ROC-Guided Survival Trees and Forests

Yifei Sun and Sy Han Chiou and Mei-Cheng Wang

Department of Biostatistics, Columbia University, New York, New York 10032, U.S.A.

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75080, U.S.A.

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

Summary: Tree-based methods are popular nonparametric tools in studying time-to-event outcomes. In this article, we introduce a novel framework for survival trees and forests, where the trees partition the dynamic survivor population and can handle time-dependent covariates. Using the idea of randomized tests, we develop generalized time-dependent ROC curves to evaluate the performance of survival trees and establish the optimality of the target hazard function with respect to the ROC curve. The tree-growing algorithm is guided by decision-theoretic criteria based on ROC, targeting specifically for prediction accuracy. While existing survival trees with time-dependent covariates have practical limitations due to ambiguous prediction, the proposed method provides a consistent prediction of the failure risk. We further extend the survival trees to random forests, where the ensemble is based on martingale estimating equations, in contrast with many existing survival forest algorithms that average the predicted survival or cumulative hazard functions. Simulations studies demonstrate strong performances of the proposed methods. We apply the methods to a study on AIDS for illustration.

Key words: Concordance index; Risk prediction; ROC curve; Survival analysis; Time-dependent covariate; Tree-based method.
1. Introduction

Tree-based methods are popular alternatives to semiparametric and parametric methods. The basic idea of trees is to partition the covariate space into subsets (nodes) where individuals belonging to the same node are alike regarding the outcome of interest. In the setting of classification and regression trees (CART) (Breiman et al., 1984), this can be achieved by a top-down and greedy approach that aims to minimize a measure of node impurity and the sum of squared deviations from the node mean, respectively; then a cost-complexity pruning algorithm is applied to select a right-sized tree. Moreover, trees are ideal candidates for ensemble methods such as bagging (Breiman, 1996) random forests (Breiman, 2001).

With the increasing focus on personalized risk prediction, tree-based methods for time-to-event data have received much attention. Survival tree has a few appealing features that make it a useful addition to the conventional survival analysis: first, partition arises in many applications where individuals in the same group share similar event risks, and the tree structure retains this natural interpretability. For example, a single survival tree can identify different prognostic groups, so that treatment or prevention strategies can be tailored for patients with different failure risks. Second, survival trees serve as building blocks for ensemble methods and can be easily transformed into powerful risk prediction tools. Gordon and Olshen (1985) first adapt the idea of CART for right-censored survival data and defined the node impurity to be the minimum Wasserstein distance between the Kaplan-Meier curves of the current node and a pure node. Other works that provided a within-node impurity or an estimate of error include Davis and Anderson (1989), LeBlanc and Crowley (1992), Zhang (1995), Jin et al. (2004), Molinaro et al. (2004) and Steingrimsson et al. (2016). As an alternative of maximizing the within-node homogeneity, Segal (1988) and LeBlanc and Crowley (1993), among others, adopted between-node separation approaches that quantify the dissimilarity between two children nodes using the log-rank statistic; other
possible splitting criteria include the Harrell’s C-statistic (Schmid et al., 2016) and the integrated absolute difference between survival functions (Moradian et al., 2017). Readers are referred to Bou-Hamad et al. (2011b) for a comprehensive review of survival trees. In practice, a small perturbation in the training data may result in a large change in a tree-based prediction function. To address the instability issue of a single tree and improve the prediction accuracy, researchers, including Hothorn et al. (2006), Ishwaran et al. (2008), Zhu and Kosorok (2012), Schmid et al. (2016), Steingrimsson et al. (2018), have developed ensemble methods that enlarge the class of models and yield reduced variance.

In many applications, the use of time-dependent covariates offers opportunities for exploring the association between a failure event and risk factors that change over time. Applying standard methods such as the Cox model is technically challenging in the choice of covariate form and can often yield biased estimation (Fisher and Lin, 1999). In contrast, tree-based methods allow the event risk to depend on the covariates in a flexible way, thus can circumvent the challenge in specifying the functional form of time-dependence. However, the aforementioned survival tree methods only deal with baseline covariates and cannot be directly applied. Bacchetti and Segal (1995) incorporated time-dependent covariates by using “pseudo-subject”, where the survival experience of one subject was viewed as survival experiences of multiple pseudo-subjects on non-overlapping intervals, and survival probability using the truncation product-limit estimator (Wang et al., 1986) was reported as the node summary. The idea of pseudo-subject was employed by most of the existing works dealing with time-dependent covariates (Huang et al., 1998; Bou-Hamad et al., 2011a; Wallace, 2014; Fu and Simonoff, 2017). However, the pseudo-subject approach may have practical limitations because one subject could be classified into multiple nodes in a tree, leading to a loss of simple interpretation and possible ambiguous prediction.

In this article, we propose a unified framework for tree-structured analysis with right cen-
sored survival outcomes and possibly time-dependent covariates. Existing survival trees are typically implemented via a greedy algorithm that maximizes the within node homogeneity or between node heterogeneity, which is not directly related to the prediction accuracy. We define time-dependent ROC curves for survival trees using the idea of randomized test and show that the target hazard function yields the highest ROC curve. The optimality of the target hazard function motivates us to use the ROC-related summary measures for (i) evaluating the prediction performance of survival trees and (ii) guiding the tree-growing procedure. We further propose a novel risk prediction forest, where the ensemble is on unbiased martingale estimating equations. Our forest algorithm shows great potential in reducing the bias compared to algorithms that directly average node summaries such as Kaplan-Meier and Nelson-Aalen estimates.

The article is organized as follows. In Section 2, we rigorously define the ROC curves to quantify the prediction performance of survival tree models. In Section 3, we develop ROC-guided splitting and pruning procedures for a right-censored survival outcome with time-dependent covariates. In Section 4, we extend the proposed survival trees to random forests to improve prediction accuracy. In Section 5, simulations studies are conducted to examine the performance of the proposed methods. In Section 6, the methods are applied to an AIDS study for illustration. We conclude the paper with a discussion in Section 7.

2. Evaluating Survival Trees with ROC Curves

In this section, we introduce a survival tree framework that can incorporate time-dependent covariates, and propose the use of ROC curves in evaluating the prediction performance. The conventional definition of incident/dynamic time-dependent ROC curve \cite{Heagerty2005} is appropriate for quantifying the prediction performance of a continuous disease marker, but may not be applicable in evaluating discrete-valued risk scores derived from a tree. For a discrete-valued marker, the ROC curve at time $t$ degenerates to a finite
number of points, hence important summary measures such as area under the curve (AUC) are not well defined. We fill in the gap by generalizing the definition of time-dependent ROC curves to evaluate continuous, discrete and mixed types of markers.

2.1 Survival trees that partition the survivor population

Suppose $T$ is a continuous survival time outcome and $Z(t)$ is a $p$-dimensional vector of possibly time-dependent covariates. Denote by $\lambda(t, z)$ the hazard function of $T$ given $Z(t) = z$, that is, $\lambda(t, z)\, dt = P\{T \in [t, t + dt) \mid Z(t) = z, T \geq t\}$. The hazard function $\lambda(t, z)$ characterizes the instantaneous risk of failure at time $t$ among survivors.

At time $t$, let $Z_t$ denote the covariate space of $Z(t)$ in the survivor population (i.e., the subpopulation satisfying $T \geq t$). We consider a tree that divides the survivor population into $M$ subgroups according to a partition on $Z_t$, denoted by $T = \{\tau_1, \tau_2, \ldots, \tau_M\}$. The elements of the partition, $\tau_1, \tau_2, \ldots, \tau_M$ are disjoint subsets of $Z_t$ satisfying $\tau_1 \cup \tau_2 \cup \cdots \cup \tau_M = Z_t$, and are called terminal nodes of the tree. A subject enters a terminal node $\tau \in T$ at time $t$ if $Z(t) \in \tau$ and $T \geq t$. For ease of discussion, we restrict the attention to a fixed time interval $[0, t_0]$ and assume $Z_t = [0, 1]^p$ for $t \in [0, t_0]$. Thus the partition $T$ can be applied on $Z_t$ for all $t \in [0, t_0]$. Since $Z(t)$ can vary over time, one subject can enter different terminal nodes at different time points. The partition $T$ induces the following model for the hazard function at $t$ given $Z(t)$,

$$\lambda_T(t, Z(t)) = \sum_{\tau \in T} I(Z(t) \in \tau)\lambda(t \mid \tau), \quad t \in [0, t_0],$$

(1)

where subjects in the same terminal node have the same hazard, and $\lambda(t \mid \tau)\, dt = P\{T \in [t, t + dt) \mid Z(t) \in \tau, T \geq t\}$ is the node-specific hazard function. Define the partition function $l_T$ so that $l_T\{z\} = \tau$ if and only if $z \in \tau$ and $\tau \in T$. The partition-based hazard function at $t$ can be written as $\lambda_T(t, z) = \lambda(t \mid l_T\{z\})$. Under our framework, the time-dependent covariates are handled similarly as in the Cox model in the sense that the hazard at time $t$ depends on the covariates at $t$. On one hand, a time-invariant partition on $[0, t_0]$ allows a
sparse model and easy interpretation of the decision rule; on the other hand, as the number of terminal nodes in $\mathcal{T}$ increases, the partition-based hazard function $\lambda_\mathcal{T}(t, z)$ approximates the true hazard function $\lambda(t, z)$. In Section 3.3, we show that the time-invariant rule is asymptotically valid. Compared to the Cox model that assumes multiplicative covariate effects and requires a correct choice of the functional form of the time-dependent covariates (Fisher and Lin [1999]), the partition-based hazard function can be more robust to model mis-specification.

For a fixed partition $\mathcal{T}$, we consider the estimation of $\lambda_\mathcal{T}$ with right-censored survival data. Let $Y = \min(T, C)$ be the observed survival time and $\Delta = I(T \leq C)$ be the failure event indicator. We use $Z^H(t) = \{Z(s), 0 \leq s \leq t\}$ to denote the covariate history up to $t$. The observed training data are $L_n = \{Y_i, \Delta_i, Z^H_i(Y_i); i = 1, \ldots, n\}$, which are assumed to be independent identically distributed replicates of $\{Y, \Delta, Z^H(Y)\}$. For $\tau \in \mathcal{T}$, define $F_{uc}\tau(t) = P\{Y \leq t, \Delta = 1, Z(Y) \in \tau\}$. By assuming independent censoring within each terminal node, that is, $P\{T \in [t, t + dt] \mid Z(t) \in \tau, T \geq t, C \geq t\} = \lambda(t \mid \tau) dt$, we have

$$\lambda(t \mid \tau) = \frac{f_{uc}\tau(t)}{P\{Z(t) \in \tau, Y \geq t\}},$$

where $f_{uc}\tau(t) = dF_{uc}\tau(t)/dt$. Define the observed counting process $N(t) = \Delta I(Y \leq t)$. We estimate $f_{uc}\tau(t)$ by the following kernel type estimator,

$$\hat{f}_{uc}\tau(t) = \int_0^\infty K_h(t-u) d\hat{F}_{uc}\tau(u),$$

where $d\hat{F}_{uc}\tau(u) = \sum_{i=1}^n I(Z_i(Y_i) \in \tau) dN_i(u)/n$, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a second order kernel function with a support on $[-1,1]$ and $h$ is the bandwidth parameter. When $t \in [0, h)$, we either set $\hat{f}_{uc}\tau(t) = \hat{f}_{uc}\tau(h)$ or use the second order boundary kernel in Müller [1991] to avoid biased estimation in the boundary region. For a fixed node $\tau$, $\hat{f}_{uc}\tau(t)$ consistently estimates $f_{uc}\tau(t)$ as $n \to \infty$, $h \to 0$ and $nh \to \infty$. We use $\hat{P}$ to denote the empirical estimate
using the training data. Thus $\lambda(t | \tau)$ is estimated by
\[
\hat{\lambda}(t | \tau) = \int_0^{\infty} K_h(t-u) d\hat{F}_{uc}(u) / \hat{P}\{Z(t) \in \tau, Y \geq t\},
\]
and $\lambda_T(t, z)$ is estimated by $\hat{\lambda}_T(t, z) = \hat{\lambda}(t | l_T\{z\})$.

**Remark 1:** If the domain of the time-dependent covariates among survivors $\{Z(t) | T \geq t\}$ changes over time, we can transform $Z(t)$ onto $[0, 1]^p$ via a one-to-one function $G_t(\cdot) : Z_t \mapsto [0, 1]^p$. Let $Z'(t) = G_t(Z(t))$ and use $h(t, z')$ to denote the hazard function of $T$ given $Z'(t) = z'$. A tree $T'$ can be constructed using the transformed covariates $Z'(t)$, and the partition-based hazard $\hat{\lambda}_{T'}$ estimates $h$. Since $\lambda(t, z) = h(t, G_t(z))$, the estimated hazard is $\hat{\lambda}_{T'}(t, G_t(z))$. For example, for the $j$th covariate $Z_j(t)$, let $\tilde{F}_{jt}(z_j) = \tilde{P}\{Z_j(t) \leq z_j | Y \geq t\}$ be its empirical cumulative distribution function in the at-risk population, then $G_t(z) = \{\tilde{F}_{1t}(z_1), \ldots, \tilde{F}_{pt}(z_p)\}^T$ is a possible choice. After transformation, the partition is time-varying on the original scale of $Z(t)$.

### 2.2 Generalized ROC curves for survival trees

Since different tree algorithms usually generate different partitions, questions arise as to how the accuracy of partition-based hazard functions can be properly measured. We propose the use of time-dependent ROC curves to evaluate $\hat{\lambda}_T(t, z)$. Heuristically, at time $t$, the ROC curve is defined on the survivor population, where a subject is considered a case if $T = t$ and a control if $T > t$. This corresponds to the incident/dynamic ROC curve in Heagerty and Zheng (2005). Let $g(\cdot) : Z_t \mapsto \mathbb{R}$ be a scalar function that summarizes information from $Z(t)$, and we predict $T = t$ or $T > t$ based on $g(Z(t))$, with a larger value being more indicative of $T = t$. Following Heagerty and Zheng (2005), the false positive rate is $\text{FPR}_t(c) = P\{g(Z(t)) > c \mid T > t\}$, the true positive rate is $\text{TPR}_t(c) = P\{g(Z(t)) > c \mid T = t\}$, and the ROC function is $\text{ROC}_t(q) = \text{TPR}_t(\text{FPR}_t^{-1}(q))$. In the literature of binary classification, it has been recognized that, when predicting a binary disease outcome with
multiple disease markers, the risk score (i.e., the probability of disease given markers) yields
the highest ROC curve \cite{GreenDandSwets1966, McIntoshandPepe2002}. For survival
outcomes, the hazard $\lambda(t, Z(t))$ can be viewed as an analog of the risk score. Following
arguments of the Neyman-Pearson Lemma, it can be shown that setting $g(\cdot) = \lambda(t, \cdot)$ yields
the highest ROC$_t$. Thus $\lambda(t, \cdot)$ with a higher ROC$_t$ curve is desired.

With a finite number of terminal nodes, $\lambda(t, z)$ at time $t$ is a discrete-valued scalar
function of $z$, thus the ROC$_t$ function for $\lambda(t, Z(t))$ becomes a finite set of points rather
than a continuous curve. More generally, if $g(Z(t))$ has a point mass at $c$, the function
ROC$_t(q)$ is not defined for $q \in (\text{FPR}_t(c), \text{FPR}_t(c^-))$. In this case, we construct a continuous
curve, denoted by ROC$_t^*$, via linear interpolation. Specifically, for $q \in (\text{FPR}_t(c), \text{FPR}_t(c^-))$, the point $(q, \text{ROC}_t^*(q))$ on the ROC$_t^*$ curve corresponds to the following prediction rule:

if $g(Z(t)) > c$, predict $T = t$; if $g(Z(t)) = c$, predict $T = t$ with probability $\{q - \text{FPR}_t(c)\}/\{\text{FPR}_t(c^-) - \text{FPR}_t(c)\}$; and if $g(Z(t)) < c$, predict $T > t$. In the special case
where $g(Z(t))$ is a continuous variable, ROC$_t^*$ reduces to ROC$_t$. The formal definition of
ROC$_t^*$ is given in the Appendix. We establish the optimality of true hazard function $\lambda(t, \cdot)$
with respect to ROC$_t^*$ in Proposition 1. The proof is given in the Supplementary Material.

Our result suggests that ROC$_t^*$ can be used to evaluate the predictive ability of $\lambda(t, \cdot)$, and
a higher ROC$_t^*$ curve is favorable.

**Proposition 1:** At time $t$, among all scalar functions $g : Z_t \mapsto \mathbb{R}$, $\lambda(t, \cdot)$ is optimal in
the sense that $g(Z(t)) = \lambda(t, Z(t))$ yields the highest ROC$_t^*$ curve.

The area under the ROC$_t^*$ curve is $\text{AUC}_t^* = \int_0^1 \text{ROC}_t^*(q) dq$, which has the interpretation
of a concordance measure \cite{Pepe2003}. For survival time data, we can shown that $\text{AUC}_t^*$ is
equivalent to

\[
\text{CON}_t(g) = P\{g(Z_1(t)) > g(Z_2(t)) \mid T_2 > T_1 = t\} + \frac{1}{2} P\{g(Z_1(t)) = g(Z_2(t)) \mid T_2 > T_1 = t\},
\]

where $\{Z_1(\cdot), T_1\}$ and $\{Z_2(\cdot), T_2\}$ are i.i.d. replicates of $\{Z(\cdot), T\}$. Based on Proposition 1
CON\(_t\) can be used as a summary measure to evaluate the predictive ability of \(g(Z(t))\) at time \(t\). It is worthwhile to point out that other summary measures of ROC curves such as the partial AUC and specific ROC points (Pepe, 2003) can also be employed. Here we focus on the discussion of CON\(_t\), and the results can be extended to other summary measures.

To evaluate \(\hat{\lambda}_T\) over time, a global measure on \([0,t_0]\) is needed. We define a scalar function \(\tilde{g} : \mathbb{R}^+ \times \mathcal{Z}_t \mapsto \mathbb{R}\) that combines \(Z(t)\) in a possibly time-dependent way. For survivors at \(t\), we use \(\tilde{g}(t,Z(t))\) to characterize the risk of \(T = t\). To derive a global measure, we integrate \(\text{CON}_t(\tilde{g}(t,\cdot))\) over \(t\) with a weight function \(\omega(t)\) and define

\[
\text{CON}(\tilde{g}) = \int_0^{t_0} \omega(t)\text{CON}_t(\tilde{g}(t,\cdot)) \, dt.
\]

Following Proposition 1, the true hazard function \(\lambda\) maximizes CON. Motivated by this fact, we propose to use CON as a guidance to build survival trees, and the goal is to construct a partition \(\mathcal{T}\) so that \(\text{CON}(\lambda_\mathcal{T})\) is as large as possible. A detailed discussion is given in Section 3.

REMARK 2: In practice, investigators can use their own weight functions \(\omega(t)\) to reflect costs of misclassification on different time points. A simple example is to set \(\omega(\cdot) = 1\). Moreover, let \(f,S\) be the marginal density and survival function of \(T\). By setting \(\omega(t) = f(t)S(t)/P(T_2 > T_1, T_1 < t_0)\), we have

\[
\text{CON}(\tilde{g}) = P\{\tilde{g}(T_1, Z_1(T_1)) > \tilde{g}(T_1, Z_2(T_1)) \mid T_2 > T_1, T_1 < t_0\} + \\
\frac{1}{2} P\{\tilde{g}(T_1, Z_1(T_1)) = \tilde{g}(T_1, Z_2(T_1)) \mid T_2 > T_1, T_1 < t_0\},
\]

which measures the probability that the subject who fails earlier has a higher risk at the failure time. In the special case where \(\tilde{g}(t,Z(t))\) is a continuous random variable that does not depend on \(t\), \(\text{CON}(\tilde{g})\) in (3) reduces to the global summary in Heagerty and Zheng (2005).

REMARK 3: The Harrell’s C-statistic (Harrell et al., 1982) has been commonly used to
quantify the capacity of a risk score at baseline in discriminating among subjects with different event times. When the event time is subject to right censoring, however, the population parameters corresponding to the Harrell’s C-statistics depend on the study-specific censoring distribution. Uno et al. (2011) studied a modified C-statistics that is consistent for a population concordance measure free of censoring under a working Cox model \( \lambda(t \mid X) = \lambda_0(t) \exp(X \beta) \), where the linear combination \( X \beta \) maximizes the limiting value of the Uno’s C-statistic. However, without the Cox model assumption, it is not clear how to combine \( X \) so that the limiting C-statistic is maximized. The proposed CON is defined in a different way so that CON is maximized when \( \tilde{g}(t, z) = \lambda(t, z) \), thus is appropriate for guiding the tree building procedure.

3. ROC-Guided Survival Trees

In this section, we develop tree-growing algorithms based on ROC\(^*_t\) and CON. We note that although the assumption

\[
P\{t \leq T < t + dt \mid Z(t) = z, T \geq t, C \geq t\} = \lambda(t, z) dt
\]  

is adopted for establishing the large-sample properties of a grown tree where the diameters of the terminal nodes tend to zero, stronger assumptions are often needed to understand the splitting criteria, especially in the early steps of the algorithm. For example, when selecting the optimal split at the root node, the log-rank splitting rule (Segal, 1988; LeBlanc and Crowley, 1993) implicitly assumes that \( C \) is independent of \( T \) within the children nodes, which is not guaranteed by (4). For ease of discussion, we assume independent censoring \( C \perp \{T, Z(\cdot)\} \) in Section 3.1 and 3.2, and assume (4) in Section 3.3.

3.1 Splitting based on CON

We first consider the use of \( \text{CON} (\lambda_T) \) as the splitting criterion. At each step, we choose to split the node that leads to the largest increase in CON of the tree. Under independent
censoring, the estimation of \( \text{CON}_t(\lambda_T(t, \cdot)) \) is developed based on the following expression,

\[
\text{CON}_t(\lambda_T(t, \cdot)) = \frac{\sum_{\lambda(t|\tau) > \lambda(t|\tau'), \tau, \tau' \in \mathcal{T}} f_{\tau}^{uc}(t) P(Y \geq t, Z(t) \in \tau')}{\sum_{\tau, \tau' \in \mathcal{T}} f_{\tau}^{uc}(t) P(Y \geq t, Z(t) \in \tau')} + \frac{\sum_{\tau \in \mathcal{T}} f_{\tau}^{uc}(t) P(Y \geq t, Z(t) \in \tau)}{2 \sum_{\tau, \tau' \in \mathcal{T}} f_{\tau}^{uc}(t) P(Y \geq t, Z(t) \in \tau')}.
\]

Therefore, replacing the unknown quantities in \( \text{CON}_t(\lambda_T(t, \cdot)) \) with the corresponding estimators in Section 2.1 yields a consistent estimator for the concordance measure, denoted by \( \hat{\text{CON}}_t(\hat{\lambda}_T(t, \cdot)) \). To estimate \( \text{CON} \), we use \( \hat{\text{CON}}(\hat{\lambda}_T) = \int_{0}^{t_0} \hat{\text{CON}}(\hat{\lambda}_T(t, \cdot)) \hat{\omega}(t) dt \), where \( \hat{\omega}(t) \) is a weight function that possibly depends on the data.

Consider the partition \( \mathcal{T} = \{ \tau_1, \tau_2, \ldots, \tau_M \} \) and a split \( s \) on the node \( \tau_j, \tau_j \in \mathcal{T} \). The partition after splitting is denoted by \( \mathcal{T}_s = \{ \tau_m, \tau_j^L, \tau_j^R; m \neq j, m = 1, \ldots, M \} \). Proposition 2 shows that \( \text{CON} \) is a proper splitting criterion and can detect the difference in hazards of two children nodes. The proof is given in the Supplementary Material.

**Proposition 2:** Splitting does not decrease the concordance. Specifically, \( \text{CON}_t(\lambda_{T_s}(t, \cdot)) \geq \text{CON}_t(\lambda_T(t, \cdot)) \), and the equality holds if and only if \( \lambda(t \mid \tau_j) = \lambda(t \mid \tau_j^L) = \lambda(t \mid \tau_j^R) \). Moreover, \( \hat{\text{CON}}_t(\hat{\lambda}_{T_s}(t, \cdot)) \geq \hat{\text{CON}}_t(\hat{\lambda}_T(t, \cdot)) \), and the equality holds if and only if \( \hat{\lambda}(t \mid \tau_j) = \hat{\lambda}(t \mid \tau_j^L) = \hat{\lambda}(t \mid \tau_j^R) \).

Based on Proposition 2, we have \( \hat{\text{CON}}(\hat{\lambda}_{T_s}) \geq \hat{\text{CON}}(\hat{\lambda}_T) \). When \( \hat{\gamma}(\cdot) > 0 \), the equality holds if and only if \( \hat{\lambda}(\cdot \mid \tau_j^L) = \hat{\lambda}(\cdot \mid \tau_j^R) \) almost everywhere on \( [0, t_0] \). The optimal split is thus defined as \( s^{opt} = \arg\max_s \hat{\text{CON}}(\hat{\lambda}_{T_s}) \). Note that the validity of Proposition 2 does not depend on the censoring distribution. In practice, \( \hat{\text{CON}}(\hat{\lambda}_T) \) may not correctly estimate \( \text{CON}(\lambda_T) \) if the independent censoring assumption is violated, but the true concordance usually increases after splitting.

One may also consider another splitting criterion where the optimal split on a node is chosen to maximize the increase of \( \text{CON} \) within the node. For a node \( \tau \), define \( S(t, \tau) = P\{Z(t) \in \tau, T \geq t\} \) and \( f(t, \tau) = \lim_{\delta_t \to 0^+} P\{T \in [t, t + \delta_t), Z(t) \in \tau\} / \delta_t \). For \( \tau \)'s children
nodes $\tau_L$ and $\tau_R$, it can be shown that $\text{CON}_t$ within $\tau$ is 0.5 before splitting and is $0.5 + |f(t, \tau_L)S(t, \tau_R) - f(t, \tau_R)S(t, \tau_L)|/f(t)S(t, \tau)$ after splitting. Hence the increase in $\text{CON}$ within $\tau$ after splitting is

$$\Delta \text{CON}_\tau = \int_0^t \frac{|f(t, \tau_L)S(t, \tau_R) - f(t, \tau_R)S(t, \tau_L)|}{f(t)S(t, \tau)} \omega(t) \, dt.$$ 

Moreover, $\Delta \text{CON}_\tau$ can be estimated by

$$\text{\hat{C}ON}_\tau = \int_0^t \frac{|\hat{f}_{\tau_L}(t)\hat{P}(Y \geq t, Z(t) \in \tau_R) - \hat{f}_{\tau_R}(t)\hat{P}(Y \geq t, Z(t) \in \tau_L)|}{\hat{f}_\tau(t)\hat{P}(Y \geq t, Z(t) \in \tau)} \hat{\omega}(t) \, dt.$$ 

The rule based on $\text{\hat{C}ON}_\tau$ can be viewed as maximizing a weighted average of $|\lambda(t \mid \tau_L) - \lambda(t \mid \tau_R)|S(t, \tau_L)S(t, \tau_R)$ over $t$. Compared with the log-rank splitting rule that maximizes a weighted average of $\lambda(t \mid \tau_L) - \lambda(t \mid \tau_R)$ over $t$, the $\Delta \text{CON}_\tau$-based rule encourages more balanced children nodes and can better detect the difference especially when hazards in the children nodes cross each other.

3.2 Pruning

Although splitting increases the concordance, a large tree often overfit the data. Following Breiman et al. (1984), we continue splitting until a tree is fully grown and prune it from the bottom up to find the most predictive subtree. As a guidance for pruning, we define a concordance-complexity measure,

$$\text{CON}_\alpha(\mathcal{T}) = \text{\hat{C}ON}(\hat{\lambda}_\tau) - \alpha |\mathcal{T}|,$$

where $|\mathcal{T}|$ is the number of terminal nodes in $\mathcal{T}$ and $\alpha$ is a complexity parameter. The second term is a penalty that controls the size of the tree. For each $\alpha$, the optimal subtree is defined as the tree that has the largest value of $\text{CON}_\alpha$ among all the possible subtrees. If more than one subtrees have the same largest value of $\text{CON}_\alpha$, we define the one with the smallest size to be the optimal subtree. In practice, $\alpha$ and the corresponding optimal subtree can be determined by cross-validation. Details of the splitting and pruning algorithm are given in the Supplementary Material.
3.3 Prediction

We use $T_n$ to denote the partition constructed based on the training sample $L_n$. Define

$$\hat{\lambda}_{T_n}(t, z) = \hat{\lambda}(t \mid l_{T_n}\{z\}).$$

Given a new observation $Z_0(t)$, we predict the hazard to be $\hat{\lambda}_{T_n}(t, Z_0(t))$. Theorem 1 summarizes the large sample property of the estimated partition-based hazard function $\hat{\lambda}_{T_n}$. The proof is given in the Supplementary Material.

**Theorem 1:** Under conditions (A1)-(A5) in the Supplementary Material, for $t \in [0, t_0]$ and any $\epsilon > 0$, as $n \to \infty$, we have

$$P\left\{ \left| \hat{\lambda}_{T_n}(t, Z_0(t)) - \lambda(t, Z_0(t)) \right| > \epsilon \mid L_n \right\} \to 0 \text{ with probability 1.}$$

Similar as most of the existing works on consistency of tree-based estimators, the convergence result is independent of the splitting and pruning algorithm. Theorem 1 is developed to provide justification on the time-invariant partition under the commonly adopted assumption that diameters of terminal nodes go to zero as $n$ increases (Breiman et al., 1984; LeBlanc and Crowley, 1993). Large-sample results incorporating the algorithm to grow the trees will be investigated in our future work.

Although the above discussion focuses on the hazard function, we can also predict survival probability when $Z(t)$ is a vector of external time-dependent covariates (Kalbfleisch and Prentice, 2011). Assume the hazard at $t$ depend on $Z^H(t)$ only through $Z(t)$. The prediction of survival probability is based on the equation $P(T \geq t \mid Z^H(t)) = \exp\{- \int_0^t \lambda(u, Z(u)) \, du\}$.

We predict the survival probability at $t$ for a subject with covariate path $Z_0(t)^H = \{Z_0(u), 0 \leq u \leq t\}$ to be

$$\hat{P}(T \geq t \mid Z_0^H(t)) = \exp \left[ - \int_0^t \sum_{i=1}^{n} I(Z_i(u) \in l_{T}\{Z_0(u)\}) \, dN_i(u) \sum_{i=1}^{n} I(Z_i(u) \in l_{T}\{Z_0(u)\}), Y_i \geq u \right]. \quad (5)$$

4. ROC-Guided Random Survival Forests

The proposed survival trees can be further transformed into powerful risk prediction tools by applying ensemble methods such as bagging (Breiman, 1996) and random forests (Breiman, 1996).
The essential idea in bagging is to average many noisy but approximately unbiased tree models to reduce the variance. Random forests improve the variance reduction of bagging by reducing the correlation between the trees via random selection of predictors in the tree-growing process. In the original random forests for regression and classification (Breiman 2001), the prediction for a new data point is the averaged prediction of all trees in the forest, and trees are grown sufficiently deep to achieve low bias. However, for right-censored survival data, averaging estimated survival or cumulative hazard functions from deeply grown trees may result in large bias, because the node-specific estimate is biased when a node contains a small number of observations. For this, we propose to average the unbiased martingale estimating equations rather than averaging node summaries from the Kaplan-Meier or Nelson-Aalen estimates (Ishwaran et al. 2008, Zhu and Kosorok 2012, Steingrimsson et al. 2016, Schmid et al. 2016). Moreover, in light of Meinshausen (2006) and Athey et al. (2018), we treat forests as a type of adaptive nearest neighbor estimator and construct forest-based local estimation for the survival or hazard functions.

Let \( T = \{ T_b \}_{b=1}^B \) be a collection of \( B \) survival trees obtained from resampling the original training data. Each tree is constructed via a tree-growing process where at each split, \( m (m < p) \) predictors are selected at random as candidates for splitting. For the \( b \)th partition \( T_b \) and a node \( \tau \in T_b \), one can solve the following unbiased estimating equation for the node-specific hazard at \( t \),

\[
\sum_{i=1}^{n} I (Z_i(t) \in \tau) \{dN_i(t) - I(Y_i \geq t)\lambda(t \mid \tau)\, dt\} = 0.
\]

Let \( l_b \{ z \} \) be the partition function for \( T_b \) so that \( l_b \{ z \} = \tau \) if and only if \( z \in \tau \) and \( \tau \in T_b \), then the \( b \)th tree induces the following estimating equation for \( \lambda(t, z) \),

\[
\sum_{i=1}^{n} I (Z_i(t) \in l_b \{ z \}) \{dN_i(t) - I(Y_i \geq t)\lambda(t, z)\, dt\} = 0. \tag{6}
\]

It is easy to see that solving Equation (6) yields the prediction of hazard at \( t \) given \( Z(t) = z \) based on one single tree \( T_b \). Note that Equation (6) can be rewritten as the following weighted
estimating equation,
\[ \sum_{i=1}^{n} w_{bi}(t, z) \{dN_i(t) - I(Y_i \geq t)\lambda(t, z) dt \} = 0. \] (7)
where \( w_{bi}(t, z) = I(Z_i(t) \in l_b \{z\}, Y_i \geq t)/\sum_{j=1}^{n} I(Z_j(t) \in l_b \{z\}, Y_j \geq t) \) and \( \sum_{i=1}^{n} w_{bi}(t, z) = 1 \). Specifically, the weight \( w_{bi}(t, z) \) is a positive constant if the \( i \)th subject is at-risk and \( Z_i(t) \) falls in the node \( l_b \{z\} \); otherwise, the weight \( w_{bi}(t, z) \) is zero.

To get forest-based prediction, we take average of the estimating functions in (7) from all the \( B \) trees and obtain the following local martingale estimating equation for \( \lambda(t, z) \),
\[ \sum_{i=1}^{n} w_i(t, z) \{dN_i(t) - I(Y_i \geq t)\lambda(t, z) dt \} = 0, \]
where the weight function is \( w_i(t, z) = \sum_{b=1}^{B} w_{bi}(t, z)/B \). The weight \( w_i(t, z) \) captures the frequency with which the \( i \)th observation \( Z_i(t) \) falls into the same node as \( z \), and \( \sum_{i=1}^{n} w_i(t, z) = 1 \). Therefore, the forest based estimator for the hazard function is
\[ \hat{\lambda}_T(t, z) = \int_{0}^{\infty} K_h(t-u) \frac{\sum_{i=1}^{n} w_i(u, z) dN_i(u)}{\sum_{i=1}^{n} w_i(u, z) I(Y_i \geq u)}. \]
When \( Z(t) \) are external time-dependent covariates, the survival probability at \( t \) given \( Z_0^H(t) = \{Z_0(u), 0 \leq u \leq t \} \) is predicted to be
\[ \exp \left\{ - \int_{0}^{t} \frac{\sum_{i=1}^{n} w_i(u, Z_0(u)) dN_i(u)}{\sum_{i=1}^{n} w_i(u, Z_0(u)) I(Y_i \geq u)} \right\}. \] (8)
To achieve good prediction performance, our algorithm relies on subsampling (without replacement) and sample-splitting techniques \cite{Athey2018}. Specifically, we divide the subsample from the original training data into two halves. The first half is used for tree-growing and the second half is used to calculate the weight. The algorithm for forest-based estimation is given in the Appendix.

5. Simulation Studies
We performed simulations to investigate the performance of the proposed methods. To cover some of the common survival models, we considered the following scenarios where the survival distribution only depends on baseline covariates \( Z \):
(I) Proportional hazards model: 
\[ \lambda(t, Z) = \lambda_0(t) \exp(\sum_{j=1}^{p} \beta_j Z_j), \]
where \( p = 10, \) 
\[ \beta = (-0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5)^\top \] and \( Z \) follows a multivariate normal with mean 0 and covariate matrix with elements \( V_{ij} = 0.75^{i-j}, i, j = 1, \ldots, 10. \)

(II) Proportional hazards model with noise variables: 
\[ \lambda(t, Z) = \lambda_0(t) \exp(\sum_{j=1}^{p} \beta_j Z_j), \]
where \( p = 10, \) 
\[ \beta = (2, 2, 0, 0, 0, 0, 0, 0, 0, 0)^\top. \]

(III) Proportional hazards model with nonlinear covariate effects:
\[ \lambda(t, Z) = \lambda_0(t) \exp\{2 \sin(2\pi Z_1) + 2|Z_2 - 0.5|\}. \]

(IV) Accelerated failure time model: 
\[ \log T = -2 + 2Z_1 + 2Z_2 + \epsilon, \] where \( \epsilon \sim N(0, 0.5^2). \)

(V) Generalized gamma family: 
\[ T = \exp(\sigma w), \ w = \log(Q^2 g)/Q, \ g \sim \text{Gamma}(Q^{-2}, 1), \] and 
\[ \sigma = 2Z_1, \ Q = 2Z_2. \]

Except for Scenario (I), the covariate \( Z \) were generated from uniform distributions over \([0, 1]^p\). In Scenarios (I), (II), and (III), we set the baseline hazard to be \( \lambda_0(t) = 2t, \) which is the hazard function of a Weibull distribution. Additionally, we generated censoring times, \( C, \) from a uniform distribution over \((0, t_c)\), where \( t_c \) was tuned to yield censoring percentages of 25\% and 50\%. As described in Remark 1, we treat the baseline covariates as time-dependent covariates through \( Z(t) \overset{\text{def}}{=} \hat{F}_t(Z), \) where \( \hat{F}_t(\cdot) \) is the empirical cumulative distribution function of the at-risk subjects. Given a new baseline observation \( Z_0, \) we set \( Z_0(t) = \hat{F}_t(Z_0) \) and predict the survival probability as in (5). We compare the proposed methods with the relative risk tree in [LeBlanc and Crowley (1992)] and the conditional inference survival tree in [Hothorn et al. (2006)], which are implemented in R ([R Core Team (2018)] functions rpart and ctree in packages rpart and party, respectively. We used ten-fold cross-validation in choosing the tuning parameter \( \alpha \) in the concordance-complexity measure as well as when selecting the right-sized tree in rpart. To evaluate the performance of different methods, we use the integrated absolute error:
\[ \frac{1}{n_0 t_0} \sum_{i=1}^{n_0} \int_0^{t_0} \left| \hat{P}(T \geq t \mid Z_i^0) - P(T \geq t \mid Z_i^0) \right| dt, \]
where \( \{Z_i^0, i = 1, \ldots, n_0\} \) are generated independently from the simulated data. We set \( n_0 = 1000 \) and \( t_0 \) to be approximately the 95\% quantile of \( Y \). When predicting survival probability, the proposed algorithm is not sensitive to kernel function \( K \) and the bandwidth parameter \( h \), so we use the Epanechnikov kernel function \( K(x) = 0.75(1 - x^2)I(|x| \leq 1) \) and set \( h = t_0/20 \) in all scenarios.

With \( n \in \{100, 200\} \) and 500 replications, the top panel of Table I reports the mean of integrated absolute error of the proposed method, rpart and ctree. For all scenarios considered, the integrated absolute errors decrease with sample size but increase with censoring percentage. On the other hand, splitting by CON and \( \Delta \text{CON}_r \) yield similar integrated absolute errors. The proposed methods are as efficient or more efficient than the competitors.

We next consider scenarios where the survival time depends on time-dependent covariates. In what follows, we considered a time-dependent covariate, \( Z_1(t) \), and a time-independent covariate \( Z_2 \), where the latter was generated from a uniform distribution over \((0, 2)\).

(VI) **Dichotomous time dependent covariate with at most one change in value:** Survival times were generated from
\[
\lambda(t, Z(t)) = \lambda_0(t) \exp \{2Z_1(t) + 2Z_2\},
\]
where \( \lambda_0(t) = 2t \), \( Z_1(t) = \theta I(t \geq U_0) + (1 - \theta)I(t < U_0) \), \( \theta \sim \text{Bernoulli}(0.5) \), and \( U_0 \) follows an exponential distribution with rate 5.

(VII) **Dichotomous time dependent covariate with multiple changes:** Survival times were generated from
\[
\lambda(t, Z(t)) = \exp \{2Z_1(t) + 2Z_2\},
\]
where \( Z_1(t) = \theta \{I(U_1 \leq t < U_2) + I(U_3 \leq t)\} + (1 - \theta)\{I(t < U_1) + I(U_2 \leq t < U_3)\} \), \( \theta \sim \text{Bernoulli}(0.5) \), and \( U_1 \leq U_2 \leq U_3 \) are the first three terms of a stationary Poisson process with rate 10.

(VIII) **Continuous time dependent covariate:** Survival times were generated from
\[
\lambda(t, Z(t)) = 0.1 \exp \{Z_1(t) + Z_2\},
\]
where \( Z_1(t) = kt + b \), \( k \) and \( b \) follow independent uniform distributions over \((1, 2)\).

In these scenarios, we continue to consider the transformation in Remark [1] to transform
In the presence of time-dependent covariate, we define the integrated absolute error as

\[
\frac{1}{n_0 t_0} \sum_{i=1}^{n_0} \int_0^{t_0} \left| \hat{P}\{T \geq t \mid Z_i^0(t)^H\} - P\{T \geq t \mid Z_i^0(t)^H\} \right| dt,
\]

where \( Z_i^0(t)^H = \{Z_i^0(u), 0 \leq u \leq t\} \) denotes the covariate history up to \( t \), and \( Z_i^0(t) \) are generated from the distribution of \( Z(t) \) independently of the training sample. To our knowledge, there is no available software for predicting the survival probability at time \( t \) based on the covariate history up to \( t \). Although there have been existing works on survival tree with time-dependent covariates, it is not clear how existing methods can be applied for prediction. We compare our methods with \texttt{rpart} and \texttt{ctree} that can only handle baseline covariates, \{\( Z_1(0), Z_2 \)\}. As expected, the lower panel of Table 1 shows that the proposed methods outperform \texttt{rpart} and \texttt{ctree} under Scenario (VI) – (VIII). In particular, our methods show a substantial improvement over \texttt{rpart} and \texttt{ctree} in the presence of a continuous time-dependent covariate (i.e., Scenario (VIII)). This indicates incorporating time-dependent covariates improves the prediction accuracy.

[Table 1 about here.]

We continue to use Scenarios (I) – (VIII) to investigate the performance of the proposed forest-based methods. To grow the trees in the forest, we considered subsampling and set the size of subsample to be 80 when \( n = 100 \) and 100 when \( n = 200 \). We randomly select \( \lceil \sqrt{p} \rceil \) variables at each splitting. For each tree, the minimum number of failure that must exist in a node for a split to be attempted is 3 and the minimum number of failures in any terminal node is 1. For each data, we set \( B = 500 \) and compute the integrated absolute error as before, but with \( \hat{P}(\cdot) \) replaced with the survival function in (8). We compare the proposed forest methods with random survival forests in Ishwaran et al. (2008) and Schmid et al. (2016), which are implemented in \texttt{R} functions \texttt{rfsrc} (Ishwaran and Kogalur, 2018) and \texttt{ranger} (Wright and Ziegler, 2017) in packages \texttt{randomForestSRC} and \texttt{ranger}, respectively. We fit
rfsrc and ranger using the default parameter settings and using smaller trees in the forest with larger minimum node sizes (i.e., with nodesize = 10 in rfsrc and min.node.size = 10 in ranger). Since rfsrc and ranger cannot handle time-dependent covariates, we fit these with the baseline covariates, \{Z_1(0), Z_2\}. Table 2 shows the averaged integrated absolute error based on 500 replications. The proposed forest methods perform better or similar to their survival tree counterparts and outperform ranger and rfsrc in almost all of the settings. As expected, our methods have a substantial advantage over the rfsrc and ranger under Scenarios (VI) – (VIII); echoing the importance to incorporate time-dependent covariates in prediction. Interestingly, for ranger and rfsrc, the default setting leads to larger error compared to single trees, and their performances are substantially improved after increasing the size of terminal nodes. We conjecture this could be due to the fact that within-node estimates are less biased with larger node sizes. In summary, our proposed methods are competitive with the existing methods when all the covariates are time-independent, and even show superior performance in the presence of time-dependent covariates.

[Table 2 about here.]

6. Application

We illustrate the proposed methods through an application to a clinical trial conducted by Terry Beirn Community Programs for Clinical Research on AIDS (Abrams et al., 1994; Fleming et al., 1995; Goldman et al., 1996). The trial was conducted to compare didanosine (ddI) and zalcitabine (ddC) treatments for HIV-infected patients who were intolerant to or had failed zidovudine treatments. Of the 467 patients recruited for the study, 230 were randomized to receive ddI treatment and the other 237 received ddC treatment. The average follow up time is 15.6 months, and 188 patients died at the end of the study. Despite having longitudinal measurements that were measured at follow-up visits, Abrams et al.
(1994) showed that the ddC treatment is as efficacious as the ddI treatment in prolonging survival time, based on a proportional hazards model with covariates measured at the baseline visit. In what follows, we apply the proposed methods to investigate time-dependent risk factors for overall survival. We included baseline covariates at randomization such as age, gender, treatment received (ddI/ddC), and AIDS diagnosis (yes/no). We also included time-dependent covariates such as CD4 count, Karnofsky score, and cumulative recurrent opportunistic infections count. The CD4 count and Karnofsky score are measured at the baseline visit and bimonthly follow-up visits. We adopt the last covariate carried forward approach between visit times when constructing these time-dependent covariates. For the opportunistic infection, we use the cumulative number of infections prior to \( t \) as the covariate value at \( t \). Variables including Karnofsky score and CD4 count are transformed into the range \([0, 1]\) using the corresponding estimated cumulative distribution functions.

Figure 1 displays the proposed ROC guided survival tree using the CON splitting criterion. The \( \Delta \text{CON}_r \) splitting criterion yielded the same tree and the result is not shown. With the concordance-complexity pruning, there are three terminal nodes in the final tree. The terminal nodes are \( \tau_1 = \{ Z(t) \mid \text{KSC}(t) \leq 0.396 \} \), \( \tau_2 = \{ Z(t) \mid \text{KSC}(t) > 0.396, \text{OP}(t) = 0 \} \), and \( \tau_3 = \{ Z(t) \mid \text{KSC}(t) > 0.396, \text{OP}(t) > 0 \} \), where \( \text{KSC}(t) \) is the transformed Karnofsky score at \( t \) and \( \text{OP}(t) \) is the cumulative number of opportunistic infection up to \( t \). The partitions \( T = \{ \tau_1, \tau_2, \tau_3 \} \) corresponds to node 2, 6, and 7, whose estimated hazard rates are plotted in Figure 2a. Figure 2a clearly shows that lower Karnofsky score is associated with higher mortality risk. For these with high Karnofsky score (e.g., \( \text{KSC}(t) \geq 0.396 \)), previous opportunistic infections are also associated with higher mortality risk. Since a patient can enter different terminal nodes at different time points, one hazard curve does not necessarily represent the hazard of one patient. For example, a patient with a high Karnofsky score and no opportunistic infection enters node 6 and have relatively low mortality risk. As time
progress, the patient could enter node 7 if the patient experiences an opportunistic infection
or enter node 2 if the patient’s Karnofsky score decreases. Either progression leads to higher
mortality risk.

We also applied the proposed forest-based method, using the same setting as in our
simulation studies. For prediction, we considered hypothetical patients with Karnofsky score
that is either increasing linearly from 60 to 90 (patient A), decreasing linearly from 80 to 40
(patient B), or constant at 80 (patient C). Holding all other covariates constant at median
(or mode for binary covariates), the predicted hazard estimates are plotted in Figure 2b with
\( B = 500 \). The figure shows that Patient A has the highest mortality risk in \( t < 0.87 \), this
is because Patient A has the lowest Karnofsky score in this duration. On the other hand,
Patient B has the highest mortality risk in \( t > 0.87 \), reflecting that the Patient B has the
highest Karnofsky score in this duration. Lastly, Patient C has the lowest mortality risk
throughout.

[Figure 1 about here.]

[Figure 2 about here.]

7. Discussion

In this article, we propose a unified framework for survival trees and forests, where ROC* and
related summary measures guide the tree-growing algorithm. Compared to existing
approaches that maximize the within node homogeneity or between node heterogeneity,
our splitting procedure has an objective directly related to the prediction accuracy. The
proposed approach can deal with time-dependent covariates, thus can provide insight into
the association between failure events and risk factors changing over time. Moreover, we
extend the survival trees to random forests, which shows great potential in improving the
prediction accuracy.
For survival outcomes, we adopt the incident/dynamic (I/D) definition of ROC to evaluate the partition-based hazard functions. When the survival tree is constructed using only the baseline covariates \( X \), one may also consider the cumulative/dynamic (C/D) definition of ROC (Heagerty et al., 2000). Specifically, at each time \( t \), subjects with \( T > t \) serve as controls and subjects with \( T \leq t \) serve as cases. With the C/D type ROC, we focus on the baseline population other than the survivor population that changes over time. Following the arguments of the Neyman-Pearson Lemma, the conditional survival function \( P(T > t \mid X) \) maximizes the following concordance measure,

\[
\text{CON}^{C/D}_{t}(\tilde{g}(t, \cdot)) = P\{\tilde{g}(t, X_2) > \tilde{g}(t, X_1) \mid T_1 \leq t < T_2 \} + \frac{1}{2} P\{\tilde{g}(t, X_2) = \tilde{g}(t, X_1) \mid T_1 \leq t < T_2 \}.
\]

Thus \( \text{CON}^{C/D}(\tilde{g}) = \int_0^T \text{CON}^{C/D}_{t}(\tilde{g}(t, \cdot)) \omega(t) \, dt \) can also be used as a guidance to build a survival tree. Compared with the C/D approach that only permits baseline covariates, the I/D approach has the advantage to allow for time-dependent covariates.

**Appendix**

*Generalized definition of time-dependent ROC*

Let \( X \) be a marker where larger value indicating higher risk of failure. We then define the incident true positive rate at time \( t \) to be

\[
\text{TPR}^*_t(c, \gamma) = P(X > c \mid T = t) + \gamma P(X = c \mid T = t), \quad \gamma \in (0, 1),
\]

and the dynamic false positive rate to be

\[
\text{FPR}^*_t(c, \gamma) = P(X > c \mid T > t) + \gamma P(X = c \mid T > t), \quad \gamma \in (0, 1).
\]

The ROC curve at time \( t \) is \( \text{ROC}^*_t(q) = \text{TPR}^*_t(\{\text{FPR}^*_t\}^{-1}(q)) \).

*Random Forests Algorithm*

Suppose the new observation of covariate history up to \( t \) is \( \{z_u, 0 \leq u \leq t\} \), and we are interested in estimating \( \lambda(t, z_t) \) or the survival probability at \( t \) given \( \{z_u, 0 \leq u \leq t\} \). For
a node $\tau$ define $s_1(\tau) = \sum_{i=1}^n I(Z_i(Y_i) \in \tau)$ and $s_2(\tau) = \min_{t \in t_{uc}} \sum_{i=1}^n I(Z_i(t) \in \tau)$, where $t_{uc}$ is the set of distinct uncensored survival times in the training data on the time interval of interest. We do not make split on $\tau$ if $s_1(\tau) < n_{min,1}$ and $s_2(\tau) < n_{min,2}$. The quantities $n_{min,1}$ and $n_{min,2}$ are pre-specified constants. For example, $n_{min,1} = 3$ and $n_{min,2} = 1$.

Step 1. Initialize the weight at time $u$ to be $w_u = (w_{1u}, w_{2u}, \ldots, w_{nu}) = (0, 0, \ldots, 0)$ for $u \in t_{uc}$.

Step 2. Draw $B$ subsamples with size $s$ from the training data. Starting from the first subsample, randomly divide the data into two evenly-sized halves, $L_1$ and $L_2$. Grow a survival tree $T_1$ with $L_1$ using CON-based splitting criterion and a random selection of $m$ features at each split. At time $u$, for observations with $Y \geq u$, return the elements in $L_2$ that fall into the same leaf as $z_u$ in the tree $T_1$, denoted by $N_u$. If $N_u \neq \emptyset$, for $e \in N_u$, update $w_u$ with $w_u[e] \leftarrow w_u[e] + 1/\mid N_u \mid$, where $\mid N_u \mid$ is the number of elements in $N_u$. If $N_u = \emptyset$, $w_u$ remains the same.

Step 3. Repeat Step 2 and update the weights for the $b$th subsample, $b = 2, \ldots, B$.

Step 4. Given the weights $w_u$ obtained from Step 2–3, calculate the hazard function using

$$\int_0^{\infty} K_h(t - u) \sum_{i=1}^n w_{iu} dN_i(u) \sum_{i=1}^n w_{iu} I(Y_i \geq u),$$

or survival function using

$$\exp \left\{ - \int_0^t \sum_{i=1}^n w_{iu} dN_i(u) \sum_{i=1}^n w_{iu} I(Y_i \geq u) \right\}.$$
The algorithm to build ROC-guided survival trees is as follows. For a node $\tau$ define $s_1(\tau) = \sum_{i=1}^{n} I(Z_i(Y_i) \in \tau)$ and $s_2(\tau) = \min_{t \in t_{uc}} \sum_{i=1}^{n} I(Z_i(t) \in \tau)$, where $t_{uc}$ is the set of distinct uncensored survival times in the training data on the time interval of interest. We attempt to make a split on node $\tau$ only if $s_1(\tau) \geq n_{\min,1}$ or $s_2(\tau) \geq n_{\min,2}$, where the quantities $n_{\min,1}$ and $n_{\min,2}$ are pre-specified constants. For example, $n_{\min,1} = 20$ and $n_{\min,2} = 5$. Moreover, we only consider the splits such that $s_1, s_2$ of the children nodes are less then $n_{\min,1}/3$ and $n_{\min,2}/3$, respectively.

**Splitting based on CON**

Step 0. Start from the tree $T_1$ with only the root node.

Step k. Given the tree $T_k$ after the $(k-1)$th split and $T_k = \{\tau_{k1}, \ldots, \tau_{kk}\}$, among the splittable terminal nodes, find the optimal split that maximize the in-sample concordance. Suppose the optimal split is on node $\tau_{kj}$, then split the node $\tau_{kj}$ into two children nodes $\tau_{kj}^L$ and $\tau_{kj}^R$ with the optimal binary rule, and set $T_{k+1} = \{\tau_{km}, \tau_{kj}^L, \tau_{kj}^R; m \neq j\}$. Stop if all the terminal nodes in $T_{k+1}$ are not splittable.

Splitting based on the increment of within-node concordance $\Delta\text{CON}_{\tau}$ follows a similar procedure and is thus omitted.

**Pruning**

Let $T_{\text{max}}$ be the fully grown tree and $K = |T_{\text{max}}|$. For $1 \leq k \leq K$, let $T_{(k)}$ be the subtree of $T_{\text{max}}$ such that $|T_{(k)}| = k$ and $T_{(k)}$ has the largest in-sample concordance among all the size-$k$ subtrees. The sequence of optimal subtrees is $C = \{T_{(1)}, T_{(2)}, \ldots, T_{(K)}\}$. For $\alpha_0 = 0$, the initial tree $T_{\alpha_0} = T_{(K)} = T_{\text{max}}$ is the optimal subtree regarding $\text{CON}_{\alpha_0}$. Define $\alpha_{T, T'} =$
\[ \overline{\text{CON}(\lambda_T)} - \overline{\text{CON}(\hat{\lambda}_T)} \]. The \( q \)th \( (q \geq 1) \) threshold parameter \( \alpha_q \) is defined as

\[
\alpha_q = \min\{ \alpha_T, T^{\alpha_{q-1}} \cdot |T| \leq |T^{\alpha_{q-1}}|, T \in \mathcal{C} \},
\]

and \( T^{\alpha_q} \) is defined as the smallest tree in \( \{ T \mid \alpha_T, T^{\alpha_{q-1}} = \alpha_q, |T| \leq |T^{\alpha_{q-1}}|, T \in \mathcal{C} \} \).

**Step 0.** Find the sequence of optimal subtrees \( \mathcal{C} \) of the fully grown tree \( T_{\text{max}} \).

**Step 1.** Calculate \( \alpha_q \) for \( q = 1, 2, \ldots \), and \( T^{\alpha_q} \) denotes the tree that only contains the root node. For \( \alpha \in [\alpha_q, \alpha_{q+1}) \) and \( q < Q \), \( T^{\alpha_q} \) is the optimal subtree.

**Step Q+1.** For \( q < Q \), set \( \beta_q = \sqrt{\alpha_q \alpha_{q+1}} \) as the representative value of the intervals \( [\alpha_q, \alpha_{q+1}) \) and \( \beta_Q = \alpha_Q \), then \( T^{\beta_q} = T^{\alpha_q} \) for \( q = 1, \ldots, Q \). Select the optimal \( \beta_q \) and the final tree using ten-fold cross validation.

2. **Proof of Proposition 1**

We show that \( \lambda(t, Z(t)) \) yields the highest \( \text{ROC}_t^* \) curve. Suppose \( \{Z(t) \mid T \geq t\} \) takes value in \( \mathcal{Z}_t \subset \mathbb{R}^p \). For a scalar function \( g : \mathcal{Z}_t \mapsto \mathbb{R} \), define \( \phi_g(z) = I(g(z) > c) + \gamma I(g(z) = c) \), then \( \text{FPR}_g^*(c, \gamma) = E\{\phi_g(Z(t)) \mid T > t\} \) and \( \text{TPR}_g^*(c, \gamma) = E\{\phi_g(Z(t)) \mid T = t\} \). We use \( P_1(z) \) to denote the probability distribution of \( Z(t) \) conditioning on \( T = t \) and use \( P_0(z) \) to denote the probability distribution of \( Z(t) \) conditioning on \( T > t \).

For any fixed \( \alpha \in (0, 1) \), we choose \( c \) and \( \gamma \) such that \( \text{FPR}_{g, t}^*(c, \gamma) = \text{FPR}_{g, t}^*(\lambda(t, \cdot), t)(c, \gamma) = \alpha \), then \( \int_{\mathcal{Z}_t} \{\phi_{\lambda(t, \cdot)}(z) - \phi_g(z)\} dP_0(z) = 0 \). Define \( S^+ = \{ z : \phi_{\lambda(t, \cdot)}(z) - \phi_g(z) > 0, z \in \mathcal{Z}_t \} \) and \( S^- = \{ z : \phi_{\lambda(t, \cdot)}(z) - \phi_g(z) < 0, z \in \mathcal{Z}_t \} \). Then we have \( \lambda(t, z) \geq c \) for \( z \in S^+ \) and \( \lambda(t, z) \leq c \) for \( z \in S^- \). Moreover, \( \lambda(t, z) \geq c \) is equivalent to \( \frac{dP_1(z)}{dP_0(z)} \geq c^* \overset{\text{def}}{=} c/\lambda(t) \) and \( \lambda(t, z) \leq c \) is equivalent to \( \frac{dP_1(z)}{dP_0(z)} \leq c^* \), where \( \lambda(t) \) is the hazard function of \( T \). Then we have \( \int_{S^+} \{\phi_{\lambda(t, \cdot)}(z) - \phi_g(z)\} dP_1(z) \geq c^* \int_{S^+} \{\phi_{\lambda(t, \cdot)}(z) - \phi_g(z)\} dP_0(z) \) and
\[ \int_{S^c} \{ \phi_{\lambda(t, \cdot)}(z) - \phi_{\gamma}(z) \} \, dP_1(z) \geq c^* \int_{S^c} \{ \phi_{\lambda(t, \cdot)}(z) - \phi_{\gamma}(z) \} \, dP_0(z), \text{ thus} \]

\[
TPR^*_{\lambda(t, \cdot), t}(c, \gamma) - TPR^*_{g,t}(c, \gamma) = E\{ \phi_{\lambda(t, \cdot)}(Z(t)) \mid T = t \} - E\{ \phi_{g}(Z(t)) \mid T = t \}
= \int_{Z_t} \{ \phi_{\lambda(t, \cdot)}(z) - \phi_{\gamma}(z) \} \, dP_1(z) \geq c^* \int_{Z_t} \{ \phi_{\lambda(t, \cdot)}(z) - \phi_{\gamma}(z) \} \, dP_0(z) \geq 0.
\]

We have proved that, for any FPR^*_{\lambda(t, \cdot), t}(c, \gamma) = FPR^*_{g,t}(c, \gamma) = \alpha, we have TPR^*_{\lambda(t, \cdot), t}(c, \gamma) \geq TPR^*_{g,t}(c, \gamma). Thus the time-dependent ROC^* curve at \( t \) of \( \lambda(t, Z(t)) \) is always higher than or the same as that of \( g(Z(t)) \) and \( \text{CON}_t(\lambda(t, \cdot)) \geq \text{CON}_t(g) \).

### 3. Proof of Proposition 2

Consider the partition \( \mathcal{T} = \{ \tau_1, \tau_2, \ldots, \tau_M \} \) and a split \( s \) on any node \( \tau_j \in \mathcal{T} \), and the partition after splitting is denoted by \( \mathcal{T}_s = \{ \tau_m, \tau_j^L, \tau_j^R \mid m \neq j, m = 1, \ldots, M \} \). In what follows, we show that \( \text{ROC}^*_{\mathcal{T}_s}(q) \geq \text{ROC}^*_{\mathcal{T}_t}(q) \) for \( q \in (0, 1) \). Without loss of generality, we assume \( \lambda(t \mid \tau_1) \geq \lambda(t \mid \tau_2) \geq \ldots \geq \lambda(t \mid \tau_M) \) and \( \lambda(t \mid \tau_j^L) < \lambda(t \mid \tau_j) < \lambda(t \mid \tau_j^R) \). We use the notation \( P(\tau \mid 0) = P(Z(t) \in \tau \mid T > t) \) and \( P(\tau \mid 1) = P(Z(t) \in \tau \mid T = t) \). Define the set of terminal nodes \( \mathcal{T}_1 = \{ \tau \in \mathcal{T} : \lambda(t \mid \tau) \leq \lambda(t \mid \tau_j^L) \} \), \( \mathcal{T}_2 = \{ \tau \in \mathcal{T} : \lambda(t \mid \tau_j^L) < \lambda(t \mid \tau) \leq \lambda(t \mid \tau_j) \} \), \( \mathcal{T}_3 = \{ \tau \in \mathcal{T} : \lambda(t \mid \tau_j) \leq \lambda(t \mid \tau) < \lambda(t \mid \tau_j^R), \tau \neq \tau_j \} \), \( \mathcal{T}_4 = \{ \tau \in \mathcal{T} : \lambda(t \mid \tau) \geq \lambda(t \mid \tau_j^R) \} \). For \( k = 1, \ldots, 4 \), define \( P(\mathcal{T}_k \mid 0) = \sum_{\tau \in \mathcal{T}_k} P(Z(t) \in \tau \mid T > t) \) and \( P(\mathcal{T}_k \mid 1) = \sum_{\tau \in \mathcal{T}_k} P(Z(t) \in \tau \mid T = t) \).

Define \( s(t \mid \tau) = \lambda(t \mid \tau) / \lambda(t) \), then \( s(t \mid \tau_1) \geq s(t \mid \tau_2) \geq \ldots \geq s(t \mid \tau_M) \) and \( s(t \mid \tau_j^L) < s(t \mid \tau_j) < s(t \mid \tau_j^R) \). The \( \text{ROC}^*_t \) for \( \lambda_{\mathcal{T}} \) can be derived as,

\[
\text{ROC}^*_t = \begin{cases} 
    s(t \mid \tau_1), & 0 < q \leq P(\tau_1 \mid 0), \\
    \sum_{k=1}^{m-1} P(\tau_k \mid 1) + s(t \mid \tau_m) \{ q - \sum_{k=1}^{m-1} P(\tau_k \mid 0) \}, & \sum_{k=1}^{m-1} P(\tau_k \mid 0) < q \leq \sum_{k=1}^{m} P(\tau_k \mid 0),
\end{cases}
\]

which is a convex function on \([0, 1] \). The \( \text{ROC}^*_t \) for \( \lambda_{\mathcal{T}_s} \) can be derived similarly. Then we have \( \text{ROC}^*_{\mathcal{T}_s}(q) = \text{ROC}^*_{\mathcal{T}_t}(q) \) on \( q \in [0, P(\mathcal{T}_4 \mid 0)] \) and \( q \in [1 - P(\mathcal{T}_4 \mid 0), 1] \).
We also define \( \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) \) such that

\[
\widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) = \begin{cases} 
\sum_{k=1}^{j-1} P(\tau_k | 1) + s(t | \tau_j^R)(q - \sum_{k=1}^{j-1} P(\tau_k | 0)), & \sum_{k=1}^{j-1} P(\tau_k | 0) < q \leq \sum_{k=1}^{j-1} P(\tau_k | 0) + P(\tau_j^R | 0), \\
\sum_{k=1}^{j-1} P(\tau_k | 1) + P(\tau_j^R | 0) + s(t | \tau_j^R)(q - \sum_{k=1}^{j-1} P(\tau_k | 0) - P(\tau_j^R | 0)), & \sum_{k=1}^{j-1} P(\tau_k | 0) + P(\tau_j^R | 0) < q < \sum_{k=1}^{j} P(\tau_k | 0), \\ \text{ otherwise.}
\end{cases}
\]

Then it can be easily seen that \( \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) > \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) \) for \( \sum_{k=1}^{j-1} P(\tau_k | 0) < q < \sum_{k=1}^{j} P(\tau_k | 0) \) (or equivalently, for \( P(\mathcal{T}_4 | 0) + P(\mathcal{T}_3 | 0) < q < 1 - P(\mathcal{T}_4 | 0) - P(\mathcal{T}_2 | 0) \) and \( \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) = \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) \) otherwise. Following similar arguments as in the proof of Proposition 1, we can show that

(i) If \( \mathcal{T}_3 = \emptyset \), \( \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) = \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) > \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) \) for \( q \in (P(\mathcal{T}_4 | 0), P(\mathcal{T}_4 | 0) + P(\tau_j^R | 0)) \).

(ii) If \( \mathcal{T}_3 \neq \emptyset \), \( \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) > \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) \geq \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) \) for \( q \in (P(\mathcal{T}_4 | 0), P(\mathcal{T}_4 | 0) + P(\mathcal{T}_3 | 0) + P(\tau_j^R | 0), 1 - P(\mathcal{T}_3 | 0)) \).

(iii) If \( \mathcal{T}_2 = \emptyset \), \( \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) = \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) > \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) \) for \( q \in (P(\mathcal{T}_4 | 0) + P(\mathcal{T}_3 | 0) + P(\tau_j^R | 0), 1 - P(\mathcal{T}_3 | 0)) \).

(iv) If \( \mathcal{T}_2 \neq \emptyset \), \( \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) > \widetilde{\text{ROC}}_{\lambda_{\tau}(t, \cdot), t}(q) \geq \text{ROC}_{\lambda_{\tau}(t, \cdot), t}(q) \) for \( q \in (P(\mathcal{T}_4 | 0) + P(\mathcal{T}_3 | 0) + P(\tau_j^R | 0), 1 - P(\mathcal{T}_3 | 0)) \).

Therefore, if \( \lambda(t | \tau_j^R) > \lambda(t | \tau_j^L) \), after splitting, \( \text{ROC}_{\tau}(t, \cdot) \) increases on the interval \( (P(\mathcal{T}_4 | 0), 1 - P(\mathcal{T}_3 | 0)) \) and stays the same on \( [0, P(\mathcal{T}_4 | 0)] \) and \( [1 - P(\mathcal{T}_3 | 0), 1] \), and thus \( \text{CON}_{t}(\lambda_{\tau}(t, \cdot)) > \text{CON}_{t}(\lambda_{\tau}(t, \cdot)) \). If \( \lambda(t | \tau_j^R) = \lambda(t | \tau_j^L) \), it is easy to see that the \( \text{ROC}_{\tau}(t, \cdot) \) does not change after splitting, thus \( \text{CON}_{t}(\lambda_{\tau}(t, \cdot)) = \text{CON}_{t}(\lambda_{\tau}(t, \cdot)) \).

By replacing all the quantities above with their estimates, we can prove the second part.
of Proposition 2. Specifically, define

$$\hat{f}(t, \tau) = \int_0^\infty K_h(t-u) \hat{E}\{I(Z(Y) \in \tau) dN(u)\},$$

Then $\hat{\lambda}(t | \tau) = \hat{f}(t, \tau)/\hat{S}(t, \tau), \hat{P}(\tau|1) = \hat{f}(t, \tau)/\sum_{\tau \in T} \hat{f}(t, \tau), \hat{P}(\tau|0) = \hat{S}(t, \tau)/\sum_{\tau \in \hat{T}} \hat{S}(t, \tau).$

Without loss of generality, we assume $\hat{\lambda}(t | \tau_1) \geq \hat{\lambda}(t | \tau_2) \geq \ldots \geq \hat{\lambda}(t | \tau_M)$ and $\hat{\lambda}(t | \tau_j^R) < \hat{\lambda}(t | \tau_j^R).$ Define $\tilde{s}(t | \tau) = \hat{\lambda}(t | \tau)/\hat{\lambda}(t),$ where $\hat{\lambda}(t) = \sum_{\tau \in T} \hat{f}(t, \tau)/\sum_{\tau \in T} \hat{S}(t, \tau).$

Then it can be shown that

$$\text{ROC}^*_\hat{\lambda}(t; r, t)(q) = \begin{cases} \tilde{s}(t | \tau_1)q, & 0 < q \leq \hat{P}(\tau_1|0), \\ \sum_{k-1}^{m-1} \hat{P}(\tau_k|1) + \tilde{s}(t | \tau_m)(q - \sum_{k=1}^{m-1} \hat{P}(\tau_k|0)), & \sum_{k=1}^{m-1} \hat{P}(\tau_k|0) < q \leq \sum_{k=1}^{m} \hat{P}(\tau_k|0), \end{cases}$$

is the in-sample estimate of $\text{ROC}^*_t$ of $\lambda_T$ and $\text{CON}_t(\hat{\lambda}_T(t, \cdot)) = \int_0^1 \text{ROC}^*_\hat{\lambda}_T(t; r, t)(q)dq.$ Following similar arguments as before, we can prove the second part of Proposition 2.

4. Proof of Theorem 1

(A1) We assume the censoring time $C$ satisfies $P(t \leq T < t + dt | Z(t) = z, T \geq t, C \geq t) = \lambda(t, z) dt.$

(A2) The time-dependent covariate $Z(t)$ is left-continuous in $t.$ There exists a constant $c_1$ such that $f_{Z(t)|Y \geq t}(z)P(Y \geq t) > c_1$ for $z \in [0, 1]^p,$ where $f_{Z(t)|Y \geq t}(z)$ is the density of $Z(t)$ given $Y \geq t.$ At time $t,$ $Z(t)$ is distributed to a bounded density $f_t(z).$

(A3) The function $f^{uc}(t | z) = \lim_{\delta_t \to 0^+} P\{T \in [t, t + \delta_t), \Delta = 1 | Z(T) = z\}/\delta_t$ is second order differentiable with respect to $t,$ and $\sup_{t \in [0, 1], z \in [0, 1]^p} \vert \partial^2 f^{uc}(t | z)/\partial t^2 \vert < c_2$ for some constant $c_2.$

(A4) For the tree $T_n,$ its size $k$ grows as $o\left(\frac{nh}{\log n}\right).$ For any $\gamma > 0,$ the diameters of the nodes satisfies $\mu(z : \text{diam}(l_{T_n}(z)) > \gamma) \to 0$ with probability 1, where $\mu$ is the Lebesgue measure.

(A5) The bandwidth $h$ satisfies $h = n^{-\alpha},$ $0 < \alpha < 1.$

Let $\mu(\tau)$ be the Lebesgue measure of the set $\tau \subset [0, 1]^p.$ Define $f(t, z), S(t, z)$ such that $f(t, z) dt dz = P\{T \in [t, t + dt), Z(t) \in [z, z + dz]\},$ and $S(t, z) dz = P\{T \geq t, Z(t) \in [z, z + dz]\},$ where $\mu(\tau)$ is the Lebesgue measure of the set $\tau \subset [0, 1]^p.$ Following similar arguments as before, we can prove the second part of Proposition 2.
\[ [z, z + dz] \}. \] We first define the following quantities,

\[
\hat{r}_n(t, z) = \frac{\sum_{i=1}^{n} \int_0^\infty K_h(t - u)I(Z_i(Y_i) \in l_n\{z\}) \mathrm{d}N_i(u)}{n \mu(l_n\{z\})},
\]

\[
r(t, z) = P(C \geq t \mid T = t, Z(t) = z)f(t, z),
\]

\[
\hat{r}_{2n}(t, z) = \frac{\sum_{i=1}^{n} I(Y_i \geq t, Z_i(t) \in l_n(z))}{n \mu(l_n\{z\})},
\]

\[
r_2(t, z) = P(C \geq t \mid T \geq t, Z(t) = z)S(t, z),
\]

where the terms with \( ^\sim \) are the corresponding estimates, and \( l_n \) is the partition function based on \( T_n \), that is, \( l_n \overset{\text{def}}{=} l_{T_n} \). It can be easily seen that

\[
\hat{\lambda}_{T_n}(t, z) = \frac{\hat{r}_n(t, z)}{\hat{r}_{2n}(t, z)}, \quad \lambda(t, z) = \frac{r(t, z)}{r_2(t, z)}.
\]

To study the large sample property of \( \hat{\lambda}_{T_n} \), we first prove the strong \( L_1 \)-consistency of \( \hat{r}_n \) and \( \hat{r}_{2n} \).

In what follows, let \( X = Z(Y) \) and \( W = Z(t) \). Let \( P \) denote the joint distribution of \( O = \{X, W, Y, \Delta\} \) and \( \mathcal{O} = [0, 1]^p \times [0, 1]^p \times \mathbb{R}^+ \times \{0, 1\} \) denote the sample space of \( O \). For \( g : \mathcal{O} \mapsto \mathbb{R} \), define \( Pg = \int g(o)P(\mathrm{d}o) \) and \( \hat{P}_n g = \frac{1}{n} \sum_{i=1}^{n} g(O_i) \). We now prove for each fixed \( t \), \( \int_{[0, 1]^p} |\hat{r}_n(t, z) - r(t, z)| \, \mathrm{d}z = o^*(1) \), where a sequence \( U_1, U_2, \ldots \) of random variables defined on the same probability space as \( (X, W, Y, \Delta) \) is said to be of order \( o^*(1) \), written as \( U_n = o^*(1) \), if as \( n \to \infty \), \( U_n \to 0 \) with probability 1.

Define \( F^{uc}(t \mid z) = E\{N(t) \mid X = z\} \) and \( f^{uc}(t \mid z) = \frac{dF^{uc}(t \mid z)}{dt} \), then \( r(t, z) = f^{uc}(t \mid z)f_X(z) \). Define

\[
\tilde{r}_n(t, z) = \frac{\nu(t, l_n\{z\})}{\mu(l_n\{z\})},
\]

where \( \nu(t, \tau) \overset{\text{def}}{=} \int_{z \in \tau} r(t, z) \, \mathrm{d}z = Pf^{uc}(t \mid X)I(X \in \tau) \) for \( \tau \subset [0, 1]^p \).

It is clear that

\[
\int_{[0, 1]^p} |\tilde{r}_n(t, z) - r(t, z)| \, \mathrm{d}z \leq \int_{[0, 1]^p} |\tilde{r}_n(t, z) - \tilde{r}_n(t, z)| \, \mathrm{d}z + \int_{[0, 1]^p} |\tilde{r}_n(t, z) - r(t, z)| \, \mathrm{d}z \overset{\text{def}}{=} I_1 + I_2.
\]
Given a class of real-valued function $\mathcal{G}$ and a partition family $\Pi$ on $\mathcal{O}$, define the function class

$$
\mathcal{G} \circ \Pi = \left\{ \sum_{\tilde{\tau}_m \in T} g_m(x, w, y, \delta) I_{\tilde{\tau}_m} : T = \{ \tilde{\tau}_m \} \in \Pi, g_m \in \mathcal{G} \right\}, \quad j = 1, 2,
$$

where $I_{\tilde{\tau}}(x, w, y, \delta) = 1$ iff $(x, w, y, \delta) \in \tilde{\tau} \subset \mathcal{O}$. Moreover, define $V_n(\mathcal{G} \circ \Pi) = \sup_{\eta \in \mathcal{G} \circ \Pi} \left| \hat{P}_n(\eta) - P(\eta) \right|$. The following Lemmas are useful for establishing the asymptotic properties of $I_1$ and $I_2$.

**Lemma 1** (? Proposition 2):

$$
\sup_{g \in \mathcal{G}} \sup_{T \in \Pi} \sum_{\tilde{\tau}_m \in T} \left| \hat{P}_n(g_m I_{\tilde{\tau}_m}) - P(g_m I_{\tilde{\tau}_m}) \right| \leq V_n(\mathcal{G}' \circ \Pi)
$$

where $\mathcal{G}' = \mathcal{G} \cup -\mathcal{G}$ contains every function in $\mathcal{G}$ and its additive inverse.

Consider a partition family $\Pi^k_X$ on $\mathcal{O}$ such that $\Pi^k_X$ contains all axis parallel partitions based on $X$ and $\Pi^k_X$ has $k$ terminal nodes. Specifically, each split is based on a single coordinate of $X$, say, $X_q$, and takes the form $\{ \text{Is } X_q > c? \}$, $c \in [0, 1]$, $q \in \{1, \ldots, p\}$. Let $\Pi^{k,n}_X$ denote the axis parallel partition family with $k$ terminal nodes, and each split of $T \in \Pi^{k,n}_X$ takes the form $\{ \text{Is } X_q > c_j? \}$, $j = 1, \ldots, n'$, $c_j$ are distinct data values of $X_q$ in the training sample, and there are at most $n$ different splits on $X_q$. Since each tree with $k$ terminal nodes has $k - 1$ splits, thus there are no more than $(np)^{(k-1)}$ partitions in $\Pi^{k,n}_X$. In other words, if we only consider axis-parallel partitions, $\{X_i; \quad i=1, \ldots, n\}$ can be partitioned in at most $(np)^{(k-1)}$ ways. Similarly, we can define a partition family $\Pi^k_W$ on $\mathcal{O}$ such that $T \in \Pi^k_W$ contains $k$ terminal nodes, the splits in $T$ are based on a single coordinate of $W$ and take the form $\{ \text{Is } W_q > c? \}$, $c \in [0, 1]$, $q \in \{1, \ldots, p\}$. We also define $k$-node partition family $\Pi^{k,n}_W$ where the splits take the form $\{ \text{Is } W_q > c? \}$ and $c$ takes distinct data values of $W_q$ in the training sample. Moreover, if $g \in \mathcal{G}$ does not depend on $w$, it can be easily seen that $V_n(\mathcal{G} \circ \Pi^k_X) = V_n(\mathcal{G} \circ \Pi^{k,n}_X)$; if $g \in \mathcal{G}$ does not depend on $x$, we have $V_n(\mathcal{G} \circ \Pi^k_W) = V_n(\mathcal{G} \circ \Pi^{k,n}_W)$.

**Lemma 2** (Bracketing number of $\mathcal{G} \circ \Pi^{k,n}_X$ and $\mathcal{G} \circ \Pi^{k,n}_W$): Let $\mathcal{G}$ be a function class with
finite number of elements. For every \( \epsilon > 0 \),

\[
N_{[]} (\epsilon, \mathcal{G} \circ \Pi_{X}^{k,n}, L_1(P)) \leq (np)^{(k-1)}|\mathcal{G}|^{k},
\]

\[
N_{[]} (\epsilon, \mathcal{G} \circ \Pi_{W}^{k,n}, L_1(P)) \leq (np)^{(k-1)}|\mathcal{G}|^{k},
\]

where \( |\mathcal{G}| \) is the number of elements in \( \mathcal{G} \).

**Proof.** For any \( f \in \mathcal{G} \circ \Pi_{X}^{k,n} \), we have \( f = \sum_{m=1}^{k} g_m I_{\tilde{\tau}_m} \) and \( \{\tilde{\tau}_m\} \in \Pi_{X}^{k,n} \). There exists a \( \epsilon \)-bracket \( \{ (\sum_{m=1}^{k} g_m - \epsilon/2) I_{\tilde{\tau}_m}, \sum_{m=1}^{k} g_m + \epsilon/2) I_{\tilde{\tau}_m} \} \) such that \( f \) lies in the bracket. For a fixed \( T \in \Pi_{X}^{k,n} \), there are \( |\mathcal{G}|^{k} \) functions in \( \mathcal{G} \circ \Pi_{X}^{k,n} \) based on \( T \). Since the number of partitions in \( \Pi_{X}^{k,n} \) is no more than \( (np)^{(k-1)} \), there are at most \( (np)^{(k-1)}|\mathcal{G}|^{k} \) functions in \( \mathcal{G} \circ \Pi_{X}^{k,n} \). Hence we have \( N_{[]} (\epsilon, \mathcal{G} \circ \Pi_{X}^{k,n}, L_1(P)) \leq (np)^{(k-1)}|\mathcal{G}|^{k} \).

**Lemma 3** (? , Theorem 2.1): Suppose that for all \( n \geq 1 \) we have for all \( g \in \mathcal{G}_n \), \( P|g| \leq A(n) \) and \( \|g\|_{\infty} \leq B(n) \). Let \( R(n) = n/A(n)B(n) \). Suppose there is a non-decreasing function \( D(n) \) such that \( D(n) = o(R(n)) \) and \( N_{[]} (\epsilon, \mathcal{G}_n, L_1(P)) \leq \exp\{E(\epsilon)D(n)\} \) for all \( \epsilon > 0 \), where \( E(\epsilon) \) is a constant depending only on \( \epsilon \) and \( E(\epsilon) < \infty \) for all \( \epsilon > 0 \). If \( \sum_{i=1}^{n} \exp\{cR(n)\} < \infty \) for all \( c < 0 \), then \( \sup_{g \in \mathcal{G}_n} \left| \hat{P}_n g - Pg \right| = o^*(1) \).

**Lemma 4:** Let \( g \) be a function on \( \mathcal{O} \) that possibly depend on \( n \) and satisfies \( P|g| \leq A(n) \) and \( \|g\|_{\infty} \leq B(n) \). Let \( R(n) = n/A(n)B(n) \). Suppose \( \sum_{i=1}^{n} \exp\{cR(n)\} < \infty \) for all \( c < 0 \) and \( k \log n = o(R(n)) \).

(i) If \( g(x, w, y, \delta) \) does not depend on \( w \), then

\[
\sup_{T \in \Pi_{X}^{k,n}} \sum_{\tilde{\tau} \in T} \left| \hat{P}_n (g I_{\tilde{\tau}}) - P(g I_{\tilde{\tau}}) \right| = o^*(1).
\]

(ii) If \( g(x, w, y, \delta) \) does not depend on \( x \), then

\[
\sup_{T \in \Pi_{W}^{k,n}} \sum_{\tilde{\tau} \in T} \left| \hat{P}_n (g I_{\tilde{\tau}}) - P(g I_{\tilde{\tau}}) \right| = o^*(1).
\]
Proof. We now prove (i), then (ii) can be proved with similar arguments. Following Lemma 1, we have
\[
\sup_{T \in \Pi_k^X} \sum_{\tau \in T} \left| \hat{P}_n(g I_{\tau}) - P(g I_{\tau}) \right| \leq V_n(\{g, -g\} \circ \Pi^k_X) = V_n(\{g, -g\} \circ \Pi^{k,n}_X).
\]
Following Lemma 2, the function class \( \mathcal{G}_n = \{g, -g\} \circ \Pi_{X}^{k,n} \) has bracketing number \( N_0(\epsilon, \mathcal{G}_n, L_1(P)) \leq \exp \{ (k - 1) \log(np) + k \log 2 \} \). Thus by Lemma 3, \( V_n(\{g, -g\} \circ \Pi^{k,n}_X) = o^*(1) \) and we have proved \( \sup_{T \in \Pi_k^X} \sum_{\tau \in T} \left| \hat{P}_n(g I_{\tau}) - P(g I_{\tau}) \right| = o^*(1) \).

Given Lemma 1-4, we now prove \( I_1 = o^*(1) \). Let \( g_1(O) = \int_0^\infty K_h(t - u) \, dN(u) \), then
\[
P\{g_1(O)I(X \in \tau)\} = P \int_0^\infty K_h(t - u) \, dN(u)I(X \in \tau) = P \int_0^\infty K_h(t - u) \, dF^\text{uc}(u | X)I(X \in \tau).
\]
Moreover, we have
\[
|P\{g_1(O)I(X \in \tau)\} - \nu(t, \tau)| = P \int_0^1 K(s) f^\text{uc}(t - sh | X) \, dsI(X \in \tau) - P\{f^\text{uc}(t | X)I(X \in \tau)\} \leq M_1 h^2 P\{I(X \in \tau)\}. \tag{9}
\]
Let \( \mathcal{T}_n \) be the partition created using the learning sample, and the number of terminal nodes is \( k \).
\[
I_1 = \sum_{\tau \in \mathcal{T}_n} \left| \frac{1}{n} \sum_{X_i \in \tau} \int_0^\infty K_h(t - u) dN_i(u) - \nu(t, \tau) \right| = \sum_{\tau \in \mathcal{T}_n} \left| \hat{P}_n\{g_1(O)I(X \in \tau)\} - \nu(t, \tau) \right| \leq \sum_{\tau \in \mathcal{T}_n} \left| \hat{P}_n\{g_1(O)I(X \in \tau)\} - P\{g_1(O)I(X \in \tau)\} \right| + \sum_{\tau \in \mathcal{T}_n} \left| P\{g_1(O)I(X \in \tau)\} - \nu(t, \tau) \right| \leq \sum_{\tau \in \mathcal{T}_n} \left| \hat{P}_n\{g_1(O)I(X \in \tau)\} - P\{g_1(O)I(X \in \tau)\} \right| + M_1 h^2 \quad \text{By (9)}
\]
Applying Lemma 4, we have \( I_1 = o^*(1) \). We next prove \( I_2 = o^*(1) \). Note that \( r(t, z) \) may not
be continuous in \( z \), and we use \( r'(t, z) \) to approximate \( r(t, z) \). Specifically, given a small \( \epsilon > 0 \), there exists a uniformly continuous function \( r' \) on \([0, 1]^p\) such that \( \int_{[0,1]^p} |r(t,z) - r'(t,z)| \, dz < \epsilon \). Moreover, let \( \gamma \) be chosen so that \( |r'(t,z_1) - r'(t,z_2)| < \epsilon \) whenever \( \|z_1 - z_2\| \leq \gamma \). Let \( \nu'(t,\tau) = \int_{z \in \tau} r'(t, z) \, dz \), we define

\[
\nu'(t,\tau) = \frac{\nu'(t, l_n\{z\})}{\mu(l_n\{z\})}.
\]

Then we have

\[
I_2 = \int_{[0,1]^p} |\tilde{r}_n(t, z) - r(t, z)| \, dz \\
\leq \int_{[0,1]^p} |\tilde{r}_n(t, z) - r_n'(t, z)| \, dz + \int_{[0,1]^p} |r_n'(t, z) - r'(t, z)| \, dz + \int_{[0,1]^p} |r'(t, z) - r(t, z)| \, dz. \tag{10}
\]

The first term in \((10)\) is

\[
\int_{[0,1]^p} |\tilde{r}_n(t, z) - r_n'(t, z)| \, dz = \sum_{\tau \in T_n} |\nu(t, \tau) - \nu'(t, \tau)| \leq \int_{[0,1]^p} |r(t, z) - r'(t, z)| \, dz < \epsilon.
\]

It is easy to see that for fixed \( t \), there exist \( M_2 \) such that \( r'(t, z) \leq M_2 \). The second term in \((10)\) is

\[
\int_{[0,1]^p} |r_n'(t, z) - r'(t, z)| \, dz = \sum_{\tau \in T_n} \int_{\tau} \left| r'(t, z) - \frac{\nu'(t, \tau)}{\mu(\tau)} \right| \, dz \\
= \sum_{\tau \in T_n} \mu(\tau)^{-1} \int_{\tau} |r'(t, z)\mu(\tau) - \nu'(t, \tau)| \, dz \\
\leq \sum_{\tau \in T_n} \mu(\tau)^{-1} \int_{\tau \times \tau} |r'(t, z_1) - r'(t, z_2)| \, dz_1 \, dz_2 \\
\leq \sum_{\text{diam}(\tau) \leq \gamma} \mu(\tau) \epsilon + \sum_{\text{diam}(\tau) > \gamma} 2 \int_{\tau} |r'(t, z)| \, dz \\
\leq \epsilon + 2M_2\mu\{z : \text{diam}(l_n\{z\}) > \gamma\}.
\]

Thus \( I_2 \leq 3\epsilon + 2M_2\mu\{z : \text{diam}(l_n\{z\}) > \gamma\} \), and we have \( I_2 = o^*(1) \).

To prove the consistency of \( \tilde{r}_{2n} \), we define

\[
\tilde{r}_{2n}(t, z) = \frac{\nu_2(t, l_n\{z\})}{\mu(l_n\{z\})},
\]

where \( \nu_2(t, \tau) = P\{S_C(t \mid T \geq t, W) \mid S(t, W) I(W \in \tau)\} = \int_{z \in \tau} r_2(t, z) f_W(z) \, dz \) and \( S_C(t \mid
\( T \geq t, z \) \( \overset{\text{def}}{=} \) \( P(C \geq t \mid T \geq t, W = z) \).

\[
\int_{[0,1]^p} |\tilde{r}_2(t, z) - r_2(t, z)| \, dz \leq \int_{[0,1]^p} |\tilde{r}_{2n}(t, z) - \tilde{r}_{2n}(t, z)| \, dz + \int_{[0,1]^p} |\tilde{r}_{2n}(t, z) - r_2(t, z)| \, dz \leq I_3 + I_4.
\]

Define \( g_2(O) = I(Y \leq t) \), then \( \nu_2(t, \tau) = P\{g_2(O)I(W \in \tau)\} \).

\[
I_3 = \sum_{\tau \in \mathcal{T}_n} \left| \frac{1}{n} \sum_{W_i \in \tau} I(Y_i \leq t) - \nu_2(t, \tau) \right|
= \sum_{\tau \in \mathcal{T}_n} \left| \hat{P}_n\{g_2(O)I(W \in \tau)\} - \nu_2(t, \tau) \right|
\leq \sum_{\tau \in \mathcal{T}_n} \left| \hat{P}_n\{g_2(O)I(W \in \tau)\} - P\{g_2(O)I(W \in \tau)\} \right|
\leq \sup_{\tau \in \mathcal{T}_n} \sum_{\tau \in \mathcal{T}_n} \left| \hat{P}_n(g_2I_\tau) - P(g_2I_\tau) \right|
= o^*(1). \quad (\text{Lemma 4})
\]

With similar argument as in proving \( I_2 = o^*(1) \), we can prove \( I_4 = o^*(1) \).

So far we have proved that \( \int_{[0,1]^p} |\tilde{r}_n(t, z) - r(t, z)| \, dz = o^*(1) \) and \( \int_{[0,1]^p} |\tilde{r}_{2n}(t, z) - r_2(t, z)| \, dz = o^*(1) \). Suppose \( Z_0(t) \) is a \( p \)-dimensional random vector that has a bounded density \( f_0(z) \) on \([0, 1]^p\) and \( Z_0(t) \) is independent of training data \( \mathcal{L}_n \), then \( E\{|\tilde{r}_{2n}(t, Z_0(t)) - r_2(t, Z_0(t))| \mid \mathcal{L}_n\} = o^*(1) \) and \( E\{|\tilde{r}_n(t, Z_0(t)) - r(t, Z_0(t))| \mid \mathcal{L}_n\} = o^*(1) \). By Markov’s inequality, for any \( \epsilon > 0 \), it can be shown that \( P\{|\tilde{r}_n(t, Z_0(t)) - r_n(t, Z_0(t))| > \epsilon \mid \mathcal{L}_n\} = o^*(1) \) and \( P\{|\tilde{r}_{2n}(t, Z_0(t)) - r_{2n}(t, Z_0(t))| > \epsilon \mid \mathcal{L}_n\} = o^*(1) \). By standard arguments, for any \( \epsilon > 0 \), we have \( P\left( \left| \frac{\tilde{r}_n(t, Z_0(t)) - r_n(t, Z_0(t))}{\tilde{r}_{2n}(t, Z_0(t)) - r_{2n}(t, Z_0(t))} \right| > \epsilon \mid \mathcal{L}_n\right) = o^*(1) \). Hence Theorem 1 is proved.

**References**

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., et al. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine* **330**, 657–662.
Athey, S., Tibshirani, J., and Wager, S. (2018). Generalized random forests. *The Annals of Statistics* Forthcoming.

Bacchetti, P. and Segal, M. R. (1995). Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS. *Lifetime Data Analysis* **1**, 35–47.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011a). Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Statistical Modelling* **11**, 429–446.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011b). A review of survival trees. *Statistics Surveys* **5**, 44–71.

Breiman, L. (1996). Bagging predictors. *Machine learning* **24**, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine* **8**, 947–961.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health* **20**, 145–157.

Fleming, T., Neaton, J., Goldman, A., DeMets, D., Launer, C., Korvick, J., and Abrams, D. (1995). Insights from monitoring the cpcra didanosine/zalcitabine trial. terry beirn community programs for clinical research on aids. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **10**, S9–S18.

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

Goldman, A. I., Carlin, B. P., Crane, L. R., Launer, C., Korvick, J. A., Deyton, L., and
Abrams, D. I. (1996). Response of cd4 lymphocytes and clinical consequences of treatment using ddi or ddc in patients with advanced hiv infection. *Journal of Acquired Immune Deficiency Syndromes* **11**, 161–169.

Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports* **69**, 1065–1069.

Green Dand Swets, J. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley: New York.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* **247**, 2543–2546.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics* **7**, 355–373.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* **15**, 651–674.

Huang, X., Chen, S., and Soong, S.-J. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics* **54**, 1420–1433.

Ishwaran, H. and Kogalur, U. (2018). Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.6.1.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics* **2**, 841–860.

Jin, H., Lu, Y., Stone, K., and Black, D. M. (2004). Alternative tree-structured survival
analysis based on variance of survival time. *Medical Decision Making* **24**, 670–680.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics* **48**, 411–425.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association* **88**, 457–467.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58**, 657–664.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research* **7**, 983–999.

Molinaro, A. M., Dudoit, S., and Van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis* **90**, 154–177.

Moradian, H., Larocque, D., and Bellavance, F. (2017). $l_1$ L1 splitting rules in survival forests. *Lifetime data analysis* **23**, 671–691.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521–530.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schmid, M., Wright, M. N., and Ziegler, A. (2016). On the use of Harrell’s C for clinical risk prediction via random survival forests. *Expert Systems with Applications* **63**, 450–459.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics* **44**, 35–47.
Steingrimsson, J. A., Diao, L., Molinaro, A. M., and Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in Medicine* **35**, 3595–3612.

Steingrimsson, J. A., Diao, L., and Strawderman, R. L. (2018). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association* Forthcoming.

Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105–1117.

Wallace, M. (2014). Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Statistics in Medicine* **33**, 4790–4804.

Wang, M.-C., Jewell, N. P., and Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Annals of Statistics* **14**, 1597–1605.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* **77**, 1–17.

Zhang, H. (1995). Splitting criteria in survival trees. *Statistical Modelling* **104**, 305–313.

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association* **107**, 331–340.
2) KSC(t) \leq 0.3961^* \\
3) KSC(t) > 0.3961 \\
6) OP(t) \leq 0.000^* \\
7) OP(t) > 0.000^*

**Figure 1**: Survival tree for the survivor population at time $t$ in the AIDS trial. KSC($t$) is the transformed Karnofsky score at $t$ and OP($t$) is the cumulative number of opportunistic infection up to $t$. 
(a) Smoothed hazard estimates over time for the groups represented by the terminal nodes 2, 6 and 7 in Figure 1.

(b) Hazard predictions over time for the hypothetical patients. Holding all other covariates constant, the Karnofsky score for the hypothetical patients are either A) increasing linearly from 60 to 90, B) decreasing linearly from 80 to 40, and C) constant at 80.

**Figure 2**: Mortality risks for the survivor population over time presented in the AIDS trial.
Table 1: Summaries of integrated absolute errors ($\times 1,000$). The numbers 0%, 25% and 50% correspond to different censoring rates; CON and $\Delta$CON$_r$ are the proposed methods with CON and $\Delta$CON$_r$ as the splitting criterion, respectively; \texttt{rpart} is the recursive regression survival tree implemented in R package \texttt{rpart}; \texttt{ctree} is the conditional inference survival tree implemented in R package \texttt{party}.

| Proposed Methods | CON | $\Delta$CON$_r$ | \texttt{rpart} | \texttt{ctree} |
|------------------|-----|----------------|---------------|---------------|
| $n$              | 0%  | 25% | 50% | 0%  | 25% | 50% | 0%  | 25% | 50% | 0%  | 25% | 50% |
| Scenarios with baseline covariates |
| 100              |     |     |     |     |     |     |     |     |     |     |     |     |
| I                | 122 | 126 | 126 | 122 | 126 | 125 | 118 | 127 | 130 | 115 | 124 | 125 |
| II               | 91  | 99  | 108 | 90  | 96  | 107 | 96  | 105 | 114 | 96  | 103 | 111 |
| III              | 89  | 107 | 112 | 90  | 107 | 112 | 87  | 102 | 110 | 89  | 101 | 106 |
| IV               | 100 | 117 | 120 | 99  | 115 | 118 | 90  | 111 | 124 | 95  | 118 | 133 |
| V                | 77  | 94  | 108 | 77  | 93  | 107 | 97  | 115 | 132 | 97  | 113 | 127 |
| 200              |     |     |     |     |     |     |     |     |     |     |     |     |
| I                | 110 | 112 | 119 | 120 | 120 | 119 | 113 | 121 | 122 | 111 | 119 | 120 |
| II               | 78  | 83  | 88  | 76  | 80  | 86  | 79  | 86  | 94  | 78  | 84  | 91  |
| III              | 71  | 91  | 105 | 72  | 90  | 106 | 71  | 86  | 95  | 80  | 93  | 99  |
| IV               | 74  | 94  | 108 | 73  | 100 | 103 | 77  | 95  | 107 | 75  | 96  | 112 |
| V                | 67  | 80  | 93  | 66  | 78  | 93  | 91  | 109 | 118 | 89  | 106 | 114 |
| Scenarios with a time-dependent covariate |
| 100              |     |     |     |     |     |     |     |     |     |     |     |     |
| VI               | 71  | 86  | 99  | 70  | 85  | 98  | 138 | 159 | 178 | 136 | 156 | 174 |
| VII              | 81  | 92  | 110 | 80  | 90  | 108 | 149 | 200 | 238 | 150 | 198 | 234 |
| VIII             | 87  | 89  | 92  | 86  | 90  | 93  | 388 | 309 | 212 | 402 | 352 | 259 |
| 200              |     |     |     |     |     |     |     |     |     |     |     |     |
| VI               | 54  | 59  | 73  | 55  | 58  | 79  | 128 | 148 | 167 | 126 | 144 | 163 |
| VII              | 73  | 78  | 85  | 74  | 77  | 84  | 145 | 201 | 233 | 150 | 202 | 228 |
| VIII             | 79  | 78  | 80  | 78  | 78  | 80  | 402 | 329 | 220 | 447 | 396 | 274 |
Table 2: Summaries of integrated absolute errors ($\times 1,000$). The numbers 0%, 25% and 50% correspond to different censoring rates; CON and $\Delta$CON$_\tau$ are the proposed methods with CON and $\Delta$CON$_\tau$ as the splitting criterion, respectively.

| Proposed Methods | ranger | rfsr
|------------------|--------|--------|
| CON | $\Delta$CON$_\tau$ | CON | $\Delta$CON$_\tau$ | CON | $\Delta$CON$_\tau$ | CON | $\Delta$CON$_\tau$ |
| $n$ | Sce | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% |
|---|---|---|---|---|---|---|---|---|---|
| 100 | I | 120 | 124 | 124 | 120 | 124 | 124 | 119 | 123 | 126 | 71 | 75 | 79 | 151 | 156 | 157 | 122 | 131 | 136 |
| | II | 65 | 67 | 69 | 65 | 67 | 69 | 136 | 139 | 125 | 99 | 109 | 108 | 139 | 143 | 142 | 84 | 94 | 96 |
| | III | 79 | 86 | 93 | 80 | 88 | 95 | 129 | 136 | 126 | 95 | 106 | 109 | 138 | 141 | 138 | 81 | 91 | 95 |
| | IV | 68 | 78 | 86 | 69 | 78 | 87 | 120 | 128 | 121 | 72 | 90 | 99 | 128 | 132 | 131 | 80 | 97 | 109 |
| | V | 64 | 71 | 76 | 64 | 71 | 76 | 122 | 132 | 137 | 116 | 125 | 128 | 150 | 155 | 156 | 118 | 129 | 134 |
| 200 | I | 112 | 115 | 116 | 112 | 115 | 116 | 119 | 123 | 126 | 64 | 67 | 72 | 117 | 119 | 122 | 67 | 71 | 77 |
| | II | 50 | 52 | 54 | 50 | 52 | 54 | 136 | 139 | 125 | 81 | 90 | 90 | 134 | 139 | 137 | 73 | 83 | 86 |
| | III | 65 | 71 | 78 | 67 | 73 | 79 | 129 | 136 | 126 | 77 | 85 | 88 | 134 | 137 | 133 | 66 | 75 | 78 |
| | IV | 54 | 62 | 71 | 54 | 63 | 71 | 120 | 128 | 121 | 67 | 83 | 92 | 126 | 129 | 127 | 72 | 89 | 101 |
| | V | 48 | 54 | 59 | 49 | 55 | 59 | 122 | 132 | 137 | 116 | 125 | 128 | 150 | 155 | 156 | 118 | 129 | 134 |

Scenarios with a time-dependent covariate

| $n$ | Sce | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% | 0% 25% 50% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | VI | 57 | 64 | 70 | 57 | 64 | 70 | 186 | 201 | 205 | 123 | 138 | 150 | 199 | 215 | 218 | 128 | 143 | 156 |
| | VII | 61 | 71 | 80 | 61 | 71 | 80 | 205 | 238 | 251 | 151 | 189 | 209 | 164 | 203 | 220 | 220 | 249 | 260 |
| | VIII | 68 | 74 | 75 | 69 | 74 | 76 | 408 | 357 | 262 | 206 | 197 | 174 | 240 | 229 | 197 | 405 | 354 | 261 |
| 200 | VI | 47 | 51 | 57 | 47 | 51 | 57 | 189 | 206 | 212 | 118 | 134 | 149 | 204 | 221 | 226 | 120 | 138 | 153 |
| | VII | 48 | 57 | 65 | 48 | 57 | 65 | 209 | 249 | 260 | 148 | 196 | 216 | 225 | 261 | 271 | 153 | 203 | 223 |
| | VIII | 54 | 58 | 59 | 54 | 58 | 59 | 436 | 389 | 286 | 254 | 244 | 206 | 432 | 386 | 285 | 291 | 277 | 229 |