Improving the within-Node Estimation of Survival Trees while Retaining Interpretability

Abstract

In statistical learning for survival data, survival trees are favored for their capacity to detect complex relationships beyond parametric and semiparametric models. Despite this, their prediction accuracy is often suboptimal. In this paper, we propose a new method for improving the within-node estimation and overall survival prediction accuracy, while preserving the interpretability of the survival tree. Simulation studies reveal the proposed method's superior finite sample performance compared to conventional approaches for within-node estimation in survival trees. Furthermore, we apply this method to analyze the North Central Cancer Treatment Group Lung Cancer Data and cardiovascular medical records from the Faisalabad Institute of Cardiology.

Keywords: Survival Analysis, Censored Data; Decision Trees; Interpretable Machine Learning.

1 Introduction

Survival trees have gained popularity in statistical learning literature for censored data due to their ability to detect complex relationships beyond parametric and semiparametric models (e.g., Segal, 1988; LeBlanc and Crowley, 1993; Steingrimsson et al., 2016). However, a recognized limitation of survival trees is their instability, resulting in suboptimal accuracy for survival prediction (Zhu and Kosorok, 2012; Fu and Simonoff, 2017). To address this issue, researchers often turn to tree-based ensembles, such as the random survival forest (Ishwaran et al., 2008), recursively imputed survival trees (Zhu and Kosorok, 2012), and conditional inference survival forest (Nasejje et. al., 2017). While these ensembles enhance prediction accuracy, they lead to "black-box" models, lacking interpretability and insight into the underlying predictive process (Castelvecchi, 2016; Samek and Müller, 2019). In recent years, the issue of interpretability has become central to the development and implementation of statistical learning models (Gilpin et. al. 2018), particularly in application settings where interpretability is crucial, such as defining prognosis and risk subgroups based on disease-free survival for cervical carcinoma patients (Sevin et. al., 1996) or offering practical silvicultural guidance to minimize losses in oak forest mortality (Fan et. al., 2006). In such contexts, tree-based ensembles may be impractical, and it is of interest to investigate ways to improve the prediction accuracy of survival trees while preserving their interpretability.

Current methodological research on improving survival tree prediction accuracy primarily focuses on refining splitting rules. Commonly employed splitting rules include log-rank-type statistics (e.g., Ciampi et. al., 1986; Segal, 1988), likelihood-based measurements (Ciampi et. al., 1987; Davis and Anderson, 1989; LeBlanc and Crowley, 1993), and various residuals (e.g., Therneau, Grambsch, and Fleming, 1990; Keleş and Segal, 2002). The review papers by Bou-Hamad, Larocque, and Ben-Ameur (2011) and Wang and Li (2017) provide comprehensive summaries of the existing splitting rules of survival trees. Recent advancements in survival tree methodology involve optimizing the tree construction procedure through mixed-integer optimization and local search techniques, yielding globally optimized survival trees (Bertsimas et al., 2022). In this paper, we propose a new approach, focusing on improving the within-node estimation to enhance the prediction accuracy. Our strategy adopts a super learning approach (Breiman, 1996; Van der Laan, Polley, and Hubbard, 2007), stacking nested models to improve the within-node estimation and overall survival prediction accuracy, while retaining the interpretability of the survival tree.

The paper is structured as follows: In Section 2, we provide detailed descriptions of the proposed method. In Section 3, we present simulation studies assessing its finite sample performances. In Section 4, we apply the method to analyze two real-world datasets: (1) North Central Cancer Treatment Group (NCCTG) Lung Cancer Data (Loprinzi et. al. 1994) and (2) cardiovascular medical records from the Faisalabad Institute of Cardiology (Chicco and Jurman, 2020). In Section 5, we provide concluding remarks on the method's significance and potential implications in survival tree methodological research.

2 Proposed Method

Our proposed method is presented in this section. We start with a conceptual overview (Section 2.1) and subsequently provide a rigorous description of the algorithm (Section 2.2).

2.1 Conceptual Overview

Our goal is to improve the within-node estimation of a survival tree in estimating the survival distribution (i.e., the survival function or the cumulative hazard function). Using Figure 1, we graphically illustrate the proposed method using an example with two covariates (denoted as Z_1 and Z_2, respectively). Generalization to higher dimensions is straightforward. Imagine in a practical setting, a survival tree algorithm partitions the 2-dimensional covariate space into terminal nodes A, B, C, and D (see left panel, top row of Figure 1). The conventional within-node estimation method involves separately estimating the cumulative hazard function in each terminal node. Specifically, a Nelson-Aalen estimator, based on all observations within a terminal node, is employed to estimate the survival of the corresponding terminal node. Suppose the conventional method yields event rates of 0.9, 0.3, 0.4, and 0.88 for terminal nodes A, B, C, and D, respectively.

Then intuitively, subjects in node A exhibit survival patterns similar to those in node D, and subjects in node B are akin to those in node C. The proposed method aims to construct a sequence of nested models by merging similar nodes and then combining these models to get a final estimate. Specifically, nodes A and D are merged to form the first reduced model, in which the survivals in nodes A and D are estimated jointly (see right panel, top row of Figure 1). Then, nodes B and C are merged, creating the second reduced model (see left panel, second row of Figure 1). Further merging produces the third reduced model, or in this case, just a one-sample estimate (see right panel, second row of Figure 1). Subsequently, a final estimate is obtained by taking a weighted average of all four models using a data-driven approach (i.e., super learning).

By doing so, different terminal nodes with similar survival patterns are allowed to "share information", consequently improving the prediction accuracy. Note that the reduced models are not necessarily in a tree-based structure, since widely separated terminal nodes may be merged during the process. Nevertheless, since all reduced models are nested models of the original survival tree (left panel, top row of Figure 1), their weighted combination will also retain the interpretation of the original tree.

2.2 Algorithm

In the standard right-censored setting, we let T denote the survival time, C denote the censoring time, and $\boldsymbol{Z}$ denote the p dimensional covariate vector. The observed time is $X = \min(T, C)$ and the event indicator is $\Delta = I(T \leq C)$. The training data based on n i.i.d. observations are denoted as $\mathcal{L} = \{(X_i, \Delta_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$. Let $N_i(t) = \Delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$ denote the counting process and at-risk process of subject i, respectively. The objective is to estimate $S(t|\boldsymbol{Z}) = P(T > t|\boldsymbol{Z})$, or equivalently, $\Lambda(t|\boldsymbol{Z}) = -\log S(t|\boldsymbol{Z})$.

The conventional method of within-node estimation involves using $\mathcal{L}$ to fit a survival tree, which partitions the covariate space into M terminal nodes, $R_1, \ldots, R_M$. The within-node

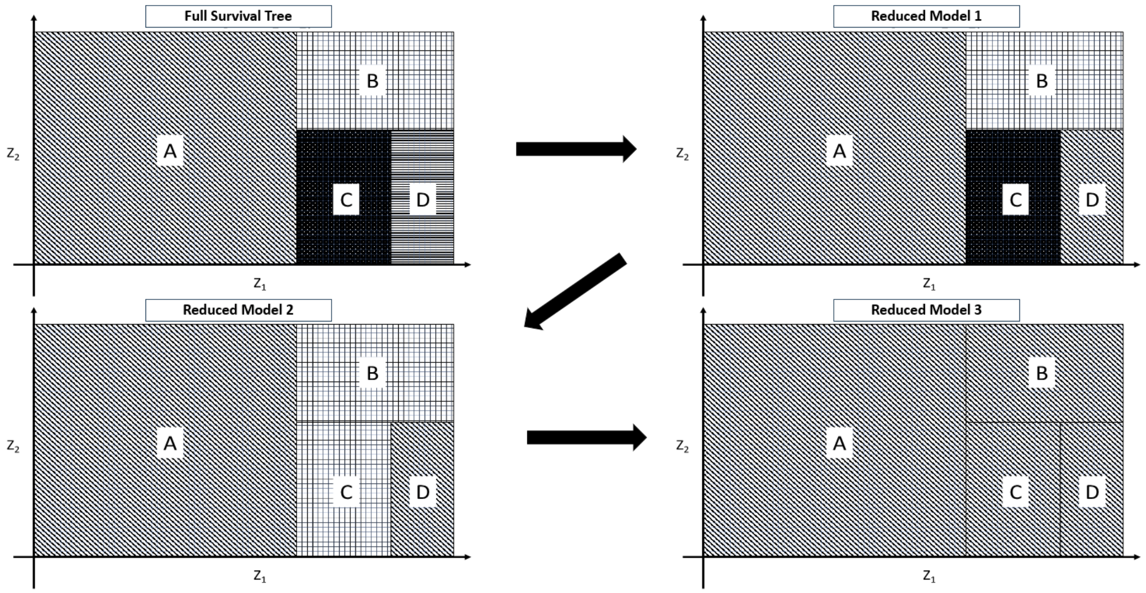


Figure 1: Graphical Illustration of the Proposed Method in 2-Dimensional Covariate Space

estimation of the cumulative hazard function for node m is estimated by the Nelson-Aalen estimator based on all observations in node m:

$$\widehat{\Lambda}^M(t|\boldsymbol{Z}_i) = \sum_{m=1}^{M} \widehat{\Lambda}^m(t|\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \in R_m), \tag{1}$$

where $\widehat{\Lambda}^m(t|\boldsymbol{Z}_i) = \frac{1}{n_m} \sum_{i=1}^{n} \int_0^t \frac{I(\boldsymbol{Z}_i \in R_m) dN_i(u)}{n_m^{-1} \sum_{j=1}^n I(\boldsymbol{Z}_j \in R_m) Y_j(u)}$, and $n_m = \sum_{i=1}^{n} I(\boldsymbol{Z}_i \in R_m)$ denotes the number of subjects in node m. Then given the new covariate information $\boldsymbol{z}_0$, the predicted cumulative hazard function is $\widehat{\Lambda}^M(t|\boldsymbol{z}_0)$ and the predicted survival function is $\widehat{S}^M(t|\boldsymbol{z}_0) = \exp(-\widehat{\Lambda}^M(t|\boldsymbol{z}_0))$.

By contrast, the proposed method of within-node estimation goes as follows:

1. Fit a survival tree using the conventional method of within-node estimation.

2. Randomly split $\mathcal{L}$ into K folds (a recommended choice of K is 10), denoted as $\mathcal{L}_{(1)}, \ldots, \mathcal{L}_{(K)}$. Let $\mathcal{L}^{(k)} = \mathcal{L} - \mathcal{L}_{(k)}$, for $k = 1, \ldots, K$. For $k = 1, \ldots, K$:

    i. Use observations in $\mathcal{L}^{(k)}$ to recalculate the within-nodes estimation of the survival tree (1), denoted as $\widehat{\Lambda}^M_{(k)}(t|\boldsymbol{Z}_i)$:

    $$\widehat{\Lambda}^M_{(k)}(t|\boldsymbol{Z}_i) = \sum_{m=1}^{M} \widehat{\Lambda}^m_{(k)}(t|\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \in R_m),$$

    where $\widehat{\Lambda}^m_{(k)}(t|\boldsymbol{Z}_i) = \frac{1}{n_m^{(k)}} \sum_{i=1}^{n} \int_0^t \frac{I(\boldsymbol{Z}_i \in R_m, i \in \mathcal{L}^{(k)}) dN_i(u)}{n_m^{(k)-1} \sum_{j=1}^n I(\mathbf{Z}_j \in R_m, j \in \mathcal{L}^{(k)}) Y_j(u)}$, and

$n_m^{(k)} = \sum_{i=1}^{n} I(\boldsymbol{Z}_i \in R_m, i \in \mathcal{L}^{(k)})$ denotes the number of subjects in $\mathcal{L}^{(k)}$ and in node m.

ii. Merge two closest terminal nodes (based on a particular splitting rule), say R_p and R_q, where $p, q \in \{1, \ldots, M\}$ and $p \neq q$, to fit the reduced model, $\widehat{\Lambda}_{(k)}^{M-1}(t|\boldsymbol{Z})$, which is nested to $\widehat{\Lambda}_{(k)}^{M}(t|\boldsymbol{Z})$, such that

$$\begin{aligned}\widehat{\Lambda}_{(k)}^{M-1}(t|\boldsymbol{Z}_i) &= \sum_{m \neq p,q} \widehat{\Lambda}_{(k)}^{m}(t|\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \in R_m) + I(\boldsymbol{Z}_i \in R_p \cup R_q) \\ &\times \frac{1}{n_p^{(k)} + n_q^{(k)}} \sum_{i=1}^{n} \int_0^t \frac{I(\boldsymbol{Z}_i \in R_p \cup R_q, i \in \mathcal{L}^{(k)}) dN_i(u)}{(n_p^{(k)} + n_q^{(k)})^{-1} \sum_{j=1}^{n} I(\boldsymbol{Z}_j \in R_p \cup R_q, j \in \mathcal{L}^{(k)}) Y_j(u)}.\end{aligned}$$

iii. Treat R_p and R_q as a single node. Continue merging two closest nodes of $\widehat{\Lambda}_{(k)}^{M-1}(t|\boldsymbol{Z})$ to fit $\widehat{\Lambda}_{(k)}^{M-2}(t|\boldsymbol{Z})$. Then, continue the procedure to get a sequence of nested models, denoted as $\widehat{\Lambda}_{(k)}^{M}(t|\boldsymbol{Z}), \ldots, \widehat{\Lambda}_{(k)}^{1}(t|\boldsymbol{Z})$, where

$$\widehat{\Lambda}_{(k)}^{1}(t|\boldsymbol{Z}_i) = \frac{1}{n^{(k)}} \sum_{i=1}^{n} \int_0^t \frac{I(i \in \mathcal{L}^{(k)}) dN_i(u)}{n^{(k)-1} \sum_{j=1}^{n} I(j \in \mathcal{L}^{(k)}) Y_j(u)},$$

where $n^{(k)} = \sum_{i=1}^{n} I(i \in \mathcal{L}^{(k)})$.

3. for $k = 1, \ldots, K$, calculate the out-of-fold predicted cumulative hazard function using models $\widehat{\Lambda}_{(k)}^{M}(t|\boldsymbol{Z}), \ldots, \widehat{\Lambda}_{(k)}^{1}(t|\boldsymbol{Z})$. Denote the out-of-fold predicted cumulative hazard as $\widetilde{\Lambda}^{M}(t|\boldsymbol{Z}), \ldots, \widetilde{\Lambda}^{1}(t|\boldsymbol{Z})$.

4. Compute the optimal stacking weights $\widehat{w}_1, \ldots, \widehat{w}_M$ by maximizing cross-validated C-index (Harrell et. al., 1982). Specifically,

$$\{\widehat{w}_1, \ldots, \widehat{w}_M\} = \text{argmax}_{w_1, \ldots, w_M} C\left(\mathcal{L}, \sum_{m=1}^{M} w_m \widetilde{\Lambda}^{m}(\tau|\boldsymbol{Z})\right), \tag{2}$$

where τ denotes the end-of-study time, and

$$C(\mathcal{L}, M(\boldsymbol{Z})) = \frac{\sum_{i=1}^{n} \Delta_i \sum_{j=i+1}^{i} I(X_i < X_j) I(M(\boldsymbol{Z}_i) > M(\boldsymbol{Z}_j))}{\sum_{i=1}^{n} \Delta_i \sum_{j=i+1}^{i} I(X_i < X_j)}$$

denote the C-index, where $M(\boldsymbol{Z})$ is some risk score in which a high value of $M(\boldsymbol{Z})$ indicates a greater probability of developing the event of interest. The optimization in (2) is under the constrains such that $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$ for all $m = 1, \ldots, M$.

4. The final estimation of the survival tree is

$$\widehat{\Lambda}(t|\boldsymbol{Z}) = \sum_{m=1}^{M} \widehat{w}_m \widehat{\Lambda}^m(t|\boldsymbol{Z}),$$

   where the structure of $\widehat{\Lambda}^m(t|\boldsymbol{Z})$ is determined by sequentially merging the terminal nodes based on all observations in $\mathcal{L}$, and $\widehat{\Lambda}^m(t|\boldsymbol{Z})$ is fitted based on all observations in $\mathcal{L}$.

Then given the new covariate information $\boldsymbol{z}_0$, the proposed method of within-node estimation yields the predicted cumulative hazard function as $\widehat{\Lambda}(t|\boldsymbol{z}_0)$ and the predicted survival function is $\widehat{S}(t|\boldsymbol{z}_0) = \exp(-\widehat{\Lambda}(t|\boldsymbol{z}_0))$.

3 Simulation Studies

We perform simulation studies to evaluate the finite sample performance of the proposed methods across various scenarios. Using a sample size of 100, we generate the covariate vector $\boldsymbol{Z} = (Z_1, \ldots, Z_{10})^\top$ from a 10-dimensional multivariate normal distribution with a mean vector $\mathbf{0}$ and a variance-covariance matrix Σ, where the (i, j)-th component $\Sigma_{i,j} = 0.6^{|i-j|}$, for all $i, j = 1, \ldots, p$. We consider event rates of 50%, 80%, and 90%. Survival times are generated from a Weibull proportional hazards model with three types of covariate effects: (1) tree-based structure, (2) linear covariate effect, and (3) nonlinear covariate effect. For tree-based structure, the survival time is generated from

$$\Lambda(t|\boldsymbol{Z}) = t^{0.7} \exp(3(Z_1 > 0) - 2(Z_2 < 0) + 3(Z_4 < 0)(Z_6 < 0) - 0.4(Z_7 > 0)(Z_3 > 0)). \quad (3)$$

For linear covariate effect, the survival time is generated from

$$\Lambda(t|\boldsymbol{Z}) = t^{0.7} \exp(3Z_1 - 4Z_2 + 2.5Z_3 + 3Z_4 - Z_6 - 4Z_7 + 1.5Z_8 + 3Z_9 - 2Z_{10}). \quad (4)$$

For nonlinear covariate effect, the survival time is generated from

$$\Lambda(t|\boldsymbol{Z}) = t^{0.7} \exp(0.3Z_1 - 0.2Z_2Z_3 + 0.3Z_4Z_6 - 4Z_7^3 + Z_3Z_5 - 0.25Z_8 - 3Z_1Z_9Z_{10}). \quad (5)$$

Censoring times are generated from a Uniform[0, c_{max}] distribution truncated at τ_{max}, where $\tau_{max} < c_{max}$. c_{max} and τ_{max} are chosen based on the targeted event rates.

When fitting survival trees, we use the log-rank test as the splitting rule and apply the default stopping rules for tree growing. When applying the proposed method for within-node estimation, we also use the log-rank test to assess terminal node similarity, where smaller log-rank test statistics indicate greater similarity between nodes. Prediction accuracy is compared between survival trees with conventional within-node estimation (i.e., solely based on observations within each terminal node) and survival trees with the proposed within-node estimation method. For each scenario, a test dataset is generated using the same mechanism as the training dataset, and the C-index based on the test dataset is used to measure the goodness of prediction.

We perform 500 simulations per scenario and present results in Table 1. Reported are the median, lower quartile (Q1), and upper quartile (Q3) of C-indices for survival trees, comparing conventional and proposed methods for within-node estimation. Additionally, we calculate the percentage improvement in median C-index by the proposed method compared to the conventional one. Results in Table 1 consistently show the proposed method outperforming, with median C-index improvement ranging from 3% to 10% across all scenarios. Notably, for tree-based covariate effect (equation 3) at a 90% event rate, the proposed method enhances median C-index from 0.51 to 0.56, a 9.8% improvement. In linear covariate effect (equation 4) at a 50% event rate, the proposed method increases median C-index from 0.61 to 0.63, a 3.3% rise. In the case of nonlinear covariate effect (equation 5) with a 80% event rate, the proposed method raises median C-index from 0.57 to 0.60, a 5.3% increase. Similarly, Q1 and Q3 of the C-index using the proposed method consistently surpass those of survival trees with conventional within-node estimation.

Table 1: Median (Q1, Q3) of C-Indices of Survival Trees with Conventional and Proposed Methods of within-Node Estimation in Various Simulation Settings (% Improvement measured by the % Increase in Median C-Index by the Proposed Method)

		within-Node Estimation Method		
Covariate Effect	Event Rate	Conventional	Proposed	% Improvement
Tree-Based	90%	0.51 (0.47, 0.56)	0.56 (0.51, 0.60)	9.8%
	80%	0.55 (0.49, 0.60)	0.60 (0.56, 0.64)	9.1%
	50%	0.62 (0.57, 0.66)	0.65 (0.61, 0.69)	4.8%
Linear	90%	0.52 (0.48, 0.55)	0.56 (0.53, 0.59)	7.7%
	80%	0.54 (0.49, 0.58)	0.59 (0.55, 0.62)	9.3%
	50%	0.61 (0.57, 0.64)	0.63 (0.60, 0.66)	3.3%
Nonlinear	90%	0.53 (0.47, 0.59)	0.56 (0.51, 0.61)	5.7%
	80%	0.57 (0.50, 0.64)	0.60 (0.55, 0.65)	5.3%
	50%	0.64 (0.59, 0.69)	0.69 (0.64, 0.72)	7.8%

4 Applications

4.1 North Central Cancer Treatment Group Lung Cancer Data

The North Central Cancer Treatment Group (NCCTG) Lung Cancer Data (Loprinzi et. al. 1994) includes 228 patients, aiming to evaluate the survival probabilities of lung cancer patients after diagnosis. The dataset encompasses demographic information and performance assessments, including sex, age, Eastern Cooperative Oncology Group (ECOG) performance score, Karnofsky performance score, caloric intake at meals, and weight loss in the last six months (in pounds). Among the 228 patients, 63 cases are right-censored.

To compare the proposed method against the conventional within-node estimation approach, we divide the dataset by a 6:4 ratio. The training set includes 60% of the original data, and the test set includes the remaining 40%. Utilizing the training data, we constructed a survival tree using both the conventional within-node estimation and the proposed method. Subsequently, the test set were employed to evaluate their predictive accuracy. Results indicate that the proposed within-node estimation method yields a C-index of 0.581, surpassing the conventional method's C-index of 0.534, representing a 9% improvement.

In Figure 2, North Central Cancer Treatment Group Lung Cancer Data analysis results are presented. The upper segment illustrates the survival tree structure, while the lower segment presents the estimated survival function utilizing the proposed within-node estimation method for each terminal node. The survival tree is interpretable through sequential "if-then" clauses:

(i) If patient caloric intake at meals is ≤ 500, then terminal node 1 is reached, with a median survival time of approximately 250 days and a survival probability of around 0.10 at 600 days.

(ii) If (i) is false and the patient's ECOG performance score is ≤ 0.5, then terminal node 2 is reached, with a median survival time of about 450 days and a survival probability of approximately 0.35 at 600 days.

(iii) If neither (i) nor (ii) is true and the patient's weight loss in the past six months exceeds 14.5 pounds, then terminal node 5 is reached, with a median survival time of around 350 days and a survival probability of about 0.30 at 600 days.

(iv) If none of the conditions (i)-(iii) hold and the patient is male, then terminal node 3 is reached, with a median survival time of approximately 250 days and a survival probability of around 0.15 at 600 days.

(v) If none of the conditions (i)-(iv) hold, the patient reaches terminal node 4, with a median survival time of about 350 days and a survival probability of around 0.30 at 600 days.

4.2 Cardiovascular Medical Records from the Faisalabad Institute of Cardiology

We also analyze the cardiovascular medical records obtained from the Faisalabad Institute of Cardiology (Chicco and Jurman, 2020), including data from 299 patients with heart failure. The cohort consists of 105 females and 194 males, aged between 40 and 95 years. All subjects exhibited left ventricular systolic dysfunction, placing them in New York Heart Association (NYHA) classes III or IV due to prior heart failures. The dataset contains 11 covariates: age, anemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, sex, platelets, serum creatinine, serum sodium, and smoking. Notably, anemia, high blood pressure, diabetes, sex, and smoking are binary variables. Additionally, 96 out of 299 patients experienced the event of interest (death).

To facilitate model evaluation, we partition the dataset into training and test sets using a 6:4 ratio and assess the performance of conventional and proposed within-node estimation methods. The proposed method demonstrated a C-index of 0.671, surpassing the conventional method, which yielded a C-index of 0.625. This represents a 7% improvement in the prediction accuracy.

Figure 3 presents the analysis results. The survival tree can be interpreted as follows:

(i) If the patient's ejection fraction is lower than 32.5%, then terminal 1 is reached, with a survival probability of 0.30 at 250 days.

(ii) If (i) is false and if the patient's creatinine exceeds 1.25 mg/dL, then terminal 4 is reached, with a survival probability of 0.47 at 250 days.

(iii) If (i) and (ii) do not hold and if the creatinine phosphokinase of the patient is less than 577 mcg/L, then terminal 2 is reached, with a survival probability of 0.82 at 250 days.

(iv) If none of the conditions (i)-(iii) hold, then terminal 3 is reached, with a survival probability of 0.77 at 250 days.

5 Discussion

In this paper, we propose a super learning strategy to improve the within-node estimation and overall prediction accuracy of survival trees. Previous research indicates that achieving satisfactory performance in survival trees necessitates a relatively large sample size because the survival estimation for each terminal node relies solely on observations within that node (Fu and Simonoff, 2017). Our method promotes "information sharing" among terminal nodes exhibiting similar survival patterns, thus allowing more efficient utilization of information for a specified sample size. The degrees of information shared among terminal nodes are also determined by a data-driven process. Our method offers distinct advantages. First, the proposed method helps the survival tree improve within-node estimation and overall prediction accuracy while maintaining its original interpretation. Second, the proposed method has versatile applicability, as it can be integrated into survival trees using any splitting rules.

The application of the proposed method that improves the within-node estimation and overall prediction accuracy of survival trees will make a significant impact on biomedical and healthcare research, especially in advancing personalized medicine and optimizing healthcare delivery. Accurate predictions enable tailored treatment plans, efficient resource allocation, and informed decision-making for patients. This enhances defining prognosis, supports early interventions, and aids in risk stratification for public health planning. For example, in oncology, a survival tree with improved prediction accuracy can help identify patients with specific genetic markers or tumor characteristics that influence response to treatment. This information guides oncologists in tailoring therapies, minimizing side effects, and optimizing the chances of successful outcomes. Additionally, accurate survival predictions aid in the selection of eligible participants for clinical trials, accelerating the development of targeted therapies and fostering advancements in cancer research. Overall, heightened prediction accuracy contributes to better patient outcomes, resource efficiency, and the ongoing evolution of precision medicine in the healthcare landscape.

There are also future directions for further research. First, it will be of interest to extend the proposed method to the classification and regression trees (CART; Breiman, 2017) for modeling continuous and categorical outcomes and investigating their finite sample performances. Additionally, extending the method to more complex survival analysis settings, such as competing risks (Andersen et. al., 2012; Kretowska, 2018), recurrent events (Cai and Schaubel, 2003; Sparapani et. al., 2020), left truncation (Guo, 1993; Fu and Simonoff, 2017), and multivariate survival analysis (Gill, 1993; Su and Fan, 2004), would be another prospective direction. Finally, in many epidemiological studies and biomedical research, the observed data are not i.i.d. (e.g., complex survey sampling (Lohr, 2021), outcome dependent sampling (Ding et. al., 2017)). Further research is needed to incorporate the study design into the proposed method in these contexts.

References

[1] Andersen, P. K., Geskus, R. B., de Witte, T., & Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology,***41**(3), 861-870.

[2] Bertsimas, D., Dunn, J., Gibson, E., & Orfanoudaki, A. (2022). Optimal survival trees. *Machine Learning,111*(8), 2951-3023.

[3] Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011). A review of survival trees.

[4] Breiman, L. (1996). Stacked regressions. *Machine learning,***24**, 49-64.

[5] Breiman, L. (2017). *Classification and regression trees.* Routledge.

[6] Cai, J., & Schaubel, D. E. (2003). Analysis of recurrent event data. *Handbook of statistics,*bf 23, 603-623.

[7] Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News,***538**(7623), 20.

[8] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making,***20**(1), 1-16.

[9] Ciampi, A., Thiffault, J., Nakache, J. P., & Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis,***4**(3), 185-204.

[10] Ciampi, A., Chang, C. H., Hogg, S., & McKinney, S. (1987). Recursive partition: A versatile method for exploratory-data analysis in biostatistics. *Biostatistics: Advances in Statistical Sciences Festschrift in Honor of Professor VM Joshi's 70th Birthday Volume V,* 23-50.

[11] Davis, R. B., & Anderson, J. R. (1989). Exponential survival trees. *Statistics in medicine,***8**(8), 947-961.

[12] Ding, J., Lu, T. S., Cai, J., & Zhou, H. (2017). Recent progresses in outcome-dependent sampling with failure time data. *Lifetime data analysis,* **23**, 57-82.

[13] Fan, Z., Kabrick, J. M., & Shifley, S. R. (2006). Classification and regression tree based survival analysis in oak-dominated forests of Missouri's Ozark highlands. *Canadian Journal of Forest Research,* **36**(7), 1740-1748.

[14] Fu, W., & Simonoff, J. S. (2017). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics,* **18**(2), 352-369.

[15] Gill, R. D. (1993). Multivariate survival analysis. *Theory of Probability & Its Applications,* **37**(2), 284-301.

[16] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. *In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

[17] Guo, G. (1993). Event-history analysis for left-truncated data. *Sociological methodology,* 217-243.

[18] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama,* **247**(18), 2543-2546.

[19] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics,* **2**(3), 841-860.

[20] Keleş, S., & Segal, M. R. (2002). Residual-based tree-structured survival analysis. *Statistics in medicine,* **21**(2), 313-326.

[21] Kretowska, M. (2018). Tree-based models for survival data with competing risks. *Computer Methods and Programs in Biomedicine,* **159**, 185-198.

[22] LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association,* **88**(422), 457-467.

[23] Lohr, S. L. (2021). *Sampling: design and analysis.* CRC press.

[24] Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., ... & Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology,* **12**(3), 601-607.

[25] Nasejje, J. B., Mwambi, H., Dheda, K., & Lesosky, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC medical research methodology,* **17**(1), 1-17.

[26] Samek, W., & Müller, K. R. (2019). Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning,* 5-22.

[27] Segal, M. R. (1988). Regression trees for censored data. *Biometrics,* 35-47.

[28] Sevin, B. U., Lu, Y., Bloch, D. A., Nadji, M., Koechli, O. R., & Averette, H. E. (1996). Surgically defined prognostic parameters in patients with early cervical carcinoma: a multivariate survival tree analysis. *Cancer: Interdisciplinary International Journal of the American Cancer Society,***78**(7), 1438-1446.

[29] Sparapani, R. A., Rein, L. E., Tarima, S. S., Jackson, T. A., & Meurer, J. R. (2020). Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics,***21**(1), 69-85.

[30] Steingrimsson, J. A., Diao, L., Molinaro, A. M., & Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in medicine,***35**(20), 3595-3612.

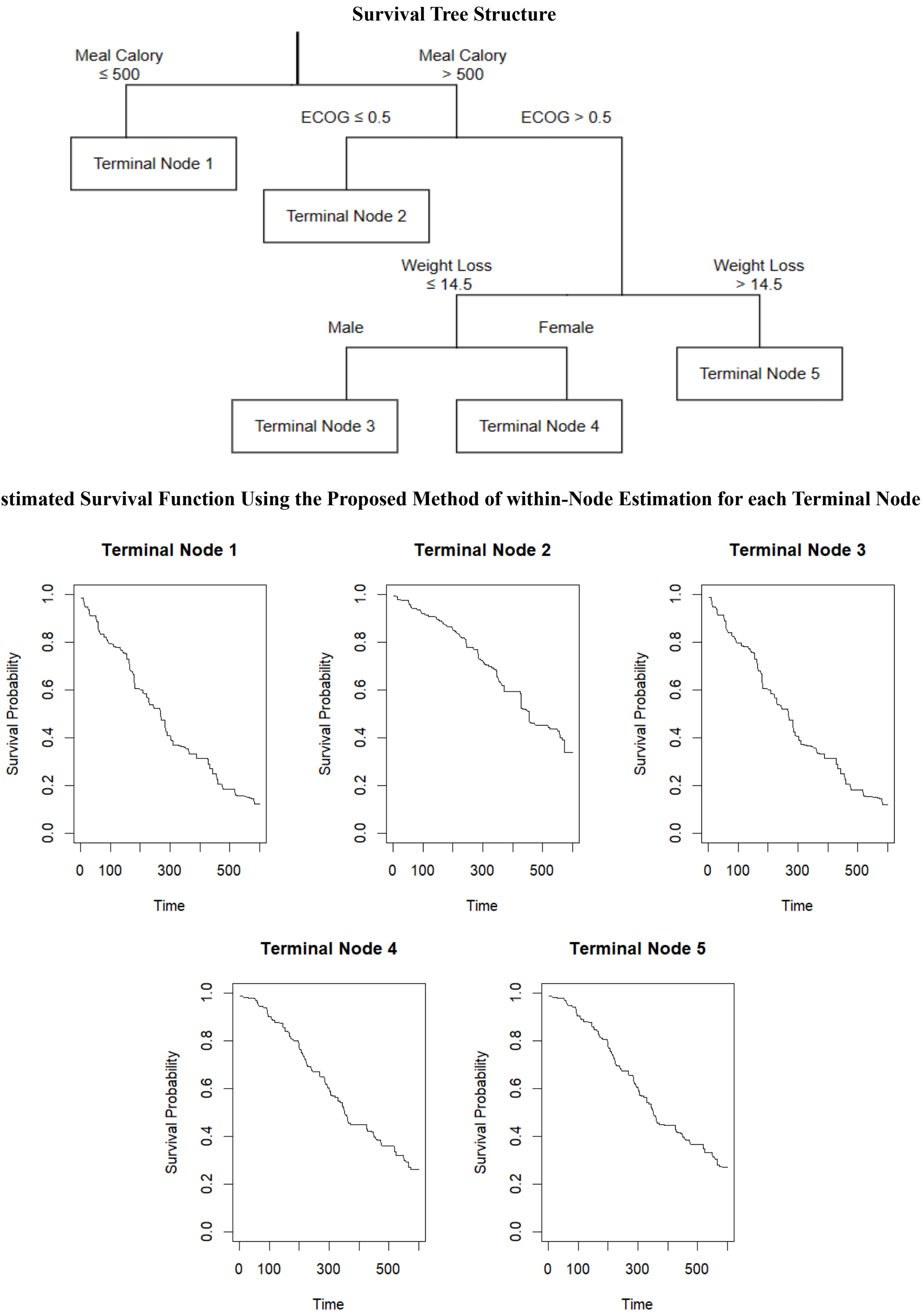
[31] Su, X., & Fan, J. (2004). Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics,***60**(1), 93-99.

[32] Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika,***77**(1), 147-160.

[33] Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology,***6**(1).

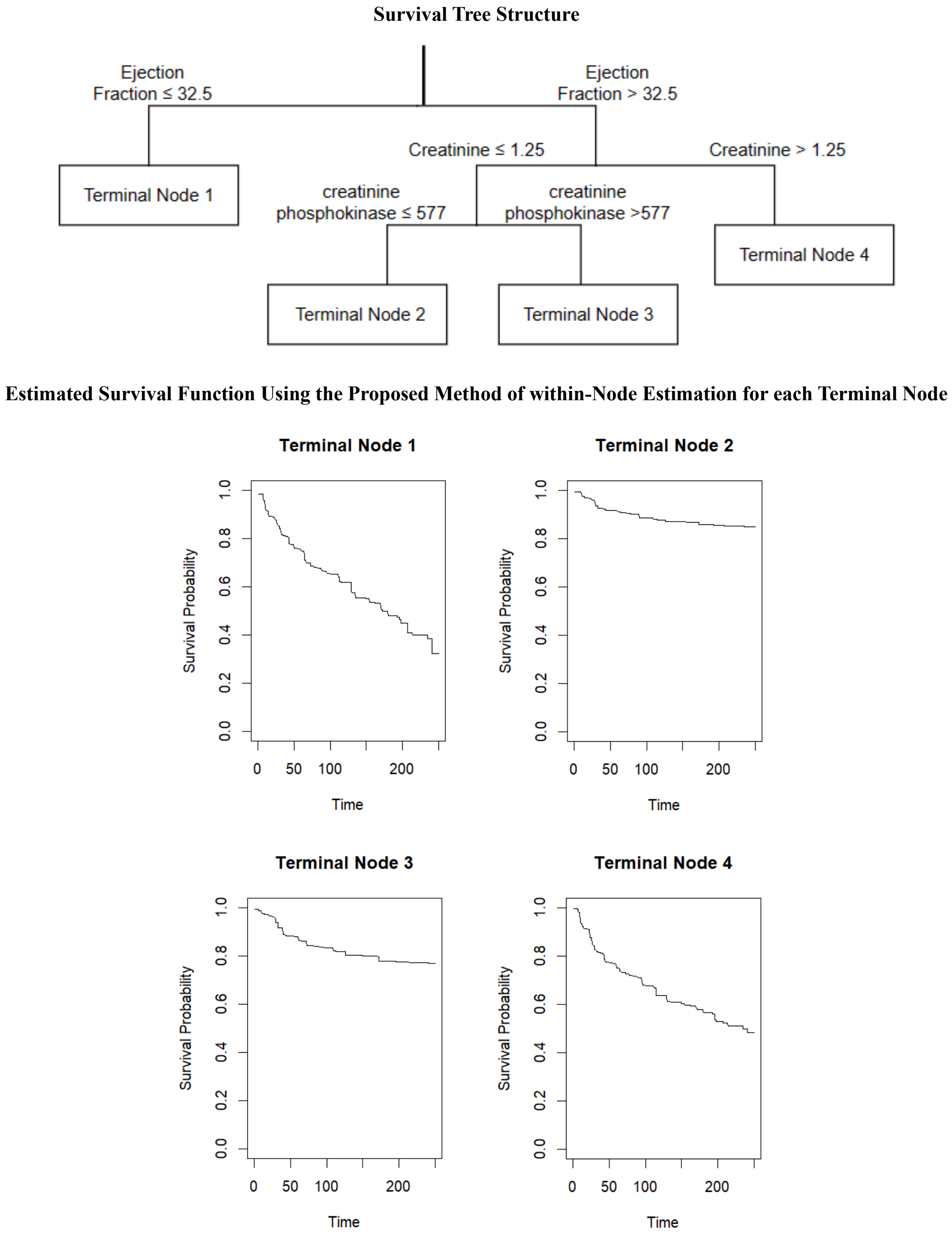
[34] Wang, H., & Li, G. (2017). A selective review on random survival forests for high dimensional data. *Quantitative bio-science,***36**(2), 85.

[35] Zhu, R., & Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association,***107**(497), 331-340.


Survival Tree Structure

Estimated Survival Function Using the Proposed Method of within-Node Estimation for each Terminal Node

Figure 2: Analysis Results for North Central Cancer Treatment Group Lung Cancer Data



Survival Tree Structure

Estimated Survival Function Using the Proposed Method of within-Node Estimation for each Terminal Node

Figure 3: Analysis Results for Cardiovascular Medical Records from the Faisalabad Institute of Cardiology