 5} : \begin{aligned} Z_1 &\sim I(Z' \leq m/2) \times \text{Uni}[0, 0.5] + I(Z' > m/2) \times \text{Uni}[0.5, 1], \\ Z_2 &\sim I(Z' \leq m/2) \times \text{Ber}(0.2) + I(Z' > m/2) \times \text{Ber}(0.8), \\ Z_3 &\sim \text{Uni}[0, 1], \\ Z_4 &\sim \text{Ber}(0.5) \end{aligned}$$

Finally, in the sixth pattern, Z_1, Z_2 , and Z_4 are randomly generated in the same way as in the first pattern, and Z_3 is generated as follows:

$$C - \text{Model 6} : \begin{aligned} Z_1 &\sim \text{Uni}[0, 1], \\ Z_2, Z_4 &\sim \text{Ber}(0.5), \\ Z_3 &\sim I(Z' \leq m/2) \times \text{Uni}[0, 0.5] + I(Z' > m/2) \times \text{Uni}[0.5, 1] \end{aligned}$$

As explained later, Z_3 is not related to the true hazard model that generates the survival time for each data, it is only related to Z' .

C-Model 1 represents the situation that the aggregate criteria of data is not related at all to the covariates. In C-Model 2 and 3, the continuous covariate Z_1 which is related to the true hazard model has relationship to the aggregate criteria of data. The only difference of these models is that this relationship is clear or vague. C-Model 4 represents the situation that the dichotomous covariate Z_2 which is related to the true hazard model has vaguely relationship to the aggregate criteria. The degree of vagueness of the relationship, that is the setting of parameters of the distributions, has been arbitrarily selected from a range that makes sense. C-Model 5 represents the situation that the aggregate criteria is related to these both covariates. In C-Model 6, we consider the situation that the aggregate criteria is only related to the covariate Z_3 which is not related to the true hazard model.

As described in above, these six patterns for generating the covariates are highly simplified examples. In real data, more complex situations would be exist. However, our main purpose of this simulation is to examine the tendency of properties of our proposed approach, we think it is sufficient to consider these simplified situations.

Suppose we have an exponential model, the survival function of which is given by

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

where the parameter λ represents a constant hazard. We suppose three patterns of the hazard model:

$$\begin{aligned} H - \text{Model 1} &: \lambda = I(Z_1 \leq 0.5) \times 0.3 + I(Z_1 > 0.5) \times 1 \\ H - \text{Model 2} &: \lambda = I(Z_2 = \{0\}) \times 0.3 + I(Z_2 = \{1\}) \times 1 \\ H - \text{Model 3} &: \lambda = 0.3 \times \exp(Z_1 + Z_2) \end{aligned}$$

In the first model, the hazard is generated from the left and right child node models that are divided by the continuous covariate Z_1 . In the second model, the hazard is generated from the child node models that are divided by the binary covariate Z_2 . In the third model, the hazard is generated from the proportional hazard model, which is related to Z_1 and Z_2 at a time. Throughout all the models, the covariates Z_3 and Z_4 are redundant.

In the same reason of setting of C-Models, we consider the highly simplified situations for hazard models. That is, we assume the constant hazard model which is a most simple situation in survival data. Moreover, we assume the tree-structured models which have two terminal nodes only in H-Model 1 and 2. In H-Model 3, we assume the simple hazard model. The setting of values in H-Models has been arbitrarily selected from a range that makes sense.

The censoring rates that are used are approximately 50 % of those that would be obtained by using uniform random numbers. Because we are considering large datasets, the number of learning samples N is set to 5000. We studied the influence of the number of data elements used for aggregation by setting the number of concepts m to 10, 50, 100, and 200 in the setting of the first and second covariate patterns when survival data is generated by Hazard Model 1. In all the other settings, the number of concepts m is only set to 200. The simulations are repeated 300 times for every pattern.

5.2 Results

The results of the simulations are listed in Tables 4–6. All tables list the probabilities of selecting the covariates Z_1 and Z_2 for splitting the medians of threshold values that are estimated for splitting the covariate when the covariate Z_1 is selected. Tables 4, 5, and 6 represent the results of simulations when H – Model 1, 2, and 3 are used, respectively.

The survival data presented in Table 4 are generated from H – Model 1, where the hazard is divided by Z_1 . From the results of C – Model 1, if the variable Z' for aggregating the data in the proposed approach is not related to the covariates, it is difficult to detect the variable Z_1 in the case of concepts of a small size. If the size of concepts were larger, that is, if the number of samples included in a concept were smaller, the probability of detecting the variable Z_1 would become large. However, the thresholds for the splitting rule

Table 4: Results of simulations for splitting in the case of H – Model 1.

		Classical approach			Symbolic approach		
		Z_1 (%)	med. thr.	Z_2 (%)	Z_1 (%)	med. thr.	Z_2 (%)
C – Model 1	10	100.0	0.498	0.0	0.0	–	47.3
	50				0.0	–	50.7
	100				36.7	0.091	14.3
	200				62.7	0.064	1.7
C – Model 2	10	100.0	0.498	0.0	100.0	0.500	0.0
	50				100.0	0.499	0.0
	100				100.0	0.497	0.0
	200				100.0	0.497	0.0
C – Model 3	200	100.0	0.499	0.0	100.0	0.307	0.0
C – Model 4	200	100.0	0.497	0.0	38.3	0.063	40.7
C – Model 5	200	100.0	0.497	0.0	100.0	0.498	0.0
C – Model 6	200	100.0	0.499	0.0	16.3	0.064	0.0

Note: Z_1 (%): percentage of Z_1 selected for splitting, med. thr.: median of thresholds when Z_1 is selected for splitting, Z_2 (%): percentage of Z_2 elected for splitting, pattern1-6: pattern according to which covariates are generated.

Table 5: Results of simulations for splitting in the case of H – Model 2.

		Classical approach			Symbolic approach		
		Z_1 (%)	med. thr.	Z_2 (%)	Z_1 (%)	med. thr.	Z_2 (%)
C – Model 1	0.0	–		100.0	3.0	0.938	94.3
C – Model 2	0.0	–		100.0	52.0	0.500	45.7
C – Model 3	0.0	–		100.0	33.3	0.494	63.3
C – Model 4	0.0	–		100.0	0.0	–	100.0
C – Model 5	2.7	0.042		97.3	100.0	0.499	0.0
C – Model 6	0.0	–		100.0	1.3	0.063	46.3

Note: Z_1 (%): percentage of Z_1 selected for splitting, med. thr.: median of thresholds when Z_1 is selected for splitting, Z_2 (%): percentage of Z_2 selected for splitting, pattern1-6: pattern according to which covariates are generated.

Table 6: Results of simulations of splitting in the case of H – Model 3.

		Classical approach			Symbolic approach		
		Z_1 (%)	med. thr.	Z_2 (%)	Z_1 (%)	med. thr.	Z_2 (%)
C – Model 1	10.3	0.035		89.7	17.7	0.062	71.3
C – Model 2	8.3	0.051		91.7	100.0	0.500	0.0
C – Model 3	6.0	0.058		94.0	100.0	0.307	0.0
C – Model 4	9.3	0.034		90.7	0.0	–	100.0
C – Model 5	39.3	0.067		60.7	100.0	0.498	0.0
C – Model 6	10.7	0.036		89.3	7.7	0.063	30.0

Note: Z_1 (%): percentage of Z_1 selected for splitting, med. thr.: median of thresholds when Z_1 is selected for splitting, Z_2 (%): percentage of Z_2 selected for splitting, pattern1-6: pattern according to which covariates are generated.

are very small. The results of C – Model 2 show that, if the variable Z' is related to the covariates, the variable Z_1 and the threshold are detected as in the classical case regardless of the number of concepts.

The results of C – Model 3 show that, if the continuous covariate Z_1 is generated to have a relationship in which the value is “vaguely” divided by Z' , the variable Z_1 is detected as important in all simulations. The median of thresholds of Z_1 in the simulations is about 0.3, and is the reason this is considered to be

vagueness in C – Model 3. That is, based on the assumption that the individual descriptions are uniformly distributed across the interval valued variables Z_1 in the proposed approach, the value of the threshold is directly affected by the vagueness in C – Model 3.

The results of C – Model 4 show that, if the binary covariate Z_2 is generated to have a relationship in which the ratio of 0 and 1 is divided by Z' , then Z_1 and Z_2 are detected as important in a nearly identical ratio. On the other hand, if Z_1 and Z_2 are related to Z' as in the case of C – Model 5, Z_1 is detected as important. The results of C – Model 6 show that, if variable Z_3 , which is not related to the true hazard model, is generated to have a relationship in which the value is clearly divided by Z' , it is difficult to detect Z_1 .

Based on these results, we can see that the detection rate of Z_1 is strongly dependent on whether that variable is related to Z' , which is used for aggregating the data in the proposed approach. Because H – Model 1 has a very simple tree structure and the number of samples is large, i. e., 5,000, Z_1 is always detected as important in the results of the classical approach.

In Table 5, survival data are generated from H – Model 2 where the hazard is divided by Z_2 . For the same reason as in the case of H – Model 1, the classical approach well detects Z_2 for splitting. The results of C – Model 2, 3, and 5, show that when Z_1 is related to Z' , the detection rate of Z_1 for splitting is increased. On the other hand, if Z_1 is not related to Z' as in the case of C – Model 1 and 4, the detection rate of Z_2 is high. In comparison with the results in Table 4, the reason for the decreasing detection rate of the variable that is related to the true hazard model is considered to be the number of offered cutpoints. That is, in the construction of a tree-structured model, the selection of covariates for splitting tend to favor those offering more cutpoints such as a continuous variable.

In Table 6, the survival data are generated from H – Model 3, where the hazard is generated from the proportional hazard model that is related to Z_1 and Z_2 at a time. In this case, the instability of the results in the classical approach is increased. The results of C – Model 1 and 6 show that, if the covariates are not related to Z' , or if the covariate Z_3 that is not related to the hazard model is only related to Z' , the splitting in the proposed approach causes instability. On the other hand, the results of C – Model 2–6 show that the splitting rule in the proposed approach has stability whenever either or both of Z_1 and Z_2 are related to Z' . The reason for this is the following: we consider the variability of each of the units in the data to decrease by aggregating the data as symbolic data, and as a result, the instability of the splitting decreases in cases in which it is difficult to obtain stable results with the classical approach.

6 Analysis results of patients with primary brain tumors

6.1 Survival tree analysis based on classical data

For the first analysis, we constructed a survival tree based on the patient data. In this analysis, each patient was considered to be a statistical unit, and our goal was to recognize effective prognostic factors for tumor patients. There are several proposed splitting criteria for constructing a survival tree [3–10]. For comparative purposes, we used the L^1 -Wasserstein distance between Kaplan-Meier survival curves for splitting [3] because the distance measure between survival functions is used as the splitting rule for constructing a tree-structured model from the concepts in this study. We also show the result obtained by using the log-rank test statistics as the splitting rule because it is the most popular method [5, 7, 8]. Because of the large number of patients ($N=8,231$), we set a relatively large value as stopping criterion for node splitting, i. e., a minimum of 50 events in a node. Referring to the recommended value $2 \leq \alpha_c \leq 4$ in Leblanc and Crowley [4], the pruning step used the split-complexity measure with $\alpha_c=3$. We used 15-bootstrap in the selection step because the number of bootstraps that is generally used is 10 and the number of patients is large.

Figure 2: Tree generated from the patient data using the L^1 -Wasserstein distance between Kaplan-Meier survival curves for splitting.

Figure 3: Tree made from the patient data using the Log-rank test statistics for splitting.

Because the resulting tree has many nodes, we show a sample of the results in Figures 2 and 3. The values in each node represent the number of patients in the node, and the value in parentheses represents the number of events. From the results obtained using the L^1 -Wasserstein distance between the Kaplan-Meier survival curves for splitting, the most important prognostic factor is the initial treatment with a splitting rule of $Z_8 \in \{1, 2, 8, 9\}$. The next most important factor is the KPS after initial treatment, and following it, the initial surgery or the KPS before treatment are also found to be important. As a result, the patients are separated into 24 sub-groups.

On the other hand, when we constructed a tree-structured model using the Log-rank test statistics for splitting, the patients are separated by the rule of $Z_8 \in \{3, 5\}$. Regarding patients with $Z_8 \in \{3, 5\}$, age is the next important factor with a splitting point of 39. In the other group, the initial surgery is the next most important factor. The KPS after initial treatment is the third most important factor for the patients in three grandchild groups. In the other grandchild group, the splitting is led by the initial treatment again. The final tree model has 27 terminal nodes.

6.2 Survival tree analysis based on aggregate data

In the second analysis, we aggregated the data based on three criteria and constructed tree-structured models using the proposed approach. The first criterion was patients' neighborhood, and 49 concepts were constructed from the patient data. The second criterion was the institution where the patient was treated, and 107 concepts were constructed. The third criterion is the tumor histology according to the WHO classification, and 122 concepts were constructed from the patient data. The survival functions of these concepts, estimated by eq. (3), are shown in Figures 4, 5, and 6.

Figure 4: Estimated survival function obtained from aggregate data (neighborhood) (vertical axis: $\hat{S}(x)$, horizontal axis: days).

Figure 5: Estimated survival function obtained from aggregate data (institution) (vertical axis: $\hat{S}(x)$, horizontal axis: days).

We analyzed the aggregate data by setting the node splitting stop condition to a minimum of three observed frequencies for an event. The penalty for the split-complexity measure was $\alpha_c = 3$, and 15-bootstrap was used in the selection step.

The tree structure for the aggregate by patient neighborhood is given in Figure 7. The values in each node represent the number of observed frequencies for concepts in the node, and the value in parentheses represents the number of observed frequencies for the events. The concepts are divided by only one rule: whether the initial radiotherapy is after surgery or not. For both terminal nodes, the survival curves estimated by eq. (3) are shown in Figure 8. From this result, it is clear that the concepts that perform

Figure 6: Estimated survival function obtained from aggregate data (tumor histology) (vertical axis: $\hat{S}(x)$, horizontal axis: days).

Figure 7: Tree made from the 49 concepts aggregated by patient neighborhoods.

Figure 8: Kaplan-Meier survival curves for each terminal node of Figure 7 (vertical axis: $\hat{S}(x)$, horizontal axis: days).

initial radiotherapy after surgery in many cases have a poor prognosis. In other words, if we hypothesize that many patients receive treatment at a hospital in their own neighborhood, the areas that do not use irradiation treatment after surgery in many cases have a good prognosis for tumor patients. A small-size tree is obtained from this aggregation data because the differences of prognostic factors between neighborhoods of patients are not large, and effective factors for prognosis cannot be detected from the data.

The analysis of the 107 concepts that were aggregated by the institution of the patients produces the tree shown in Figure 9. The estimated survival curves for each terminal node are shown in Figure 10. The resulting tree has five terminal nodes, and the most important factor is whether the initial surgery of patients involves a biopsy or not. Institutions with a high proportion of biopsy treatments for initial surgery

Figure 9: Tree made from the 107 concepts aggregated by the institution that treated the patients.

Figure 10: Kaplan-Meier survival curves for each terminal node of Figure 9 (vertical axis: $\hat{S}(x)$, horizontal axis: days).

showed a poor prognosis. The 8.7 observed frequencies for all concepts are included in that node (node 1), including the 3.8 observed frequencies for the event. The second most important factor is defined by the initial treatment. If a concept has a high probability of an initial treatment of surgery and chemoradiotherapy, a poor prognosis is indicated. The number of observed frequencies included in that node (node 2) is 15.6. Similarly, the initial radiotherapy and KPS after initial treatment are important factors for the splitting rule in the resulting model. Interestingly, in the fourth level of the tree, the concepts that have a high probability of a high KPS after initial treatment (node 5) have good prognoses. The important factors obtained by this analysis are similar to the analysis based on classical variables, although the most important factor for each model is different.

The analysis of the 122 concepts that were aggregated by tumor histology produced the tree shown in Figure 11. The estimated survival curves for each terminal node are shown in Figure 12. In this case, the six terminal nodes are included in the model, and four variables are given as important factors associated with

Figure 11: Tree made from the 122 concepts aggregated by tumor histology according to the WHO classification.

Figure 12: Kaplan-Meier survival curves for each terminal node of Figure 11 (vertical axis: $\hat{S}(x)$, horizontal axis: days).

the prognosis of patients. Age is defined as the most important prognostic factor; the histology class that includes very young children shows poor prognoses up to 50 days after the observation started. On the other hand, the subsequent 50 days, for the class of patients 52 years old or older, where the initial treatment was surgery plus chemoradiotherapy or some other treatment, showed the worst prognosis in many cases. As another prognostic factor, the histology class that gave patients chemotherapy in many situations showed a worse prognosis than the other classes.

Note that, in sharp contrast to the analysis based on individual patients, initial radiotherapy Z_{10} is considered an important prognostic factor in all models obtained by aggregate data. Its variable does not appear until at least the fourth level in the trees obtained from classical data. Moreover, the most important factors vary according to the design of the statistical unit for analysis. This represents the importance of deciding what unit to use as the statistical unit. If our goal is an analysis based on aggregate data rather than on each patient, our approach allows us to obtain a new understanding of the data with the ease of interpretation allowed by a hierarchically structured model.

7 Conclusion

In this paper, we propose a new approach to the analysis of aggregate survival data based on a tree-structured model. For each aggregate of the data (concept), the approach considers the symbolic variables and constructs a tree-structured model in which a concept can be included in several terminal nodes. If a concept is separated in a node, the survival functions for both child nodes are estimated by the observed frequencies of the concepts included in those nodes. By constructing a model using this condition, we expect to obtain a more flexible model while retaining the ease of interpretation afforded by a tree-structured model.

By using simple simulation studies, we studied the properties of the proposed approach in several situations. We obtained three principal results. First, if the criterion for aggregating data is not related to the hazard model, the obtained splitting rule has instability with the possibility of returning results that are unrelated to the true hazard model. Second, the fact that the selection of covariates for splitting tends to favor those offering more cutpoints is problematic, and this problem extends to the use of the classical approach for constructing a tree-structured model. Finally, the advantage of the proposed approach is the possibility that the instability associated with the tree-structured model in the classical approach is reduced by aggregating the data and by conducting the analysis based on the symbolic representation in some situations.

As an actual example of the proposed approach, we constructed a tree-structured model based on brain tumor patient data. For comparison, we first performed an analysis using an ordinary survival tree method on each patient. We next aggregated the patient data based on three criteria and constructed tree-structured models using the proposed approach. As a result, the most important factors in the resulting models varied according to the analysis, and we were able to obtain new interpretations of the data.

In this study, we treated observation survival time, censored time, and all the numerical and continuous symbolic predictors as interval valued variables. For ease of illustration of the proposed method, we also assumed the individual descriptions are uniformly distributed across those intervals. However, from a more practical point of view, we need to process these variables under more flexible assumptions. For example, we would prefer to assume a normal distribution of patients' weight aggregated by each address. Another example would be the preferred use of the empirical distributions of variables in circumstances under which it is difficult to predict the specific distribution for some variable. Then, our proposed approach could easily be extended to accommodate these situations by changing the formula for calculating observed frequencies to take into account the probabilities of individual descriptions across the interval.

The main issue with the framework of the tree-based method, is the instability of the resulting model. As mentioned in the simulation section, the splitting variable and splitting rule strongly depend on the distribution of observations, and as a result, the structure of the tree would immediately change if small changes were made to the learning samples. An effective way of addressing this concern would be to take into account the ensemble procedure. The use of a technique such as random forest would reduce the problem of instability, and as a result, the prediction accuracy of the model would be expected to increase. A good review of existing survival tree algorithms and recent survival ensemble methods has been published Zhou and McArdle [24].

Of course, the extension of our proposed approach to incorporate the use of these techniques can be considered. However, an extension of this nature would have to overcome a significant hurdle in that the calculation of $G(t)$ is very time consuming. One approach to overcome this problem would be to consider a method that only performs splitting by $H(t)$ rather than performing pruning for constructing each tree. In addition, it is difficult to use the ensemble method to interpret the proposed model. Because our proposed approach has the advantage of producing results that are easy to interpret based on the tree structure, the reduction of instability based on the ensemble method is not recommended when both the prediction and the obtained model itself are important. Therefore, the extension of our proposed approach to address the above issues and to overcome the instability of the result remains a future challenge.

Reducing the number of units by the aggregation of a very large number of data values has an advantage in practical analysis. For example, if values are missing from the data, or if a variable of the data has a large number of categories, we need to address these issues. Another advantage of analysis based on aggregate data is that it can highlight new aspects of the data based on the variable on which the researchers wish to focus. We believe that our proposed approach offers advantages in terms of the ease of interpretation and flexibility based on the fact that a single concept can be included in multiple terminal nodes.

References

1. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. *J Am Stat Assoc* 1963;58:415–34.
2. Breiman L, Friedman JH, Olshen RA, Stone C. *Classification and regression trees*. CA: Wadsworth, 1984.
3. Gordon L, Olshen RA. Tree-structured survival analysis. *Cancer Treat Rep* 1985;69:1065–9.
4. Leblanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc* 1993;88:457–67.
5. Ciampi A, Hogg SA, McKinney S, Thiffault J. RECPAM A computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. methods and program features. *Comput Methods Program Biomed* 1988;26:239–56.
6. Davis RB, Anderson JR. Exponential survival trees. *Stat Med* 1989;8:947–61.
7. Leblanc M, Crowley J. Relative risk trees for censored survival data. *Biometrics* 1992;48:411–25.
8. Segal MR. Regression trees for censored data. *Biometrics* 1988;44:35–47.
9. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residual for survival models. *Biometrika* 1990;77:147–60.
10. Zhang HP. Splitting Criteria in Survival Trees. In *Statistical Modelling Proceedings of the 10th International Workshop on Statistical Modeling*, 1995:305–14.
11. Shimokawa A, Kawasaki Y, Miyaoka E. Comparison of splitting methods on survival tree. *Int J Biostat* 2015;11:175–88.
12. Quantin C, Billard L, Touati M, Andreu N, Cottin Y, Zeller M, et al. Classification and regression trees on aggregate data modeling: an application in acute myocardial infarction. *J Probab Stat* 2011;2011:1–19.
13. Bertrand P, Goupil F. Descriptive statistics for symbolic data. In: Bock HH, and Diday E. editors. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Berlin: Springer-Verlag, 2000:106–24.
14. Billard L, Diday E. *Symbolic data analysis: conceptual statistics and data mining*. Chichester: Wiley, 2006.
15. Bock HH, Diday E. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Berlin: Springer, 2000.
16. Diday E, Noirhomme-Fraiture M. *Symbolic data analysis and the SODAS software*. Chichester: Wiley, 2008.
17. Billard L, Diday E. From the statistics of data to the statistics of knowledge: symbolic data analysis. *J Am Stat Assoc* 2003;98:470–87.
18. Mballo C, Diday E. Decision trees on interval valued variables. *Elect J Symbolic Data Anal* 2005;3:8–18.
19. Shimokawa A, Kawasaki Y, Miyaoka E. Construction of regression trees on interval-valued symbolic variables. *J Jpn Soc Comput Stat* 2014;27:61–79.
20. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457–81.
21. Narita Y, Shibui S; Committee of Brain Tumor Registry of Japan Supported by the Japan Neurosurgical Society. Trends and outcomes in the treatment of Gliomas based on data during 2001–2004 from the brain tumor registry of Japan. *Neurol Medico-Chirurgica* 2015;55:286–95.
22. The Committee of the Brain Tumor Registry of Japan. Report of brain tumor registry of Japan (2001–2004) 13th edition. *Neurol Medico-Chirurgica* 2014;54:1–102.
23. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychol Methods* 2009;14:323–48.
24. Zhou Y, McArdle JM. Rationale and applications of survival tree and survival ensemble methods. *Psychometrika* 2015;80:811–33.