TREE-BASED SURVIVAL MODELS AND PRECISION MEDICINE

Yifan Cui

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2018

Approved by:

Shankar Bhamidi

Jan Hannig

Michael R. Kosorok

Donglin Zeng

Kai Zhang

ABSTRACT

YIFAN CUI: Tree-based survival models and precision medicine
(Under the direction of Dr. Michael R. Kosorok and Dr. Jan Hannig)

Random forests have become one of the most popular machine learning tools in recent years. The main advantage of tree- and forest-based models is their nonparametric nature. My dissertation mainly focuses on a particular type of tree and forest model, in which the outcomes are right censored survival data. Censored survival data are frequently seen in biomedical studies when the true clinical outcome may not be directly observable due to early dropout or other reasons.

We first carry out a comprehensive analysis of survival random forest and tree models and show the consistency of these popular machine learning models by developing a general theoretical framework. Our results significantly improve the current understanding of such models and this is the first consistency result of tree- and forest-based regression estimator for censored outcomes under high-dimensional settings. In particular, the consistency results are derived through analyzing the splitting rules and establishing an adaptive concentration bound of the variance component, which may also shed light on the theoretical analysis of other random forest models.

In the second part, motivated by tree-based survival models, we propose a fiducial approach to provide pointwise and curvewise confidence intervals for the survival functions. On each terminal node, the estimation is essentially a small sample and maybe heavy censoring problem. Most of the asymptotic methods of estimating confidence intervals have coverage problems in many scenarios. The proposed fiducial based pointwise confidence intervals maintain coverage in these situations. Furthermore, the average length of the proposed pointwise confidence intervals is often shorter than the length of competing methods that maintain coverage.

In the third topic, we show one application of tree-based survival models in precision medicine. We extend the outcome weighted learning to right censored survival data without requiring either inverse probability of censoring weighting or semi-parametric modeling of the censoring and failure times. To accomplish this, we take advantage of the tree based approach to nonparametrically

impute the survival time in two different ways. We also illustrate the proposed method on a phase III clinical trial of non-small cell lung cancer.

I dedicate this dissertation work to my parents,

Chaojie Cui and Hong Ruan,

and my grandparents,

Zhentian Ruan and Shihua Zhang,

who have loved and supported me throughout my life.

ACKNOWLEDGEMENTS

I am very grateful to all of those who contributed in some way to the work in this thesis. First and foremost, my deepest gratitude is to my advisors, Dr. Michael R. Kosorok and Dr. Jan Hannig, for their enthusiasm, guidance, encouragement, and support. Their mentoring has been both inspirational and cheered me during my venture. I have been extremely lucky to have two advisors who cared so much about my work, and lent invaluable assistance to my study and research. It is much more than statistics that I learned from them. I would not have been able to achieve this accomplishment without them. I would also like to thank Dr. Kosorok for providing me the opportunity to visit the University of Cambridge to conduct research on adaptive clinical trials.

I would also like to thank my committee members Professor Shankar Bhamidi, Donglin Zeng, and Kai Zhang for reading the manuscript and offering insightful comments which have led to significant improvements of the thesis.

I would like to thank Dr. Ruoqing Zhu, Dr. Mai Zhou, and Dr. Stefan Wager. My collaboration with them has been a very enjoyable part of my graduate study.

I am also thankful to all other faulty members, students and staff in the Department of Statistics and Operations Research and Department of Biostatistics.

My sincere thanks also go to Dr. James Hung, Dr. Peiling Yang, and Dr. Semhar Ogbagaber, for offering me the summer research opportunities at the Food and Drug Administration and leading me to work on sequential parallel comparison designs.

Last but not the least, I would like to thank my parents. They have helped me throughout the process of this dissertation by giving encouragement and providing emotional support. They are the most important people in my world and I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

In this chapter, we outline the contributions in the subsequent development of the thesis.

1.1 Some asymptotic results for survival tree and forest models

In Chapter 2, we develop a theoretical framework and asymptotic results for survival tree and forest models under right censoring. We first investigate the method from the aspect of splitting rules, where the survival curves of the two potential child nodes are calculated and compared. We show that existing approaches lead to a potentially biased estimation of the within-node survival and cause non-optimal selection of the splitting rules. This bias is due to the censoring distribution and the non i.i.d. sample structure within each node. Based on this observation, we develop an adaptive concentration bound result for both tree and forest versions of the survival tree models. The result quantifies the variance component for survival forest models. Furthermore, we show with two specific examples how these concentration bounds, combined with properly designed splitting rules, yield consistency results. The two examples are: 1) a finite dimensional setting with random splitting rules; and 2) an infinite dimensional case with marginal signal checking. The development of these results serves as a general framework for showing the consistency of tree- and forest-based survival models.

1.2 Nonparametric generalized fiducial inference for survival functions under censoring

Fiducial Inference, introduced by Fisher in the 1930s, has a long history, which at times aroused passionate disagreements. However, its application has been largely confined to relatively simple parametric problems. In Chapter 3, we present what might be the first time fiducial inference, as generalized by Hannig et al. [64], is systematically applied to estimation of a nonparametric

survival function under right censoring. We find that the resulting fiducial distribution gives rise to surprisingly good statistical procedures applicable to both one sample and two sample problems. In particular, we use the fiducial distribution of a survival function to construct pointwise and curvewise confidence intervals for the survival function, and propose tests based on the curvewise confidence interval. We establish a functional Bernstein-von Mises theorem, and perform thorough simulation studies in various scenarios with different levels of censoring. The proposed fiducial based confidence intervals maintain coverage in situations where asymptotic methods often have substantial coverage problems. Furthermore, the average length of the proposed confidence intervals is often shorter than the length of competing methods that maintain coverage. Finally, the proposed fiducial test is more powerful than various types of log-rank tests and sup log-rank tests in some scenarios. We illustrate the proposed fiducial test comparing chemotherapy against chemotherapy combined with radiotherapy using data from the treatment of locally unresectable gastric cancer.

1.3 Tree based weighted learning for estimating ITRs with censored data

Estimating individualized treatment rules is a central task for personalized medicine. [135] and [133] proposed outcome weighted learning to estimate individualized treatment rules directly through maximizing the expected outcome without modeling the response directly. In Chapter 4, we extend the outcome weighted learning to right censored survival data without requiring either inverse probability of censoring weighting or semiparametric modeling of the censoring and failure times as done in [138]. To accomplish this, we take advantage of the tree based approach proposed in [141] to nonparametrically impute the survival time in two different ways. The first approach replaces the reward of each individual by the expected survival time, while in the second approach only the censored observations are imputed by their conditional expected failure times. We establish consistency and convergence rates for both estimators. In simulation studies, our estimators demonstrate improved performance compared to existing methods. We also illustrate the proposed method on a phase III clinical trial of non-small cell lung cancer.

CHAPTER 2

Some asymptotic results for survival tree and forest models

2.1 Introduction

Random forests [19] have become one of the most popular machine learning tools in recent years. Many extensions of random forests [106, 73, 26, 93] have seen tremendous success in statistical and biomedical related fields [85, 105, 14, 60, 100, 114, 71] in addition to many applications to artificial intelligence and machine learning problems.

The main advantage of tree- [21] and forest-based models is their nonparametric nature. However, the theoretical properties have not been fully understood yet to date, even in the regression settings, although there has been a surge of research on understanding the statistical properties of random forests in classification and regression. [83] is one of the early attempts to connect random forests to nearest neighbor predictors. Later on, a series of work including [12, 11, 54] and [92] established theoretical results on simplified tree-building processes or specific aspects of the model. More recently, [142] established consistency results based on an improved splitting rule criteria; [125] analyzed the confidence intervals induced from a random forest model; [84] established connections with Bayesian variable selection in the high dimensional setting; [110] showed consistency of the original random forests model on an additive structure; and [126] studied the variance component of random forests and established corresponding concentration inequalities. For a more comprehensive review of related topics, we refer to [13].

In this chapter, we focus on the theoretical properties of a particular type of tree- and forest-model, in which the outcomes are right censored survival data [50]. Censored survival data are frequently seen in biomedical studies when the true clinical outcome may not be directly observable due to early dropout or other reasons. Random forest based survival models have been developed to handle censored outcomes, including [70, 69, 73, 141, 115] and many others. However, there are few established theoretical results despite the popularity of these methods in practice, especially

in genetic and clinical studies. For a general review of related topics, including single-tree based survival models , we refer to [16]. To the best of our knowledge, the only consistency result to date is given in [72] who considered the setting where all predictors are categorical. Hence, in this chapter, we attempt to lay out a theoretical framework for tree- and forest-based survival models in a general setting, including when the number of dimensions diverges with the sample size. Furthermore, we establish consistency under several specific models. Without the risk of ambiguity, we refer to all considered models as tree-based survival models, while the established results apply to both single-tree and forest versions.

The chapter is organized as follows: in Section 2.2, we introduce tree-based survival models and some basic notations. Section 2.3 is devoted to demonstrating a fundamental property of the survival tree model associated with splitting rule selection and terminal node estimation. A concentration inequality of the Nelson-Aalen [1] estimator based on non-identically distributed samples is established. Utilizing this result, we derive adaptive concentration bounds for tree-based survival models in Section 2.4. Furthermore, in Section 2.5, we establish consistency and a variance bound for two particular choices of splitting rules, one of which are infinite dimensional cases, and one of which is finite dimensional. Details of proofs are given in the appendices, and a summary of notation is given before the appendices for convenience.

2.2 Tree-based survival models

The essential ingredient of tree-based survival models is recursive partitioning. A d-dimensional feature space $\mathcal{X}$ is partitioned into terminal nodes, or more precisely, mutually exclusive and exhaustive subsets. We denote $\boldsymbol{A} = \{\mathcal{A}_u\}_{u\in\mathcal{U}}$ to be the collection of these terminal nodes returned by fitting a single tree, where $\mathcal{U}$ is a set of indices, and hence $\mathcal{X} = \bigcup_{u\in\mathcal{U}} \mathcal{A}_u$ and $\mathcal{A}_u \cap \mathcal{A}_l = \emptyset$ for any $u \neq l$. We also call $\boldsymbol{A}$ a partition of the feature space $\mathcal{X}$. In a traditional tree-building process [21], where binary splitting rules are used, all terminal node are (hyper)rectangles, i.e., $\mathcal{A} = \bigotimes_{j=1}^{d}(a_j, b_j]$. Other possibilities can also be considered. For example, linear combination splits [93, 77, 142] may result in more complex structures of terminal nodes. However, regardless of the construction of the trees, the terminal node estimates are obtained by treating the within-node observations as

identically distributed. Before giving a general algorithm of tree-based survival models, we first introduce some notation.

Following the standard notation in the survival analysis literature, let $\{X_i, Y_i, \delta_i\}_{i=1}^n$ be a set of n i.i.d. copies of the covariates, observed survival time, and censoring indicator, where the observed survival time $Y_i = \min(T_i, C_i)$, and $\delta_i = \mathbb{1}(T_i \leq C_i)$. We assume that each T_i follows a conditional distribution $F_i(t) = \text{pr}(T_i \leq t \mid X_i)$, where the survival function is denoted $S_i(t) = 1 - F_i(t)$, the cumulative hazard function $\Lambda_i(t) = -\log\{S_i(t)\}$, and the hazard function $\lambda_i(t) = \text{d}\Lambda_i(t)/\text{d}t$. The censoring time C_i's is assumed to follow the conditional distribution $G_i(t) = \text{pr}(C_i \leq t \mid X_i)$, where a non-informative censoring mechanism, $T_i \perp C_i \mid X_i$, is assumed.

In any tree-based survival model, terminal node estimation is a crucial part. For any node $\mathcal{A}_u$, this can be obtained through the Kaplan-Meier [74] estimator for the survival function or the Nelson-Aalen [99, 1] estimator of the cumulative hazard function based on the within-node data. Our focus in this chapter is on the following Nelson-Aalen estimator

$$\widehat{\Lambda}_{\mathcal{A}_u}(t) = \sum_{s \leq t} \frac{\sum_{i=1}^n \mathbb{1}(\delta_i = 1)\mathbb{1}(Y_i = s)\mathbb{1}(X_i \in \mathcal{A}_u)}{\sum_{i=1}^n \mathbb{1}(Y_i \geq s)\mathbb{1}(X_i \in \mathcal{A}_u)}, \tag{2.1}$$

and the associated Nelson-Altshuler estimator [4] for the survival function when needed:

$$\widehat{S}_{\mathcal{A}_u}(t) = \exp\big\{ -\widehat{\Lambda}_{\mathcal{A}_u}(t)\big\}.$$

Hence a single tree model can be expressed by a collection of doublets $\{\mathcal{A}_u, \widehat{\Lambda}_{\mathcal{A}_u}\}_{u \in \mathcal{U}}$. In an ensemble survival tree method [73, 141], a set of B trees are fitted to the data. In practice, $B = 1000$ is used in the popular `R` package `randomForestSRC` as the default value. Hence the forest, or a collection of partitions, $\{\{\mathcal{A}_u^b, \widehat{\Lambda}_{\mathcal{A}_u^b}\}_{u \in \mathcal{U}_b}\}_{b=1}^B$ indexed by b is constructed. These trees are constructed with a bootstrap sampling mechanism or with the entire training data, in addition to a variety of types of randomness injected [55] to the vanilla random forests [19]. To facilitate later arguments, we conclude this section by providing a high-level outline (Algorithm 1) for fitting a survival forest model. Many of the details of the splitting rule component are deferred to later sections.

Algorithm 1: Pseudo algorithm for tree-based survival models

Input: Training dataset $\mathcal{D}_n$, terminal node size k, number of trees B;
Output: $\{\{\mathcal{A}_u^b, \widehat{\Lambda}_{\mathcal{A}_u^b}\}_{u\in\mathcal{K}_b}\}_{b=1}^B$

1 **for** $b = 1$ ***to*** B **do**
2 Initiate $\mathcal{A} = \mathcal{X}$, a bootstrap sample $\mathcal{D}_n^b$ of $\mathcal{D}_n$, $\mathcal{K}_b = \emptyset$, $u = 1$;
3 At a node $\mathcal{A}$, if $\sum_{X_i\in\mathcal{D}_n^b} \mathbb{1}(X_i \in \mathcal{A}) < k$, proceed to Line 5. Otherwise, construct a splitting rule such that $\mathcal{A} = \mathcal{A}_{\text{left}} \cup \mathcal{A}_{\text{right}}$, where $\mathcal{A}_{\text{left}} \cap \mathcal{A}_{\text{right}} = \emptyset$. ;
4 Send the two child nodes $\mathcal{A}_{\text{left}}$ and $\mathcal{A}_{\text{right}}$ to Line 3 separately;
5 Conclude the current node $\mathcal{A}$ as a terminal node $\mathcal{A}_u^b$, calculate $\widehat{\Lambda}_{\mathcal{A}_u^b}$ using the within-node data, and update $\mathcal{K}_b = \mathcal{K}_b \cup \{u\}$ and $u = u + 1$;
6 **end**
7 **return** $\{\{\mathcal{A}_u^b, \widehat{\Lambda}_{\mathcal{A}_u^b}\}_{u\in\mathcal{K}_b}\}_{b=1}^B$

2.3 The splitting rule and its biasedness

One central idea throughout the survival tree and forest literature is to construct certain goodness-of-fit statistics that evaluate the impurity reduction across many candidate splitting rules. The best splitting rule is then selected and implemented to partition the node. This essentially resembles the idea in a regression tree setting where the mean differences or equivalently the variance reduction is used as the criterion. The most popular criteria in survival tree models is constructed through the log-rank statistic [59, 28, 81, 43, 73, 141] and other nonparametric comparisons of two curves, such as the Kolmogorov-Smirnov, Wilcoxon-Gehan and Tarone-Ware type of statistics [29, 111]. Other ideas include likelihood or model based approaches [27, 36, 87, 2, 120, 43], inverse probability of censoring weighting (IPCW) [94, 69], and non-standard criteria such as [134] and [Kręłowska]. [16] provides a comprehensive review of the methodological developments of survival tree models.

The literature on the theoretical analysis of survival tree based methods seems to be somewhat sparse. One of the more recent general results, as mentioned in the introduction, is [72], who established uniform consistency of random survival forests [73] by assuming a discrete feature space as can happen, for example, when the covariates are genotyping data. The idea can be extended to many other specific survival tree models, however, the discrete distribution assumption of the feature space is not satisfied in most applications. The major difficulty of the theoretical developments in a general setting is the highly complex nature of the splitting rules and their interference with the entire tree structure.

2.3.1 Within-node estimation

To begin our analysis, we start by investigating the Kaplan-Meier (KM) and the Nelson-Altshuler (NA) estimators of the survival function. There are two main reasons that we revisit these classical methods: first, these methods are wildly used for terminal node estimation in fitted survival trees. Hence, the consistency of any survival tree model inevitably relies on their asymptotic behavior; second, the most popular splitting rules, such as the log-rank, Wilcoxon-Gehan and Tarone-Ware statistics, are all essentially comparing the KM curves across the two potential child nodes, which again plays an important role in the consistency results. We note that although other splitting criteria exist, our theoretical framework can be extended to address their particular properties. Without making restrictive distributional assumptions on the underlying model, our results shows that the currently implemented splitting rules, not surprisingly, are affected by the underlying censoring distribution, and are essentially biased, in the sense that they may not select the most important variable to split on asymptotically. Furthermore, we exactly quantify this biased estimator by developing a concentration bound around its true mean.

Noticing that the KM and the NA estimators are the two most commonly used estimators, the following Lemma bounds their difference through an exact inequality regardless of the underlying data distribution. The proof follows mostly from [34], and is given in the Appendices.

Lemma 1. *Let $\widehat{S}_{KM}(t)$ and $\widehat{S}_{NA}(t)$ be the Kaplan-Meier and the Nelson-Altshuler estimators, respectively, obtained using the same set of samples $\{Y_i, \delta_i\}_{i=1}^n$. Then we have,*

$$|\widehat{S}_{KM}(t) - \widehat{S}_{NA}(t)| < \widehat{S}_{KM}(t)\frac{4}{\sum_{i=1}^n \mathbb{1}(Y_i \geq t)},$$

for any observed failure time point t such that $\widehat{S}_{KM}(t) > 0$.

The above result suggests that calculating the difference between two KM curves is asymptotically the same as using the NA estimator as long as we only calculate the curve up to a time point where the sample size is sufficiently large. For this purpose, we make the following assumption throughout the chapter to guarantee with large probability that $\widehat{S}_{NA}(t) = \widehat{S}_{KM}(t) + O(1/n)$ across all terminal nodes:

Assumption 1. *There exists fixed positive constants $\tau < \infty$ and $M \in (0, 1)$, such that*

$$\text{pr}(Y_i \geq \tau \mid X_i) \geq M,$$

uniformly for all $X_i \in \mathcal{X}$.

Note that similar assumptions are commonly used in the survival analysis literature, for examples, $\text{pr}(T \geq \tau) > 0$ in [50], and $\text{pr}(C = \tau) > 0$ in [95]. The above assumption is a straightforward extension due to the partitioning nature of tree models.

2.3.2 A motivating example

Noting that the splitting rule selection process essentially compares the survival curves computed from two child nodes, we take a closer look at this process. In fact, most studies of the large sample property of the KM estimator assume that the observations are i.i.d. [22, 56], or at least one set of the failure times or censoring times are i.i.d. [139]. However, this is almost always not true for tree-based methods at any internal node because both T_i's and C_i's typically depend on the covariates. The question is whether this affects the selection of the splitting rule. A simulation study can be utilized to better demonstrate this issue.

Consider the split at a particular node. We generate three random variables: $X^{(1)}$, $X^{(2)}$ and $X^{(3)}$ from a multivariate normal distribution with mean 0 and variance Σ, where the diagonal elements of Σ are all 1, and the only nonzero off diagonal element is $\Sigma_{12} = \Sigma_{21} = 0.8$. The failure distribution of T is exponential with mean $\exp(1.25 \cdot X^{(1)} + X^{(3)} - 2)$. We consider two censoring distributions for C: the first one is an exponential distribution with mean 1 for all subjects, i.e., they are identically distributed; and the second one is an exponential distribution with mean equal to $\exp(3 \cdot X^{(2)})$. The splitting rule is searched for by maximizing the log-rank test statistics between the two potential child nodes $\{X^{(j)} \leq c, X \in \mathcal{A}\}$ and $\{X^{(j)} > c, X \in \mathcal{A}\}$, and the cutting point c is searched for throughout the entire range of the variable. In an ideal situation, one would expect the best splitting rule to be constructed using $X^{(1)}$ with large probability, since it carries the most signal. This is indeed the case as shown in the first row of Table 2.1 for the i.i.d. censoring case, but not so much for the dependent censoring case. The simulation is done with $n = 1000$ and repeated 1000 times. While this only demonstrates the splitting process on a single node, the consequence of

this on the consistency of the entire tree is much more involved since the entire tree structure can be altered by the censoring distribution. It is difficult to draw a definite conclusion at this point, but the impact of the censoring distribution is clearly demonstrated.

Table 2.1: Probability of selecting the splitting variable.

Censoring distribution	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
G_i identical	0.950	0.004	0.046
G_i depends on $X_i^{(2)}$	0.256	0.028	0.716

2.3.3 Survival estimation based on independent but non-identically distributed observations

It now seems impossible to analyze the consistency without exactly quantifying the within node estimation performance. We look at two different quantities corresponding to the two scenarios used above. The first one is an averaged cumulative hazard function within any node $\mathcal{A}$:

$$\Lambda_{\mathcal{A}}(t) = \frac{1}{\mu(\mathcal{A})} \int_{x \in \mathcal{A}} \Lambda(t \mid x) \mathrm{dP}(x), \tag{2.2}$$

where P is the distribution of X, and $\mu(\mathcal{A}) = \int_{x \in \mathcal{A}} \mathrm{dP}(x)$ is the measure of node $\mathcal{A}$. Clearly, since in the first case, the censoring distribution is not covariate dependent, we are asymptotically comparing $\Lambda_{\mathcal{A}}(t)$ on the two child nodes, which results in the selection of the first variable. This should also be considered as a rational choice since $X^{(1)}$ contains more signal at the current node.

In the second scenario, i.e., the dependent censoring case, the within-node estimator $\widehat{\Lambda}_{\mathcal{A}}(t)$ does not converge to the $\Lambda_{\mathcal{A}}(t)$ in general, which can be inferred from the following theorem. As the main result of this section, Theorem 1 is interesting by its own right for understanding tree-based survival models, since it establishes a bound of the survival estimation under independent but non-identically distributed samples, which is a more general result than [139]. It exactly quantifies the estimation performance for each potential child node, hence is also crucial for understanding splitting rules generally. This theorem can be found in an unpublished technical report by Mai Zhou at the University of Kentucky.

Theorem 1. *Let $\widehat{\Lambda}(t)$ be the Nelson-Aalen estimator of the cumulative hazard function from a set of n independent samples $\{Y_i, \delta_i\}_{i=1}^n$ subject to right censoring, where the failure and censoring distributions (not necessarily identical) are given by F_i's and G_i's. Under Assumption 1, we have for $n \geq 288/(\epsilon_1^2 M^4)$,*

$$\mathrm{pr}\Big(\sup_{t<\tau}\Big|\widehat{\Lambda}(t) - \Lambda_n^*(t)\Big| > \epsilon_1\Big) < 16(n+2)\exp\Big\{\frac{-nM^4\epsilon_1^2}{288}\Big\}, \tag{2.3}$$

where

$$\Lambda_n^*(t) = \int_0^t \frac{\sum[1-G_i(s)]\mathrm{d}F_i(s)}{\sum[1-G_i(s)][1-F_i(s)]}. \tag{2.4}$$

The proof is deferred to Appendix. Based on Theorem 1, if we restrict ourselves to any node $\mathcal{A}$, the difference between the within-node estimator $\widehat{\Lambda}_{\mathcal{A}}(t)$ and

$$\Lambda_{\mathcal{A},n}^*(t) = \int_0^t \frac{\sum_{X_i\in\mathcal{A}}[1-G_i(s)]\mathrm{d}F_i(s)}{\sum_{X_i\in\mathcal{A}}[1-G_i(s)][1-F_i(s)]} \tag{2.5}$$

is bounded above, where $\Lambda_{\mathcal{A},n}^*(t)$ is some version of the underlying true cumulative hazard contaminated by the censoring distribution. Noting that $\Lambda_{\mathcal{A},n}^*(t)$ also depends on the sampling points X_i's, we further develop Lemma 14 in the Appendix to verify that $\Lambda_{\mathcal{A},n}^*(t)$ and its expected version $\Lambda_{\mathcal{A}}^*(t)$ are close enough, where

$$\Lambda_{\mathcal{A}}^*(t) = \int_0^t \frac{E_{X\in\mathcal{A}}[1-G(s\mid X)]\mathrm{d}F(s\mid X)}{E_{X\in\mathcal{A}}[1-G(s\mid X)][1-F(s\mid X)]}. \tag{2.6}$$

It is easy to verify that the difference between $\Lambda_{\mathcal{A},n}^*(t)$ and $\Lambda_{\mathcal{A}}(t)$ will vanish if the F_i's are identical within a node $\mathcal{A}$ (a sufficient condition):

$$\begin{aligned}\Lambda_{\mathcal{A},n}^*(t) &= \int_0^t \frac{\sum_{X_i\in\mathcal{A}}[1-G_i(s)]}{\sum_{X_i\in\mathcal{A}}[1-G_i(s)]}\frac{\mathrm{d}F(s)}{1-F(s)} \\ &\qquad\qquad\qquad\qquad (\text{if } F_i\equiv F \text{ for all } X_i\in\mathcal{A}) \\ &= \int_0^t \frac{\mathrm{d}F(s)}{1-F(s)} = \frac{1}{\mu(\mathcal{A})}\int_{x\in\mathcal{A}}\int_0^t \frac{\mathrm{d}F(s)}{1-F(s)}\mathrm{dP}(x) = \Lambda_{\mathcal{A}}(t).\end{aligned} \tag{2.7}$$

As we demonstrated in the simulation study above, comparing $\widehat{\Lambda}_{\mathcal{A}}(t)$ between the two child nodes may lead to a systematically different selection of splitting variables than using $\Lambda_{\mathcal{A}}(t)$ which can't be known a priori. The main cause of the differences between these two quantities is that the NA estimator treats each node as a homogeneous group, which is typically not true. Another simple interpretation is that although the conditional independence assumption $T \perp C \mid X$ is satisfied, we have instead at each internal node that

$$T \not\perp C \mid \mathbb{1}(X^{(j)} < c)$$

is almost always true for any j and c, causing a nonidentifiability problem.

Exactly quantifying the statistical behavior of each internal node in the entire survival tree or forest is difficult due to the fact that some subtle changes in the censoring distribution G may completely alter the entire tree structure. Of course, such a difficulty only arises when the splitting rule is highly data dependent as happens, e.g., with the log-rank test statistic. When the splitting rule is independent of the observed data, the analysis becomes much easier. We will provide the results under this random splitting rule setting in Section 2.5. An analog of this result for the uncensored regression and classification settings was proposed by [20], and further analyzed by [83, 11, 6] and many others. Another situation where consistency can be derived is when the splitting rules find almost always the correct variable to split. To look closer at this setting, we consider two high dimensional cases in Section 2.5, and show that the marginal screening type of splitting rules will lead to consistency. To establish these results, we use the variance-bias breakdown, and start by analyzing the variance component of a survival tree estimator in the next section.

Remark 2.3.1. Another kind of inconsistency can be caused by non-marginal underlying failure models. In the regression setting, this is well documented through, for example, the "checker-board" structure used in [88], [11] and [142]. It is easy to see that the failure distribution in a survival model can be chosen similarly to cause inconsistency under the regular marginal splitting rule. However, the mechanism of their causes is fundamentally different from the issue that we described above which is solely due to the underlying censoring distribution.

2.4 Adaptive concentration bounds of survival trees and forests

In this section, we focus on quantifying the survival tree model from a new angle, namely, the adaptive concentration [126] of each terminal node estimator to the true within-node expectation. In the sense of the variance-bias breakdown, this section is to quantify a version of the variance component of a tree-based model estimator. To be precise, with large probability, our main results bound $\left|\widehat{\Lambda}_{\mathcal{A}}(t)-\Lambda^*_{\mathcal{A},n}(t)\right|$ across all potentially possible terminal nodes $\mathcal{A}$ in a fitted tree or forest. The adaptiveness comes from the fact that the target of the concentration is the censoring contaminated version $\Lambda^*_{\mathcal{A},n}(t)$, which is adaptively defined for each node $\mathcal{A}$ with the observed samples, rather than as a fixed universal value.

The results in this section have many implications. This bound is essentially the variance part in a bias-variance break down of an estimator, and is satisfied regardless of the splitting rule selection. Hence, we can then analyze the bias part to show the consistency of a survival tree model. Furthermore, following our framework, the consistency results for any survival tree model can simply be established by checking several conditions on the splitting rules. Although this may still pose challenges in certain situations, our unified framework is largely applicable to most existing methods. Some additional definitions and notations are needed as we proceed.

2.4.1 Additional definitions

Following our previous assumptions on the underlying data generating model, we observe a set of n i.i.d. samples $\mathcal{D}_n = \{X_i, Y_i, \delta_i\}_{i=1}^n$. We view each tree as a partition of the feature space, denoted $\boldsymbol{A} = \{\mathcal{A}_u\}_{u\in\mathcal{U}}$, where the $\mathcal{A}_u$'s are non-overlapping hyper-rectangular terminal nodes. The following definition of a valid partition, which we owe to [126], is used to restrict the partition $\boldsymbol{A}$ being constructed. Here we state the definition again:

Definition 2.1 Valid tree and forest partitions [126]. A tree partition $\boldsymbol{A}$ is $\{\alpha, k\}$-valid if it satisfies two conditions: 1). For each splitting, the child node contains at least a fraction $\alpha \in (0, 0.5)$ of the training samples in its parent node; and 2). Each terminal node contains at least k training examples. For the training data $\mathcal{D}$, we denote the set of all $\{\alpha, k\}$-valid tree partitions by $\mathcal{V}_{\alpha,k}(\mathcal{D})$. In addition, define the collection $\{\boldsymbol{A}^{(b)}\}_{b=1}^B$ as a valid forest partition if all of the B partitions $\boldsymbol{A}^{(b)}$'s are valid. Then the set of all such valid forest partitions is denoted as $\mathcal{H}_{\alpha,k}(\mathcal{D})$.

The following definition is essentially the tree model estimator of the cumulative hazard function obtained form a partition $\boldsymbol{A}$. When $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D})$, we will call the induced estimator a *valid survival tree*. The regression version of their definitions can be found in [126].

Definition 2.2 Valid survival tree. Given the observed data $\mathcal{D}_n$, a valid survival tree estimator of the cumulative hazard function is induced by a valid partition $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$ with $\boldsymbol{A} = \{\mathcal{A}_u\}_{u \in \mathcal{U}}$:

$$\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) = \sum_{u \in \mathcal{U}} \mathbb{1}(x \in \mathcal{A}_u) \widehat{\Lambda}_{\mathcal{A}_u}(t), \tag{2.8}$$

where each $\widehat{\Lambda}_{\mathcal{A}_u}(t \mid x)$ is defined in Equation (2.1).

When an ensemble of trees are fitted, we define a *valid survival forest*:

Definition 2.3 Valid survival forest. A valid survival forest $\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}$ is defined as the average of B valid survival trees induced by a collection of valid partitions $\{\boldsymbol{A}_{(b)}\}_1^B \in \mathcal{H}_{\alpha,k}(\mathcal{D}_n)$:

$$\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\Lambda}_{\boldsymbol{A}_{(b)}}(t \mid x). \tag{2.9}$$

In the following, we define the censoring contaminated survival tree and forest, which are the asymptotic versions of the corresponding within-node average estimators of the cumulative hazard function. Note that by Theorem 1, these averages are censoring contaminated versions $\Lambda^*_{\mathcal{A},n}(t)$, but not the true averages $\Lambda_{\mathcal{A}}(t)$.

Definition 2.4 Censoring contaminated survival tree and forest. Given the observed data $\mathcal{D}_n$ and $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$, the corresponding censoring contaminated survival tree is defined as

$$\Lambda^*_{\boldsymbol{A},n}(t \mid x) = \sum_{u \in \mathcal{U}} \mathbb{1}(x \in \mathcal{A}_u) \Lambda^*_{\mathcal{A}_u,n}(t), \tag{2.10}$$

where each $\Lambda^*_{\mathcal{A}_u,n}(t)$ is defined by Equation (2.5). Furthermore, let $\{\boldsymbol{A}_{(b)}\}_1^B \in \mathcal{H}_{\alpha,k}(\mathcal{D}_n)$. Then the censoring contaminated survival forest is given by

$$\Lambda^*_{\{\boldsymbol{A}_{(b)}\}_1^B,n}(t \mid x) = \frac{1}{B} \sum_{b=1}^{B} \Lambda^*_{\boldsymbol{A}_{(b)},n}(t \mid x). \tag{2.11}$$

2.4.2 Main result

In order to obtain the adaptive concentration bound for survival trees, we need to bound

$$\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda^*_{\boldsymbol{A},n}(t \mid x)$$

for all valid partitions $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$. We first specify several regularity assumptions. The first assumption is a bound on the dependence of the individual features. Note that in the literature, uniform distributions are often assumed [12, 11] on the covariates, which implies independence. To allow dependency among covariates, we assume the following, which has also been considered in [126].

Assumption 2. *Covariates $X \in [0,1]^d$ are distributed according to a density $p(\cdot)$ satisfying $1/\zeta \leq p(x) \leq \zeta$ for all x and some $\zeta \geq 1$.*

We also set a restriction on the tuning parameter—the minimum terminal node size k—so that it grows with n and dimension d via the following rate:

Assumption 3. *Assume that k is bounded below so that*

$$\lim_{n\to\infty} \frac{\log(n)\max\{\log(d), \log\log(n)\}}{k} = 0. \tag{2.12}$$

Then we have the adaptive bound for our tree estimator in the following theorem. The proof is presented in Appendix.

Theorem 2. *Suppose the training samples (X_i, Y_i, δ_i) satisfy Assumptions 1 and 2, and the rate of the sequence (n, d, k) satisfies Assumption 3. Then all valid trees concentrate on censoring contaminated tree:*

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \boldsymbol{A}\in\mathcal{V}_{\alpha,k}(\mathcal{D}_n)} \left|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda^*_{\boldsymbol{A},n}(t \mid x)\right| \leq M_1\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

with probability larger than $1 - 2/\sqrt{n}$, for some universal constant M_1.

In addition, in a high dimensional setting, i.e. $\liminf_{n\to\infty}(d/n) > 0$, Theorem 2 can be simplified as follows:

Corollary 1. *Suppose the training samples (X_i, Y_i, δ_i) satisfy Assumptions 1 and 2, and the rate of sequence (n, d, k) satisfies Assumption 3 and $\liminf_{n\to\infty}(d/n) > 0$. Then all valid trees concentrate on censoring contaminated trees:*

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \boldsymbol{A}\in\mathcal{V}_{\alpha,k}(\mathcal{D}_n)} \left|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda^*_{\boldsymbol{A},n}(t \mid x)\right| \leq M_1\sqrt{\frac{\log(n)\log(d)}{k\log((1-\alpha)^{-1})}},$$

with probability larger than $1 - 2/\sqrt{n}$, for some universal constant M_1.

Remark 2.4.1. In a moderately high dimensional setting, i.e. $d \sim n$, the rate is $\log(n)/k^{1/2}$. In an ultra high dimensional setting, for example, $\log(d) \sim n^{\vartheta}$, where $0 < \vartheta < 1$, the rate is close to $n^{\vartheta}/k^{1/2}$. The rate that k grows with n cannot be too slow in order to achieve the bound in the ultra high dimensional setting. This is somewhat intuitive since if k grows slowly then we are not able to bound all possible nodes defined in 2.1.

The above theorem and corollary hold for all single tree partitions in $\mathcal{V}_{\alpha,k}(\mathcal{D}_n)$. Consequently, we have the following results for the forest estimator. The proof is deferred to Appendix.

Corollary 2. *Suppose Assumptions 1-3 hold. Then all valid forests concentrate on the censoring contaminated forest probability with larger than $1 - 2/\sqrt{n}$,*

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \{\boldsymbol{A}_{(b)}\}_1^B\in\mathcal{H}_{\alpha,k}(\mathcal{D}_n)} \left|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid x) - \Lambda^*_{\{\boldsymbol{A}_{(b)}\}_1^B,n}(t \mid x)\right| \leq M_1\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

for some universal constant M_1. Furthermore, if $\liminf_{n\to\infty}(d/n) \to \infty$,

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \{\boldsymbol{A}_{(b)}\}_1^B\in\mathcal{H}_{\alpha,k}(\mathcal{D}_n)} \left|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid x) - \Lambda^*_{\{\boldsymbol{A}_{(b)}\}_1^B,n}(t \mid x)\right| \leq M_1\sqrt{\frac{\log(n)\log(d)}{k\log((1-\alpha)^{-1})}},$$

with probability larger than $1 - 2/\sqrt{n}$*, for some universal constant* M_1*.*

The results established in this section essentially address the variation component in a fitted random forest. We chose not to use the true within-node population averaged quantity $\Lambda^*_{\mathcal{A}}(t)$ (see Equation 2.6), or its single tree and forest versions as the target of the concentration. This is because such a result would require bounded density function of the failure time T. However, when $f(t)$ is bounded, the results can be easily generalized to $\Lambda^*_{\mathcal{A}}(t)$. Lemma 2 provides an analog of Theorem 1 in this situation.

The next section establishes consistency of several specific models. Intuitively, if a particular splitting rule leads to "nicely behaved" terminal nodes across the entire tree or forest, then consistency results can be derived. For example, for a finite dimensional case, "nicely behaved" terminal nodes essentially require that the diameter of each terminal node shrinks to 0 (in the language of [38]), while in a high-dimensional case, we would require that the diameters of all important variables (see definition in Section 2.5.2 below) shrink to 0.

2.5 Consistency of survival tree and forest models

With the above established concentration inequalities, we are now in a position to discuss consistency results under several scenarios and particular choices of splitting rules. We note that there is no existing splitting rule which can universally handle all underlying models, hence it is more realistic to discuss several different specific scenarios. Of course, the choice of the corresponding splitting rule would then depend on the particular scenario which is not known a priori. However, this is still both theoretically and practically important for understanding the model since there are currently no practical guideline. In addition, the analysis strategy serves as a general framework for showing consistency results for any tree- and forest-based survival model. We consider two specific scenarios: 1) a finite dimensional case where the splitting rule is chosen randomly; and 2) an infinite dimensional case using the difference of Nelson-Aalen estimators as the splitting rule. Throughout this section, to streamline our presentation, we assume that the covariates X is uniformly distributed.

2.5.1 Consistency of survival forest when dimension d is fixed

In this setting, we assume the dimension of the covariates space is fixed and finite. At each internal node we choose the splitting variable randomly. When the splitting variable is chosen, we choose the splitting point at random such that both two child nodes contain at least a proportion α of the samples in the parent node. To prove the consistency of the forest, we need to bound the bias term

$$\sup_{t<\tau} E_X \left| \Lambda^*_{\{\boldsymbol{A}_{(b)}\}_1^B, n}(t \mid X) - \Lambda(t \mid X) \right|,$$

and combine the results with the variance aspect. It should be noted that in Section 2.4, we did not treat the tree- and forest- structures ($\boldsymbol{A}$ and $\{\boldsymbol{A}_{(b)}\}_1^B$) as random variables. Instead, they were treated as elements of the valid structure sets. However, in this section, once a particular splitting rule is specified, these structures become random variables associated with certain distributions induced from the splitting rule. When there is no risk of ambiguity, we inherit the notation $\widehat{\Lambda}_{\boldsymbol{A}}$ to represent a tree estimator, where the randomness of $\boldsymbol{A}$ is understood as part of the randomness in the estimator itself. A similar strategy is applied to the forest version of the estimator. Before presenting the consistency results, we introduce an additional smoothness assumption on the hazard function:

Assumption 4. *For any fixed time point t, the cumulative hazard function $\Lambda(t \mid x)$ is L_1-Lipschitz continuous in terms of x, and the hazard function $\lambda(t \mid x)$ is L_2-Lipschitz continuous in terms of x, i.e., $|\Lambda(t \mid x_1) - \Lambda(t \mid x_2)| \leq L_1||x_1 - x_2||$ and $|\lambda(t \mid x_1) - \lambda(t \mid x_2)| \leq L_2\|x_1 - x_2\|$, respectively, where $\|\cdot\|$ is the Euclidean norm.*

We are now ready to state our main consistency results for the proposed survival tree model. Theorem 3 provides the point-wise consistency result. The proof is presented in Appendix.

Theorem 3. *Under the assumptions 1–4, the proposed survival tree model with random splitting rule is consistent, i.e., for each x,*

$$\sup_{t<\tau} \left|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda(t \mid x)\right| = O\left(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \left(\frac{k}{n}\right)^{\frac{c_1}{d}}\right),$$

with probability approaching to 1, where the constants $0 < c_2, c_4 < 1$, $c_3 = (1-2\alpha)/8$ *and* $c_1 = \frac{c_3(1-c_2)(1-c_4)}{\log_{1-\alpha}(\alpha)}$. *Consequently,*

$$\begin{aligned}&\sup_{t<\tau} E_X|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid X) - \Lambda(t \mid X)| \\ &= O\bigg(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d}} + \log(k)w_n\bigg),\end{aligned}$$

where

$$w_n = \frac{2}{\sqrt{n}} + d\exp\Big\{-\frac{c_2^2\log_{1/\alpha}(n/k)}{2d}\Big\} + d\exp\Big\{-\frac{(1-c_2)c_3c_4^2\log_{1/\alpha}(n/k)}{2d}\Big\}.$$

Remark 2.5.1. The first part $\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}}$ comes from the concentration bound results and the second part $(\frac{k}{n})^{\frac{c_1}{d}}$ comes from the bias part. We point out that the optimal rate is obtained by setting $k = n^{\frac{c_3}{c_3+1/2d\log_{1-\alpha}(\alpha)}}$, and then the optimal rate is close to $n^{-\frac{c_3}{2[c_3+1/2d\log_{1-\alpha}(\alpha)]}}$. If we always split at the middle point at each internal node, then the optimal rate degenerates to $n^{-\frac{1}{d+2}}$, which is the same rate as in [32].

The consistency result can be easily extended to survival forests with B trees. Theorem 4 presents an integrated version, which can be derived from Theorem 3.

Theorem 4. *Under the Assumptions 1-4, the proposed survival forest is consistent, i.e.*

$$\begin{aligned}&\lim_{B\to\infty}\sup_{t<\tau} E_X|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid X) - \Lambda(t \mid X)| \\ &= O\bigg(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d}} + \log(k)w_n\bigg),\end{aligned}$$

where w_n *is a sequence approaching to 0 as defined in Theorem 3,* $0 < c_2, c_4 < 1$, $c_3 = (1-2\alpha)/8$ *and* $c_1 = \frac{c_3(1-c_2)(1-c_4)}{\log_{1-\alpha}(\alpha)}$.

2.5.2 Consistency of survival forests with a nonparametric splitting rule when dimension d is infinite

In this section, we allow the dimension of the feature space d to go to infinity with sample size n. We assume there are d_0 important features for the failure time among d covariates, i.e., the true

model has size $|\mathcal{M}| = d_0 \leq d$. We implement the splitting rule as following. A similar idea for splitting rules has been considered in the guess-and-check forest [126] in the regression setting.

Algorithm 2: Splitting rule for marginal checked survival forest

1 For a currently internal node $\mathcal{A}$ containing at least $2k$ training samples, we pick a splitting variable $j \in \{1, \ldots, d\}$ uniformly at random;

2 We then pick the splitting point $\tilde{x}$ using the following rule such that both two child nodes contain at least proportion α of the samples in their parent node:

$$\tilde{x} = \arg\max_{x} \Delta(x),$$

where $\Delta(x) = \int_0^\tau \big|\widehat{\Lambda}_{\mathcal{A}_j^+(x)}(t) - \widehat{\Lambda}_{\mathcal{A}_j^-(x)}(t)\big|\mathrm{d}t$, $\mathcal{A}_j^+(x) = \{X : X^j \geq x\}$, and $\mathcal{A}_j^-(x) = \{X : X^j < x\}$, $X^{(j)}$ is the j-th dimension of X;

3 If either there is already a successful split on the variable j or the following inequality holds:

$$\Delta(\tilde{x}) \geq 2M_3\tau\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

for a universal constant M_3 then we split at $\tilde{x}$ along the j-th variable. If not, we randomly sample another variable out of the remaining variables and proceed to Step 2). When there are no remaining feasible variables, we randomly select an index out of d to proceed to a split.

Lemma 3 and 4 show that a d dimensional survival forest based on the above splitting rule is equivalent to a d_0 dimensional survival forest with probability approaching to 1. $\Lambda^*_{\mathcal{A},n}(t)$ is an essential tool to prove Lemma 3 and 4. Notice that $\Lambda^*_{\mathcal{A},n}(t)$ is a sample version of the asymptotic distribution of the terminal node $\mathcal{A}$. In Lemma 2, we show the bound of the difference of $\Lambda^*_{\mathcal{A},n}(t)$ and its integrated version $\Lambda^*_{\mathcal{A}}(t)$ across all valid nodes $\mathcal{A}$, where $\Lambda^*_{\mathcal{A}}(t)$ is as defined in Equation 2.6. The proof is given in Appendix.

Lemma 2. *Assume the density function of the failure time $f(t \mid x)$ is bounded by L for each x. The difference between $\Lambda^*_{\boldsymbol{A},n}(t)$ and $\Lambda^*_{\boldsymbol{A}}(t)$ is bounded by*

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \boldsymbol{A}\in\mathcal{V}_{\alpha,k}(\mathcal{D}_n)} \big|\Lambda^*_{\boldsymbol{A},n}(t \mid x) - \Lambda^*_{\boldsymbol{A}}(t \mid x)\big| \leq M_2\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

with probability larger than $1 - 1/\sqrt{n}$.

Now we are ready to present Lemma 3 and Lemma 4. The proof is shown in Appendix.

Lemma 3. *The probability that the proposed survival tree ever splits on a noise variable is smaller than* $3/\sqrt{n}$.

To establish consistency, we need one additional assumption about monotonicity of the failure distribution and the effect size of the censoring distribution.

Assumption 5. *Monotonicity of* dF. *Without loss of generality, assume that* $f = dF$ *is monotone increasing with respect to* X. *Furthermore, there is a minimum effect size* $\ell > 0$ *such that*

$$\int_0^\tau \left| \tilde{M} \int_0^t \frac{\int_{1/2}^1 f(s \mid X^{(j)} = x^{(j)}, X^{(-j)} = x^{(-j)}) \mathrm{d}x^{(j)}}{\int_{1/2}^1 [1 - F(s \mid X^{(j)} = x^{(j)}, X^{(-j)} = x^{(-j)})] \mathrm{d}x^{(j)}} \, \mathrm{d}s \right.$$
$$\left. - \frac{1}{\tilde{M}} \int_0^t \frac{\int_0^{1/2} f(s \mid X^{(j)} = x^{(j)}, X^{(-j)} = x^{(-j)}) \mathrm{d}x^{(j)}}{\int_0^{1/2} [1 - F(s \mid X^{(j)} = x^{(j)}, X^{(-j)} = x^{(-j)})] \mathrm{d}x^{(j)}} \, \mathrm{d}s \right| \mathrm{d}t \geq \ell,$$

for all $x \in [0,1]^d$ *and all important (non-noise) variables* j. *Here,* $X^{(-j)}$ *is a sub-vector of* X *obtained by removing the* j*th entry, and* $\tilde{M}$ *stands for the lower probability bound of censoring at* τ, *i.e.,* $\mathrm{pr}(C \geq \tau \mid X) \geq \tilde{M} > 0$.

Recall in Assumption 1, we assumed that $\mathrm{pr}(Y_i \geq \tau | X_i) \geq M > 0$. Hence taking $\tilde{M}$ as M automatically satisfies the above assumption of the censoring distribution; however, $\tilde{M}$ is usually larger than M. Note that Assumption 5 essentially bounds below the signal size regardless of any dependency structures between C_i and T_i for a given subject i. However, when the G_i's in Equation 2.5 are identical, the constant $\tilde{M}$ can be removed from the assumption.

Lemma 4. *At any given internal node, if an important variable is randomly selected and has never been used before, then the probability that the proposed survival tree splits on this variable is at least* $1 - 3/\sqrt{n}$.

Based on Lemma 3 and 4, we essentially only split on d_0 dimensions with probability $1 - 3/\sqrt{n}$. The consistency holds from Theorem 3. The following result shows the consistency of the proposed survival forest. The proof is almost identical to Theorem 4:

Theorem 5. *Under the Assumptions 1-5, the proposed survival forest using the splitting rule specified in Algorithm 2 is consistent, i.e.*

$$\lim_{B\to\infty}\sup_{t<\tau} E_X|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid X) - \Lambda(t \mid X)|$$
$$= O\bigg(\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d_0}} + \log(k)w_n\bigg),$$

where w_n is a sequence approaching to 0 as defined in Theorem 3, $0 < c_2, c_4 < 1$, $c_3 = (1-2\alpha)/8$ and $c_1 = \frac{c_3(1-c_2)(1-c_4)}{\log_{1-\alpha}(\alpha)}$.

Although we have developed a result where d can grow exponentially fast with n in this section, the splitting rule implemented was not a completely the same as the practically used version because it essentially checks only the signal where the candidate variable have never been used. This is done by comparing the signal for the two potential splits $X^{(j)} < 1/2$ versus $X^{(j)} < 1/2$ at an internal node. Once a variable is first used, it will be automatically included as a candidate thereafter. This idea is essentially the same as the protected variable set used in [142], where the protected set serves as the collection of variables that have used in previous nodes.

2.6 Discussion

In this chapter, we developed several fundamental results for tree- and forest-based survival models. Firstly, we investigated the within-node Nelson-Aalen estimator of the cumulative hazard function and developed a concentration inequality for independent but non-identically distributed samples. Secondly, we extended the result to develop a concentration bound across all possible fitted tree and forest models. Lastly, we developed consistency under two specific models with corresponding splitting rule methods.

In section 2.5, our results suggest that survival tree models are able to adapt to the sparsity structure of the underlying model. This is demonstrated in the second consistency results, where the number of true important variables d_0 is much smaller than the total number $d \geq d_0$. The splitting rules inherit information from nodes in the upper level of a tree and will always consider a variable that has been used before. There is a possible extension that the signals of the splitting

variables could be repeatedly check at all internal nodes. Alternative splitting rules which can further improve this upper limit on d are of theoretical interest.

CHAPTER 3

Nonparametric generalized fiducial inference for survival functions under censoring

3.1 Introduction

Fiducial inference can be traced back to a series of articles by the father of modern statistics R. A. [46, 47, 48, 49] who introduced the concept as a potential replacement of the Bayesian posterior distribution. A systematic development of the idea has been hampered by ambiguity, as [23] describes: "The reason for this lack of agreement and the resulting controversy is possibly due to the fact that the fiducial method has been put forward as a general logical principle, but yet has been illustrated mainly by means of particular examples rather than broad requirements." Indeed, we contend that until recently fiducial inference was applied to relatively a small class of parametric problems only.

Since the mid 2000s, there has been a renewed interest in modifications of fiducial inference. [61, 62] bring forward a mathematical definition of what they call the Generalized Fiducial Distribution (GFD). Having a formal definition allowed fiducial inference to be applied to a wide variety of statistical settings [65, 129, 130, 131, 128, 63, 30, 127, 66, 80, 86].

Other related approaches include Dempster-Shafer theory [37, 40], inferential models [91], and confidence distributions [132, 109, 68]. Objective Bayesian inference, which aims at finding non-subjective model based priors can also be seen as addressing the same basic question. Examples of recent breakthroughs related to reference prior and model selection are Bayarri et al. [8], Berger et al. [9, 10]. There are many more references that interested readers can find in the review article [64].

In this paper, we apply the fiducial approach in the context of survival analysis. To our knowledge, this is the first time fiducial inference has been systematically applied to an infinite-dimensional statistical problem. However, for use of confidence distributions to address some basic

non-parametric problems see Chapter 11 of [109]. In this manuscript, we propose a computationally efficient algorithm to sample from the GFD, and use the samples from the GFD to construct statistical procedures. The median of the GFD could be considered as a substitution for the Kaplan-Meier estimator [74], which is a classical estimator in survival analysis. Appropriate quantiles of the GFD evaluated at a given time provide pointwise confidence intervals for survival function. Similarly, the confidence intervals for quantiles of survival functions can be obtained by inverting the GFD.

The proposed pointwise confidence intervals maintain coverage in situations where classical confidence intervals often have coverage problems [45]. [45, 44] construct solutions to avoid these coverage problems. It is interesting to note that the conservative version of the proposed pointwise fiducial confidence interval is equivalent to beta product confidence procedure confidence interval of [45]. The other fiducial confidence interval proposed in this paper is based on log-linear interpolation and has the shortest length among all existing methods which maintain coverage.

We also construct curvewise confidence intervals for survival functions. Based on the curvewise confidence intervals, we propose a two sample test for testing whether two survival functions are equal. The proposed test does not need the proportional hazard assumption [17], and appears to be a good replacement for the log-rank test and sup log-rank test.

We establish an asymptotic theory which verifies the frequentist validity of the proposed fiducial approach. In particular, we prove a functional Bernstein–von Mises theorem for the GFD in Skorokhod's $D[0,t]$ space. Because randomness in GFD comes from two distinct sources the proof of this results is different from the usual proof of asymptotic normality for the Kaplan-Meyer estimator. As a consequence of the functional Bernstein–von Mises theorem, the proposed pointwise and curvewise confidence intervals provide asymptotically correct coverage, and the proposed survival function estimator is asymptotically equivalent to the Kaplan-Meier estimator.

We report results of a simulation study showing the proposed fiducial methods provide competitive, and in some cases superior performance to the methods in the literature. In particular, we compare the performance of the GFD intervals with classical confidence intervals like Greenwood [122], Borkowf [15], Strawderman-Wells [117, 118], nonparametric bootstrap [41, 3], constrained bootstrap [7], Thomas-Grunkemeier method [123], constrained beta [7], and beta product confidence procedure [45, 44] in various settings with small samples and/or heavy censoring. Additionally we also consider the setting of [7] in which the data contains fewer censored observations. Next, we

report several scenarios showing the desirable power of the GFD test in comparison to 12 different types of log-rank tests implemented in the R package survMisc [35]: original log-rank test [90]; Gehan-Breslow generalized Wilcoxon log-rank test [52]; Tarone-Ware log-rank test [121]; Peto-Peto log-rank test [101]; Modified Peto-Peto log-rank test [5]; Fleming-Harrington log-rank test [67] and corresponding supremum versions [51, 42].

We apply the proposed fiducial method to test the difference between chemotherapy and chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer [75]. The proposed fiducial test has the smallest p-value compared to existing methods. We also report a small simulation study based on 500 synthetic datasets mimicking the cancer data. The proposed fiducial test is more powerful than the 12 different tests described above.

3.2 Methodology

3.2.1 Fiducial approach explained

In this section, we explain the definition of a generalized fiducial distribution. We demonstrate the definition on the problem of estimating survival functions when no censoring is present. We start by expressing the relationship between the data $\boldsymbol{Y}$ and the parameter $\boldsymbol{\theta}$ using

$$\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{U}, \boldsymbol{\theta}), \tag{3.1}$$

where $\boldsymbol{G}(\cdot, \cdot)$ is a deterministic function termed the data generating equation, and $\boldsymbol{U}$ is a random vector whose distribution is independent of $\boldsymbol{\theta}$ and completely known. Data $\boldsymbol{Y}$ could be simulated by generating a random variable $\boldsymbol{U}$ and plugging it into the data generating equation (3.1). For example, a data generating equation for the $N(\mu, \sigma^2)$ model is $Y_i = G(U_i, \mu, \sigma) = \mu + \sigma\Phi^{-1}(U_i)$, where $\boldsymbol{U} = (U_1, \cdots, U_n)$ are independent and identically distributed $U(0, 1)$ and $\Phi(y)$ is the distribution function of the standard normal distribution.

The inverse cumulative distribution function method for generating random variables provides a common data generating equation for a nonparametric independent and identically distributed model:

$$Y_i = G(U_i, F) = F^{-1}(U_i), \quad i = 1, \ldots, n, \tag{3.2}$$

where $F^{-1}(u) = \inf\{y \in \mathbb{R} : F(y) \geq u\}$ is the usual "inverse" of the distribution function $F(y)$[24]. Notice that the distribution function F itself is the parameter $\boldsymbol{\theta}$ in this infinite dimensional model. The actual observed data is generated using the true distribution function F_0.

Roughly speaking a GFD is obtained by inverting the data generating equation, and [64] proposes a very general definition of GFD. However, in order to simplify the presentation, we will use an earlier, less general version found in [61]. The two definitions are equivalent for the models considered here.

We start by denoting the inverse image of the data generating equation (3.1) by

$$Q(\boldsymbol{y}, \boldsymbol{u}) = \{\boldsymbol{\theta} \,:\, \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\}.$$

For the special case (3.2) the inverse image is

$$Q(\boldsymbol{y}, \boldsymbol{u}) = \bigcap_{i=1}^{n} \{F : F(y_i) \geq u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0\}. \tag{3.3}$$

If $\boldsymbol{y}$ is the observed data and $\boldsymbol{u}_0$ the value of the random vector $\boldsymbol{U}$ that was used to generate it, then we are guaranteed that the true parameter value $\boldsymbol{\theta}_0 \in Q(\boldsymbol{y}, \boldsymbol{u}_0)$. However, we only know a distribution of $\boldsymbol{U}$ and not the actual value $\boldsymbol{u}_0$. Notice that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}_0, \boldsymbol{\theta}_0)$ and therefore only values of $\boldsymbol{u}$ for which $Q(\boldsymbol{y}, \boldsymbol{u}) \neq \emptyset$ should be considered. Let $\boldsymbol{U}^*$ be another random variable independent of and having the same distribution as $\boldsymbol{U}$. Since the conditional distribution of $\boldsymbol{U}^* \mid \{Q(\boldsymbol{y}, \boldsymbol{U}^*) \neq \emptyset\}$ can be viewed as summarizing our knowledge about $\boldsymbol{u}_0$, the conditional distribution of

$$Q(\boldsymbol{y}, \boldsymbol{U}^*) \mid \{Q(\boldsymbol{y}, \boldsymbol{U}^*) \neq \emptyset\} \tag{3.4}$$

can be viewed as summarizing our knowledge about $\boldsymbol{\theta}_0$.

Notice that $Q(\boldsymbol{y}, \boldsymbol{u})$ is a set that can contain more than one element. We deal with this by selecting a representative from the closure of $Q(\boldsymbol{y}, \boldsymbol{u})$. The distribution of a representative selected from (3.4) is a Generalized Fiducial Distribution. Based on the theoretical results presented, the non-uniqueness caused by this somewhat arbitrary choice disappears asymptotically. A possible conservative alternative to selecting a single representative from $Q(\boldsymbol{y}, \boldsymbol{u})$ could use the theory of belief functions [37, 112].

To describe the GFD in the particular case of (3.2) we define for all $s \geq 0$, $F^L_{(\boldsymbol{y},\boldsymbol{u})}(s) = \inf\{F(s) : F \in Q(\boldsymbol{y},\boldsymbol{u})\}$ and $F^U_{(\boldsymbol{y},\boldsymbol{u})}(s) = \sup\{F(s) : F \in Q(\boldsymbol{y},\boldsymbol{u})\}$. The closure of the inverse image (3.3) is a set of all distribution functions F that stay between $F^L_{(\boldsymbol{y},\boldsymbol{u})}$ and $F^U_{(\boldsymbol{y},\boldsymbol{u})}$. Also notice that $Q(\boldsymbol{y},\boldsymbol{u})$ is not empty if and only if the order of $\boldsymbol{u}$ matches the order of $\boldsymbol{y}$, with the understanding that in the case of ties in $\boldsymbol{y}$, the u_i's corresponding to the ties could be any order.

By exchangeability, the conditional distribution $\boldsymbol{U}^* \mid \{Q(\boldsymbol{y},\boldsymbol{U}^*) \neq \emptyset\}$ is the same as the distribution of $\boldsymbol{U}^*_{[\boldsymbol{y}]}$, where $\boldsymbol{U}^*_{[\boldsymbol{y}]}$ is the independent and identically distributed U(0,1) reordered to match the order of $\boldsymbol{y}$. Thus, any distribution stochastically larger than $F^L_{(\boldsymbol{y},\boldsymbol{U}^*_{[\boldsymbol{y}]})}$ and stochastically smaller than $F^U_{(\boldsymbol{y},\boldsymbol{U}^*_{[\boldsymbol{y}]})}$ is a GFD. Sampling from this fiducial distribution is easy to implement.

We consider the following 2 main options in using the GFD for inference. The first option is to construct conservative confidence sets. For example, when designing pointwise confidence intervals for the survival function at time s, we use quantiles of $1-F^U_{(\boldsymbol{y},\boldsymbol{U}^*_{[\boldsymbol{y}]})}(s)$ for lower bounds and quantiles of $1-F^L_{(\boldsymbol{y},\boldsymbol{U}^*_{[\boldsymbol{y}]})}(s)$ for upper bounds.

The second option is to select a suitable representative of $Q(\boldsymbol{y},\boldsymbol{U}^*_{[\boldsymbol{y}]})$. When there are no ties present in the data we propose to fit a continuous distribution function by using linear interpolation for the survival function on the log scale, i.e., the distribution function $F^I_{(\boldsymbol{y},\boldsymbol{u})}(s) = 1 - e^{L(s)}$, where $L(s)$ is the linear interpolation between $(0,0), (y_{(1)}, \log u_{(1)}), \ldots, (y_{(n)}, \log u_{(n)})$, and on the interval $(y_{(n)}, \infty)$ we extrapolate by extending the line between $(y_{(n-1)}, \log u_{(n-1)})$ and $(y_{(n)}, \log u_{(n)})$. We will call this the log-linear interpolation.

As usually, we denote the GFD for survival functions $S^L_{(\boldsymbol{y},\boldsymbol{u})} = 1-F^U_{(\boldsymbol{y},\boldsymbol{u})}$, $S^U_{(\boldsymbol{y},\boldsymbol{u})} = 1-F^L_{(\boldsymbol{y},\boldsymbol{u})}$, and $S^I_{(\boldsymbol{y},\boldsymbol{u})} = 1-F^I_{(\boldsymbol{y},\boldsymbol{u})}$. For simplicity, hereinafter we omit the subindex $(\boldsymbol{y},\boldsymbol{u})$. In the rest of this paper we will also denote Monte Carlo samples of the lower bound, the upper bound, and the log-linear interpolation of the GFD for the survival function by S^L_i, S^U_i, and S^I_i $(i = 1, \ldots, m)$, respectively.

To demonstrate the fiducial distribution of this section, we draw 300 observations from $Weibull(20, 10)$. We plot a fiducial sample of survival functions $S^I_i (i = 1, \ldots, 1000)$ and the empirical survival function in the left panel of Figure 3.2.

3.2.2 Fiducial approach in survival setting

In this section, we derive the GFD for the failure distribution based on right censored data. Here we treat the situation when the failure and censoring times are independent. The same GFD is derived under a more general model that includes dependence between failure and censoring times in the Appendix.

Let failure times X_i $(i = 1, \ldots, n)$ follow the true distribution function F_0 and censoring times Z_i $(i = 1, \ldots, n)$ have the distribution function R_0. We observe partially censored data $\{y_i, \delta_i\}$ $(i = 1, \ldots n)$, where $y_i = x_i \wedge z_i$ is the minimum of x_i and z_i, $\delta_i = I\{x_i \leq z_i\}$ denotes censoring indicator.

We consider the following data generating equation,

$$Y_i = F^{-1}(U_i) \wedge R^{-1}(V_i), \quad \delta_i = I\{F^{-1}(U_i) \leq R^{-1}(V_i)\}, \tag{3.5}$$

where U_i, V_i are independent and identically distributed $U(0, 1)$ and the actual observed data were generated using $F = F_0$ and $R = R_0$. We are committing a slight abuse of notation as $\boldsymbol{Y}$ in Equation (3.1) is $(\boldsymbol{Y}, \boldsymbol{\delta})$ in Equation (3.5) and $\boldsymbol{U}$ in Equation (3.1) is $(\boldsymbol{U}, \mathbf{V})$ in Equation (3.5).

For a failure event $\delta_i = 1$, we have full information about failure time x_i, i.e., $x_i = y_i$, and partial information about censoring time z_i, i.e., $z_i \geq y_i$. In this case, just as in the previous section,

$$F^{-1}(u_i) = y_i \quad \text{if and only if} \quad F(y_i) \geq u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0.$$

For a censored event $\delta_i = 0$, we know only partial information about x_i, i.e., $x_i > y_i$, and full information on z_i, i.e., $z_i = y_i$. Similarly,

$$F^{-1}(u_i) > y_i \quad \text{if and only if} \quad F(y_i) < u_i,$$
$$R^{-1}(v_i) = y_i \quad \text{if and only if} \quad R(y_i) \geq v_i, R(y_i - \epsilon) < v_i \text{ for any } \epsilon > 0.$$

To obtain the inverse map, we start by inverting a single observation. If $\delta_i = 1$, the inverse map for this datum is

$$Q_1^{F,R}(y_i, u_i, v_i) = \{F : F(y_i) \geq u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0\} \times \{R : R^{-1}(v_i) \geq y_i\}.$$

If $\delta_i = 0$, the inverse map is

$$Q_0^{F,R}(y_i, u_i, v_i) = \{F : F(y_i) < u_i\} \times \{R : R(y_i) \geq v_i, R(y_i - \epsilon) < v_i \text{ for any } \epsilon > 0\}.$$

Combining these we obtain the complete inverse map

$$Q^{F,R}(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v}) = \bigcap_i Q_{\delta_i}^{F,R}(y_i, u_i, v_i) = Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}) \times Q^R(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{v}), \tag{3.6}$$

where

$$Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}) = \left\{ F : \begin{cases} F(y_i) \geq u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0 & \text{for all } i \text{ s.t. } \delta_i = 1 \\ F(y_j) < u_j & \text{for all } j \text{ s.t. } \delta_j = 0 \end{cases} \right\}, \tag{3.7}$$

and $Q^R(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{v})$ is analogous. Notice that the inverse of $Q^{F,R}$ in (6) is in the form of a Cartesian product. This is a direct consequence of our choice of data generating equation, and it greatly simplifies the calculation of marginal fiducial distribution for failure times.

To demonstrate the inverse (3.7), Figure 3.1 presents the survival function representation of $Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u})$ for one small data set ($n = 8$) of $X \sim Weibull(20, 10)$ censored by $Z \sim Exp(20)$, and two different values of $\boldsymbol{u}$. The circle points denote failure observations and the triangle points denote censored observations. Any survival function lying between the upper and the lower bounds is an element of the closure of $Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u})$. In particular, we plot the log-linear interpolation going through the failure observations as described in Section 3.2.1 with a modification to ensure it satisfies the lower fiducial bound. Notice that the upper fiducial bound changes at the failure times only, while the lower fiducial bound changes at all failure times and at some censoring times depending on the value of $\boldsymbol{u}$.

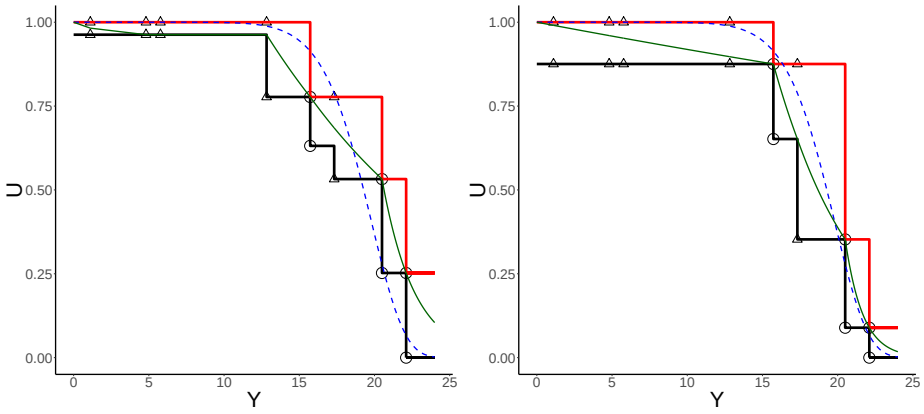


Figure 3.1: Two realizations of fiducial curves for a sample of size 8 from $Weibull(20, 10)$ censored by $Exp(20)$. Here fiducial curves refer to Monte Carlo samples S_i^L, S_i^U, and S_i^I $(i = 1, 2)$. The red curve is an upper bound and the black curve is a lower bound. The green curve is the log-linear interpolation. The circle points denote failure observations. The triangle points denote censored observations. The dashed blue curve is the true survival function of $Weibull(20, 10)$. Since the fiducial distribution reflects uncertainty we do not expect every fiducial curve to be close to the true survival function.

When defining the GFD, let $(\boldsymbol{U}^*, \mathbf{V}^*)$ be independent of and having the same distribution as $(\boldsymbol{U}, \mathbf{V})$. Because of the way the inverse (3.6) separates and the fact that $\boldsymbol{U}^*$ and $\mathbf{V}^*$ are independent, the (marginal) fiducial distribution for the failure distribution function F is

$$Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{U}^*) \mid \{Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{U}^*) \neq \emptyset\}. \tag{3.8}$$

The conditional distribution of $\boldsymbol{U}^* \mid \{Q(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{U}^*) \neq \emptyset\}$ can be sampled efficiently because it is the distribution of a particular random reordering of a sample of independent and identically distributed $U(0, 1)$. To this end we define $\mathcal{P}$ as the set of all permutations for which the permuted order statistics $\boldsymbol{u}_{(\Pi)}, \Pi \in \mathcal{P}$ satisfy $Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}_{(\Pi)}) \neq \emptyset$. Notice that the i-th element of $\boldsymbol{u}_\Pi$ is the $\Pi(i)$-th order statistics of $\boldsymbol{u}$, i.e., $\boldsymbol{u}_{(\Pi)_i} = \boldsymbol{u}_{(\Pi(i))}$. The set $\mathcal{P}$ is invariant to $\boldsymbol{u}$ as long as $\boldsymbol{u}$ has no ties. Therefore we simulate independent and identically distributed $U(0, 1)$, sort them, and then permute them using a permutation selected at random from $\mathcal{P}$.

The random permutation $\Pi \in \mathcal{P}$ can be generated sequentially starting from the smallest among the $\boldsymbol{y}$ to the largest. We start with the set $\mathcal{N} = \{1, \ldots, n\}$. At any given observation y_i, we select $\Pi(i)$ from $\mathcal{N}$ as either a) the smallest remaining value if the observed value y_i is a failure time or b) any of the remaining values selected at random if the observed value y_i is a censoring time. We then

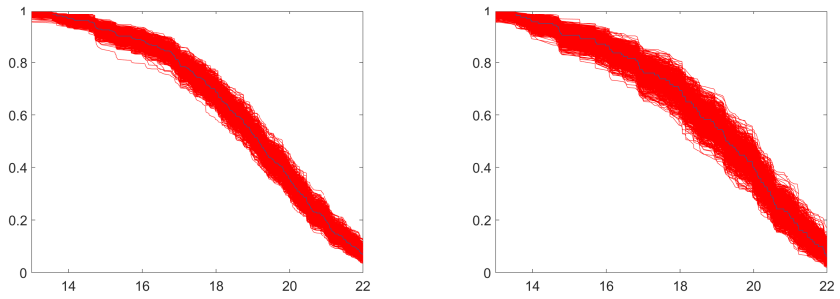


Figure 3.2: A plot of Monte Carlo realizations $S_i^I (i = 1, \ldots, 1000)$ sampled from the GFD based on a sample of 300 uncensored $Weibull(20, 10)$ observations, and the same 300 $Weibull(20, 10)$ observations censored by $Exp(20)$. The red curves are the 1000 fiducial curves, and the blue curve are the empirical survival function and the Kaplan-Meier estimator, respectively. As expected, we observe higher uncertainty in the fiducial sample under censoring.

remove the selected $\Pi(i)$ from $\mathcal{N}$ and proceed to the next smallest observation y_j until we exhaust the observations and $\mathcal{N}$.

Given $\{Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{U}^*) \neq \emptyset\}$, and the results of the first $i-1$ steps, the components of $\boldsymbol{U}^*$ not yet selected are exchangeable, which validates the proposed algorithm.

The details of this algorithm are in the Appendix. We implement the same two basic approaches to deriving statistical procedures from the GFD as in Section 3.2.1. To illustrate the fiducial distribution in the right censoring case, failure time X follows $Weibull(20, 10)$ and censoring time Z follows $Exp(20)$ with sample size 300. Censoring percentage is about 60%. We plot a fiducial sample of the survival function $S_i^I (i = 1, \ldots, 1000)$ and Kaplan-Meier estimator in the right panel of Figure 3.2. As expected, we see a wider spread of fiducial curves in the censoring case indicating higher uncertainty.

3.2.3 Inference based on fiducial distribution

In this section, we describe how to use fiducial samples for inference, specifically, point estimation, pointwise confidence intervals for survival functions and quantiles, curvewise confidence intervals, and testing. The actual numerical implementation will be based on a sample of survival functions S_i^L, S_i^U, and S_i^I $(i = 1, \ldots, m)$, i.e., the lower bound, the upper bound, and the log-linear interpolation respectively, obtained from the algorithm in the Appendix.

By Lemma 16 shown in the Appendix, the Kaplan-Meier estimator falls into the interval given by the expectation of the lower and upper fiducial bounds at any failure time t. However, instead

of using the Kaplan-Meier estimator we propose to use the pointwise median of the log-linear interpolation fiducial distribution as a point estimator of the survival function. It follows from Section 3.3 that the proposed estimator is asymptotically equivalent to the Kaplan-Meier estimator. Numerically, we estimate the median of the GFD at time x by computing a pointwise median of the fiducial sample $S_i^I(x)$ $(i = 1, \ldots, m)$. We report a simulation study in Section 3.4.1 to support this estimator.

As explained at the end of Section 3.2.1 we use two types of pointwise confidence intervals, conservative and log-linear interpolation, using quantiles of appropriate parts of the fiducial samples. For example, a 95% confidence log-linear interpolation confidence interval for $S(x)$ is formed by using the empirical 0·025 and 0·975 quantiles of $S_i^I(x)$. Similarly, a 95% conservative confidence interval is formed by taking the empirical 0·025 quantile of $S_i^L(x)$ as a lower limit and the empirical 0·975 quantile of $S_i^U(x)$ as an upper limit. Simulation results in Section 3.4.1 show that the proposed confidence intervals match or outperform their main competitors regarding coverage and length.

In order to save space, in the rest of this section we present procedures based on the log-linear interpolation sample only. A conservative version can be obtained analogously. In survival analysis, we are also interested in confidence intervals for quantile q of the survival function, where $0 < q < 1$. We obtain such a confidence interval by inverting the procedure of computing the pointwise confidence interval. Specifically, a 95% confidence interval is obtained by taking empirical 0·025 and 0·975 quantiles of the inverse of fiducial sample S_i^I evaluated at q.

Next, we discuss the use of the GFD to obtain simultaneous curvewise confidence bands. In particular, for a $1 - \alpha$ curvewise confidence set we propose using a band $\{S : \|S - M\| \leq c\}$ of fiducial probability $1 - \alpha$, where M denotes the pointwise median of the GFD, and $\|\cdot\|$ is the L_∞ norm, i.e., $\|S - M\| = \max_x |S(x) - M(x)|$. Numerically we implement this by using a fiducial sample. Let

$$l_j = \|S_j^I - \hat{M}\| = \max_x |S_j^I(x) - \hat{M}(x)|, j = 1, \ldots, m,$$

where $\hat{M}$ is the estimated pointwise median of the GFD. Then we form the 95% curvewise confidence band $\{S : \|S - \hat{M}\| \leq \hat{c}\}$, where $\hat{c}$ is the 0·95 quantile of l_j. To illustrate, we plot 95% pointwise and curvewise confidence intervals for the $Weibull(20, 10)$ example under right censoring in Figure 3.3.

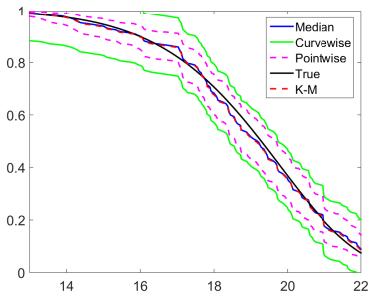


Figure 3.3: An example of 95% pointwise and curvewise confidence intervals of survival function by proposed log-linear interpolation approach.

The curvewise confidence set could be inverted for testing. The resulting test is different from the log-rank test [90] and its modifications. Based on our definition of the $1-\alpha$ fiducial band, the fiducial p-value for the two sided test

$$H_0 : S(t) = S_0(t) \text{ for all } t, \quad H_1 : S(t) \neq S_0(t) \text{ for some } t,$$

is $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(\|S^I - M\| \geq \|S_0 - M\|)$, where $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}$ stands for a fiducial probability computed for observed data $(\boldsymbol{y}, \boldsymbol{\delta})$, S^I stands for a random survival function following the log-linear interpolation GFD, and as before M is the pointwise median of the fiducial distribution. We estimate this p-value from a fiducial sample by finding the largest α for which $1-\alpha$ curvewise confidence set contains S_0. In particular, let

$$l_0 = \max_x |S_0(x) - \hat{M}(x)|, \quad l_j = \max_x |S^I_j(x) - \hat{M}(x)|, \ j = 1, \ldots, m. \tag{3.9}$$

Numerically, we approximate the p-value by the proportion of the fiducial sample satisfying $l_j \geq l_0$.

While the log-rank test is a two sided test only, the fiducial approach could also be used to define one sided tests. For example for testing

$$H_0 : S(t) \geq S_0(t) \text{ for all } t, \quad H_1 : S(t) < S_0(t) \text{ for some } t,$$

we define a fiducial p-value as the fiducial probability $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(\max_x\{S^I(x) - M(x)\} \geq \max_x\{S_0(x) - M(x)\})$.

Finally, let us consider two sample testing. For each sample, we have observed values $\boldsymbol{y}^i$ and censoring indicators $\boldsymbol{\delta}^i$, $i = 1, 2$. The two independent log-linear interpolation GFDs are denoted by $S^I_{(\boldsymbol{y}_i,\boldsymbol{\delta}_i)}$, $i = 1, 2$. When testing $H_0 : S^1 - S^2 = \Delta_0$ we define a fiducial p-value as the fiducial probability $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(\|S^I_{(\boldsymbol{y}_1,\boldsymbol{\delta}_1)} - S^I_{(\boldsymbol{y}_2,\boldsymbol{\delta}_2)} - M_D\| \geq \|\Delta_0 - M_D\|)$, where M_D is the median of the difference of the two GFDs.

Numerically, we evaluate the p-value in the same fashion as in Equation (3.9). We will compare the performance of the proposed fiducial test with the log-rank test and sup log-rank test with different weights for the two sample settings by simulation in Section 3.4.2.

3.3 Theoretical results

Recall that the GFD is a data dependent distribution $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}$ that is defined for every fixed data set $(\boldsymbol{y}, \boldsymbol{\delta})$. It can be made into a random measure $\text{pr}^*_{\boldsymbol{Y},\boldsymbol{\delta}}$ in the same way as one defines the usual conditional distribution, i.e., by plugging random variables $(\boldsymbol{Y}, \boldsymbol{\delta})$ for the observed data set. In this section, we will study the asymptotic behavior of this random measure assuming there are no ties with probability 1.

[103] prove a Bernstein-von Mises theorem for the exchangeably weighted bootstrap, of which the Bayesian bootstrap [107] is an example. However, the result of [103] is not applicable in the survival settings due to the fact that the jump sizes of F^L or F^U are not exchangeable. In this section, we study the theoretical properties of the GFD in the survival setting. For simplicity, we state the results in this section using upper fiducial bound of survival functions S^U, i.e., the lower fiducial bound of cumulative distribution functions F^L. Lemma 15 in the Appendix proves that the same results hold for S^L and S^I.

First we introduce some notations: X_i is failure time, Z_i is censoring time, Y_i is the observed minimum of failure and censoring time, and $\delta_i = I\{X_i \leq Z_i\}$ is the censoring indicator. We define the counting process

$$N_i(t) = I\{Y_i \leq t\}\delta_i, \quad \bar{N}(t) = \sum_{i=1}^{n} N_i(t),$$

and the at-risk process

$$K_i(t) = I\{Y_i \geq t\}, \quad \bar{K}(t) = \sum_{i=1}^{n} K_i(t).$$

We need the following two assumptions which are also needed for theoretical study of the Kaplan-Meier estimator [50].

Assumption 6. *There exists a function π such that, as $n \to \infty$,*

$$\sup_{0 \le t < \infty} \left| \bar{K}(t)/n - \pi(t) \right| \to 0 \quad \textit{almost surely.}$$

This assumption is very mild. For example if Y_i are independent and identically distributed, it is implied by Glivenko-Cantelli Theorem; see the discussion following Assumption 6.2.1 in [50] for more details.

Assumption 7. *F_0 is absolutely continuous.*

Let $\tilde{S}(t) = \prod_{s \le t} \{1 - \Delta \bar{N}(s)/\bar{K}(s)\}$ be the Kaplan-Meier estimator. It is well-known, see for example Theorem 6.3.1 of [50], that for any t satisfying $\pi(t) > 0$,

$$\sqrt{n}\{\tilde{F}(\cdot) - F_0(\cdot)\} \to \{1 - F_0(\cdot)\} W\{\gamma(\cdot)\} \quad \text{in distribution on } D[0,t], \tag{3.10}$$

where $\tilde{F}(t) = 1 - \tilde{S}(t)$, $\gamma(t) = \int_0^t \pi^{-1}(s) d\Lambda(s)$, W is Brownian Motion, and Λ is the cumulative hazard function.

Recall that the procedure for sampling from (3.8) in Section 3.2.2 defines a random permutation Π. Conditional on $\{Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{U}^*) \neq \emptyset\}$ and the results of the first $i-1$ steps, the distribution of the $\Pi(i)$-th order statistic $\boldsymbol{U}^*_{(\Pi(i))}$ corresponding to a failure time y_i is the minimum of $\bar{K}(y_i)$ independent random variables distributed as uniform on $(\boldsymbol{U}^*_{(\Pi(j))}, 1)$, where $\boldsymbol{U}^*_{(\Pi(j))}$ corresponds to the failure time y_j immediately preceding y_i. If y_i is the smallest failure time then set $\boldsymbol{U}^*_{(\Pi(j))} = 0$. Since $S^U(y_i) = 1 - \boldsymbol{U}^*_{(\Pi(i))}$ for all failure times, the upper bound of the GFD has a distribution that can be written as

$$S^U(t) = \prod_{s_i \le t} \{1 - \Delta \bar{N}(s_i) B_i\}, \tag{3.11}$$

where $\Delta \bar{N}(t) = \bar{N}(t) - \bar{N}(t-)$, s_i are ordered failure times, and B_i are independent $Beta(1, \bar{K}(s_i))$, respectively. Its expectation $\hat{S}(t) = E\{S^U(t)\}$ can be easily computed from (3.11) as

$$\hat{S}(t) = \prod_{s \le t} \left\{ 1 - \frac{\Delta \bar{N}(s)}{1 + \bar{K}(s)} \right\}. \tag{3.12}$$

Equation (3.12) provides us with a modification of the Kaplan-Meier estimator that also satisfies (3.10). We will use this modification throughout this section and in all the proofs that can be found in the Appendix. As our first result, we prove a concentration inequality for $S^U(t)$.

Theorem 6. *The following bound holds for any dataset with $\bar{K}(t) \geq 1$ and any $\epsilon > 0$,*

$$pr^*_{\boldsymbol{y},\boldsymbol{\delta}}\{\sup_{s\leq t}|S^U(s)-\hat{S}(s)| \geq 3\epsilon^2/n^{1/2}+\bar{N}(t)/\bar{K}(t)^{-2}\} \leq \bar{N}(t)[(1-\epsilon/n^{3/4})^{\bar{K}(t)}+0{\cdot}4^{\bar{K}(t)}+n/\{\epsilon^2\bar{K}(t)\}^2]. \quad (3.13)$$

Remark 3.3.1. Theorem 6 and Assumption 6 imply that the fiducial distribution is uniformly consistent. In particular, provided that we have a sequence of data so that $\bar{K}(t)/n \to \pi(t) > 0$, the right-hand side of (3.13) is $O(n^{-1})$ whenever $\epsilon^2 = n^{1/2}$.

Before presenting our main result we need two additional assumptions.

Assumption 8. *$\int_0^t f_n(s)/\bar{K}(s)d\bar{N}(s) \to \int_0^t f(s)\lambda(s)ds$ almost surely for any $t \in \mathcal{I} = \{t : \pi(t) > 0\}$ and $f_n \to f$ uniformly.*

Assumption 8 is reasonable since the probability of failure and censoring both happening in the $[t, t + \Delta t)$ is of a higher order $O((\Delta t)^2)$.

Assumption 9. *$\sup_{0\leq s\leq t}|\tilde{F}(s) - F_0(s)| \to 0$ almost surely for any $t \in \mathcal{I} = \{t : \pi(t) > 0\}$, where $\tilde{F} = 1 - \tilde{S}$, and $\tilde{S}$ is the Kaplan-Meier estimator.*

Remark 3.3.2. The strong consistency result of Assumption 9 has been proved for the model described in Section 3.2.2 by [57, 119]. Moreover, Assumption 9 is only needed for establishing a strong version of Theorem 7, i.e., convergence in distribution almost surely. If the Kaplan-Meier estimator only converges in probability, then the convergence mode in Theorem 7 is in distribution in probability.

The following theorem establishes a Bernstein-von Mises theorem for the fiducial distribution. In particular, we will show that the fiducial distribution of $n^{1/2}\{F^L(\cdot)-\hat{F}(\cdot)\}$, where $\hat{F}(\cdot) = 1-\hat{S}(\cdot)$ and $F^L(\cdot) = 1 - S^U(\cdot)$, converges in distribution almost surely to the same Gaussian process as in (3.10). To understand the somewhat unusual mode of convergence used here, notice that there are two sources of randomness present. One is from the fiducial distribution itself that is derived from each fixed data set. The other is the usual randomness of the data. The mode of convergence here is in distribution almost surely, i.e., the centered and scaled fiducial distribution viewed as a

random probability measure converges almost surely to the Gaussian distribution described in the right-hand side of Equation (3.10) using the weak topology on the space of probability measures.

Theorem 7. *Based on Assumptions 6–9, for any* $t \in \mathcal{I} = \{t : \pi(t) > 0\}$, $n^{1/2}\{F^L(\cdot) - \hat{F}(\cdot)\} \to \{1 - F_0(\cdot)\}W\{\gamma(\cdot)\}$ *in distribution on* $D[0,t]$ *almost surely, where* $\gamma(t) = \int_0^t \pi^{-1}(s)d\Lambda(s)$.

Notice that Theorem 7 implies that the pointwise fiducial confidence intervals are equivalent to the asymptotic confidence intervals based on the Kaplan-Meyer estimator. This fact can be also seen from Theorem 2 of [45]. This is in line with our experience with GFD in parametric settings, i.e., the fiducial procedures are asymptotically as efficient as maximum likelihood. The following corollary shows that Theorem 7 also implies that all the pointwise and curvewise confidence intervals described in Section 3.2.3 have asymptotically correct coverage. Consequently, the tests described in Section 3.2.3 also have asymptotically correct type I error.

Corollary 3. *Let* $\Psi\{\phi(\cdot)\}$ *be a map:* $D[0,t] \to \mathbb{R}$ *satisfying, there exists a function* ψ *so that*

$$\Psi\{\phi(\cdot)\} = \Psi\{-\phi(\cdot)\}, \quad \Psi\{a\phi(\cdot)\} = \psi(a)\Psi\{\phi(\cdot)\}, \tag{3.14}$$

for all $\phi \in D[0,t]$, $a > 0$, *the distribution of the random variable* $\Psi[\{1-F_0(\cdot)\}W\{\gamma(\cdot)\}]$ *is continuous and the* $(1-\alpha)$*-th quantile of this distribution is unique.*

Then, under the assumptions in Theorem 7, any set $C_{n,\alpha} = \{F : \Psi\{F(\cdot) - \hat{F}(\cdot)\} \leq \epsilon_{n,\alpha}\}$ *with* $pr^*_{\boldsymbol{y},\boldsymbol{\delta}}(C_{n,\alpha}) = 1 - \alpha$ *is a* $1-\alpha$ *asymptotic confidence set for* F_0.

3.4 Simulation study

3.4.1 Coverage of pointwise confidence intervals and mean square error of point estimators

We present comparisons of frequentist properties of the proposed fiducial confidence intervals with a number of competing methods. We will consider two basic groups of settings, one with heavy censoring from [45] and another with a moderate level of censoring from [7]. In both cases the proposed GFD intervals perform comparable to or better than the reported methods.

First we reproduce the settings in [45] that have a very high level of censoring. [45] compared their proposed beta product confidence procedure methods with a number of asymptotic methods.

These include Greenwood by logarithm transformation, the confidence interval on the Kaplan-Meier estimator using Greenwood's variance by logarithm transformation [122]; Modified Greenwood by logarithm transformation which modifies the estimator of variance for the lower limit by multipling the Greenwood's variance estimator by $K(y_i)/K(t)$ at t, where y_i is the largest observed survival less than or equal to t [122]; Borkowf by logarithm transformation, which gives wider intervals with more censoring and assumes normality on $\log(\tilde{S}(t))$, where $\tilde{S}(t)$ is the Kaplan-Meier estimator [15]; shrinkage Borkowf by logarithm transformation, which uses a shrinkage estimator of the Kaplan-Meier estimator with a hybrid variance estimator [15]; Strawderman-Wells, that uses the Edgeworth expansion for the distribution of the studentized Nelson-Aalen estimator [117, 118]; Thomas-Grunkemeier, a likelihood ratio method which depends on a constrained product-limit estimator of the survival function [123]; Constrained Beta, which refers the distribution of $\tilde{S}(t)$ to a beta distribution subject to some constraints [7]; nonparametric Bootstrap [41, 3]; Constrained Bootstrap, an improved bootstrap approximation subject to some constraints [7].

Simulation studies in [45] show that the above asymptotic methods have a coverage problem, i.e., the error rate of 95% confidence interval of all these methods is larger than 5% in their high censoring scenarios. Therefore in this setting we focus on comparing the fiducial methods with our main competing methods, which are beta product confidence procedure [45], mid-p beta product confidence procedure [44], see also Chapter 11 of [109], and Binomial-C [31], which maintain the coverage. We report the error rate of coverage and the average width of confidence intervals for fiducial methods, beta product confidence procedure using method of moment, beta product confidence procedure using Monte Carlo with samples 1000, mid-p beta product confidence procedure, and Binomial-C. We point out that Clopper-Pearson Binomial-C requires knowledge of the censoring times for each individual [45].

We consider following two scenarios in [45]. In the first scenario, failure time X is $Exp(10)$, censoring time Z is $U(0,5)$. We simulate 100000 independent datasets of size 30 and applied our methods with fiducial sample size 1000. In the second scenario, we reproduce the setting using a mixture of exponentials to mimic the pilot study of treatment in severe systemic sclerosis [98]. In particular, failure time X is a mixture of $Exp(0{\cdot}227)$ with probability 0.187 and $Exp(22{\cdot}44)$ with probability $0{\cdot}813$, censoring time Z is $U(2,8)$. We simulate 100000 independent datasets of size 34 and apply our methods with fiducial sample size 1000.

The simulation results are in Table 3.1 and Table 3.2 for each scenario, respectively. In the tables, L denotes the error rate that the true parameter is less than the lower confidence limit; U denotes the error rate that the true parameter is greater than the upper confidence limit. The two-sided error rate is obtained by adding the values in column L and U. Values less than 2·5% in individual columns, 5% in aggregate, indicate good performance. W is the average width of the confidence interval. The row labels are: FD-I the proposed method using log-linear interpolation; FD-C the proposed conservative confidence interval; BPCP-MM beta product confidence procedure using method of moment; BPCP-MC beta product confidence procedure using Monte Carlo; BPCP-MP mid-p beta product confidence procedure; BN Clopper-Pearson Binomial-C. From Table 3.1 and Table 3.2 we see that our confidence intervals using log-linear interpolation maintain the aggregate coverage, are much shorter, but may be slightly biased to the left. Not surprisingly, the performance of the proposed conservative confidence interval is similar to the beta product confidence procedure method. Recall, Table 1 and Table F·2 in [45] show all asymptotic methods mentioned above have a coverage problem in this heavily censored setup, and so are not considered here.

We also perform a simulation for the mean square error of survival functions, adopting a setting in [45]. Here, failure time is $Exp(1)$, and censoring time is $U(0, 5)$. We simulate 100000 independent datasets of size 25 and apply our fiducial methods with fiducial sample size 10000. Since the Kaplan-Meier estimator is not defined after the largest observation if it is censored, we follow [45] and define it in three ways after the last observation: KML is defined as 0, KMH is defined as the Kaplan-Meier at the last value, and KMM=0·5*KML+0·5*KMH. We evaluate mean square error at t, where $S(t) = 0{\cdot}99$, 0·9, 0·75, 0·5, 0·25, 0·1, 0·01. We report the results in Table 3.3. FD-I uses the pointwise median of the log-linear interpolation fiducial distribution as a point estimator of the survival function. BPCP-MM and BPCP-MP are associated median unbiased estimators defined in [45]. We see the proposed fiducial approach has the smallest mean square error for $S(t) = 0{\cdot}99$, 0·9, 0·75, 0·5, 0·25, 0·1, 0·01.

Our second simulation study setting comes from [7] where the data contains more exact observations. In the first scenario, survival time X follows $Exp(10)$, and censoring time Z is $Exp(50)$. In the second scenario, survival time X follows $Exp(10)$, and censoring time Z is $Exp(25)$. We plot the empirical error rates from 5000 simulations with sample size $n = 100$ of different non-asymptotic confidence intervals in the Figures 3.4 and 3.5, respectively. From the Figures 3.4, Figure 3.5, and

Table 3.1: Error rate (in percent) and average width of 95% confidence intervals for scenario 1

	t=1			t=2			t=3			t=4		
	L	U	W	L	U	W	L	U	W	L	U	W
FD-I	1·9	2·7	0·21	1·5	2·8	0·29	1·4	3·0	0·37	1·8	3·1	0·45
FD-C	0·0	1·4	0·26	0·3	1·6	0·36	0·1	1·5	0·46	0·0	1·4	0·63
BPCP-MM	0·0	1·3	0·26	0·3	1·4	0·35	0·1	1·3	0·46	0·0	1·0	0·62
BPCP-MC	0·0	1·3	0·25	0·4	1·5	0·35	0·1	1·5	0·46	0·0	1·4	0·63
BPCP-MP	0·0	2·2	0·23	0·8	2·3	0·32	0·4	2·2	0·41	0·0	2·0	0·57
BN	0·0	1·4	0·26	0·7	1·3	0·38	0·6	1·3	0·51	0·1	0·9	0·70

Table 3.2: Error rate (in percent) and average width of 95% confidence intervals for scenario 2

	t=3			t=4			t=5			t=6		
	L	U	W	L	U	W	L	U	W	L	U	W
FD-I	2·2	2·7	0·29	1·9	2·9	0·31	1·7	3·0	0·33	1·5	3·2	0·36
FD-C	1·2	1·7	0·33	0·7	1·8	0·36	0·4	1·8	0·40	0·1	1·7	0·46
BPCP-MM	1·3	1·7	0·33	0·7	1·7	0·35	0·4	1·6	0·39	0·1	1·4	0·46
BPCP-MC	1·2	1·8	0·32	0·7	2·0	0·35	0·4	1·9	0·39	0·1	1·9	0·46
BPCP-MP	1·8	2·1	0·30	1·6	2·4	0·32	0·9	2·5	0·36	0·4	2·3	0·41
BN	1·4	1·5	0·35	1·5	1·6	0·40	1·5	1·7	0·46	1·0	1·5	0·56

Table 3.3: Mean square error of survival function estimators

	$S(t)$ =0·99	$S(t)$ =0·9	$S(t)$ =0·75	$S(t)$ =0·5	$S(t)$ =0·25	$S(t)$ =0·1	$S(t)$ =0·01
FD-I	0·30	3·11	7·08	10·08	8·24	4·38	1·20
BPCP-MM	0·44	3·44	7·50	10·60	8·83	4·40	1·50
BPCP-MP	0·48	3·65	7·54	10·62	8·99	5·79	0·26
KML	0·39	3·61	7·71	10·94	9·38	6·17	0·28
KMM	0·39	3·61	7·71	10·94	9·35	5·77	0·79
KMH	0·39	3·61	7·71	10·94	9·33	5·65	2·92

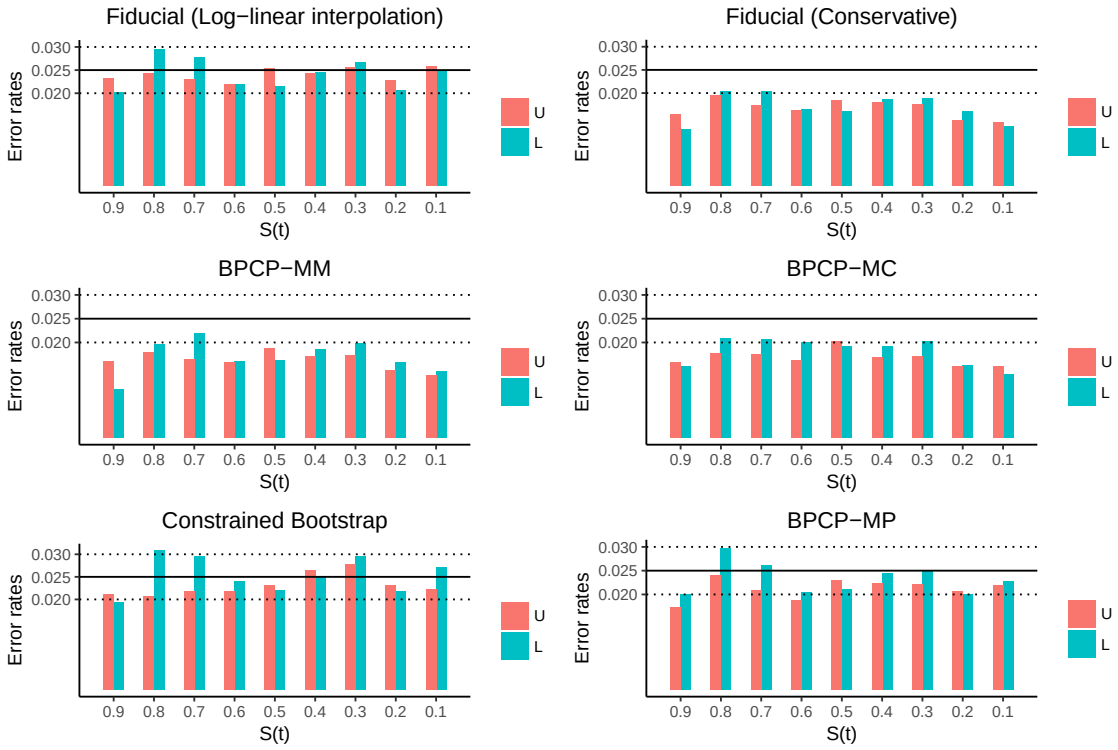


Figure 3.4: Error rates from 5000 simulations of different confidence intervals with $n = 100$, survival time follows $Exp(10)$, and censoring time follows $Exp(50)$. L denotes the error rate that the true parameter is lower than lower bound. U denotes the error rate that the true parameter is above the upper bound.

the figures in [7], we see that the fiducial confidence intervals do as well as the constrained bootstrap in these settings.

3.4.2 Comparisons between the proposed fiducial test and different types of log-rank tests for two sample testing

We compare the performance of the proposed fiducial approach with different types of tests for testing the equality of two survival functions [35]. A common approach to testing the difference of two survival curves is the log-rank test. There are several modifications of the log-rank tests that consist of re-weighing. In our tables, LR denotes the original log-rank test with weight 1 [90]; GW, i.e., Gehan-Breslow generalized Wilcoxon, denotes log-rank test weighted by the number at risk overall [52]; TW denotes log-rank test weighted by the square root of the number at risk overall [121]; PP denotes log-rank test with Peto-Peto's modified survival estimate [101]; MPP denotes log-rank test with modified Peto-Peto's survival estimate [5]; FH denotes Fleming-Harrington weighted log-rank test [67]. The supremum family of tests are designed to detect differences in survival curves

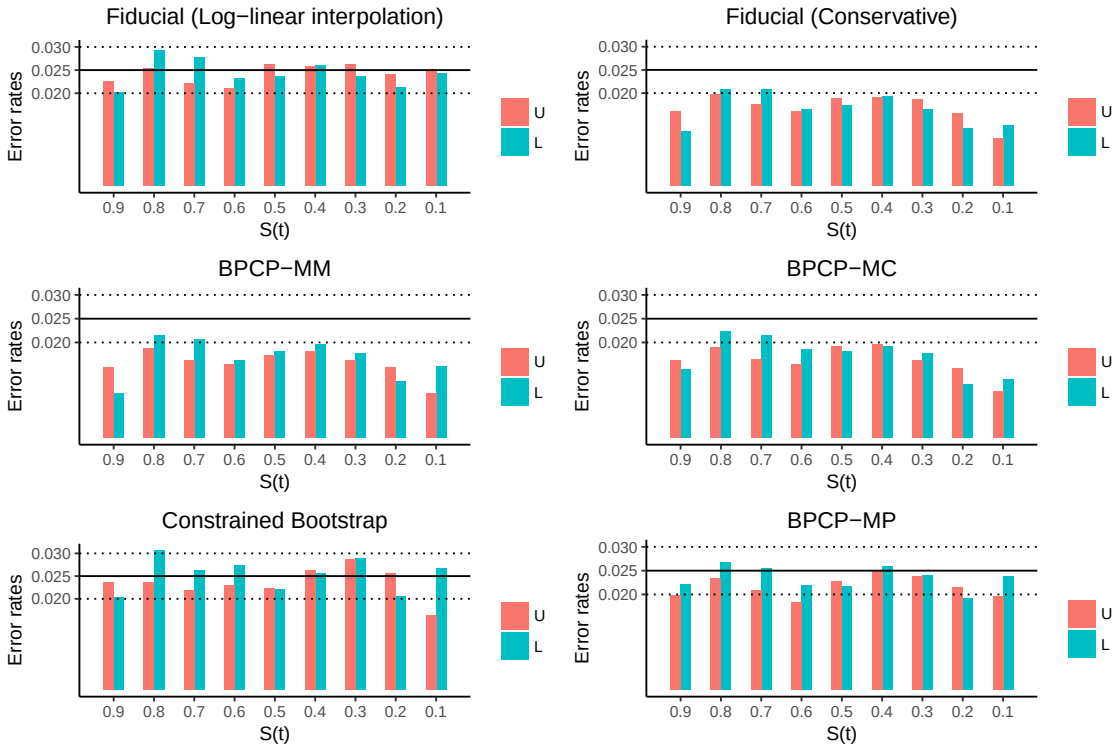


Figure 3.5: Error rates from 5000 simulations of different confidence intervals with $n = 100$, survival time follows $Exp(10)$, and censoring time follows $Exp(25)$. L denotes the error rate that the true parameter is lower than lower bound. U denotes the error rate that the true parameter is above the upper bound.

which cross [51, 42]. SLR denotes the original sup log-rank test with weight 1; SGW denotes the sup version of GW; STW denotes the sup version of TW; SPP denotes the sup version of PP; SMPP denotes the sup version of MPP; SFH denotes the sup version of FH.

Four scenarios are considered in this section. In the first scenario the null hypothesis is true. In the remaining three scenarios we consider various departures from the null hypothesis. For each scenario we simulated 500 independent datasets of size 200, and applied the proposed fiducial test with fiducial sample size 1000 as well as the 12 existing methods mentioned above. Then we calculate the percentage of p-values less than 0·05. If the null hypothesis is true, the p-value should follow uniform distribution and the percentage should be around 5%. If the null hypothesis is false, a higher percentage is preferable as it means bigger power.

In the first scenario, for the first group, failure time is $Weibull(2, 1)$ and censoring time follows $|N(0, 1)|$. The censoring percentage is approximately 55%. For the second group, failure time is again $Weibull(2, 1)$ but censoring time is $Exp(1)$. The censoring percentage is approximately 60%. We observe that p-values of all methods follow uniform distribution under H_0. Table 3.4 shows the

Table 3.4: Percentage of p-value less than 0·05 (%)

Fiducial	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
5·0	5·0	6·6	6·4	6·4	6·0	4·8	4·6	6·0	6·0	6·0	6·0	4·2

Table 3.5: Percentage of p-value less than 0·05 (%)

Fiducial	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
100	71·0	98·4	27·6	49·2	50·8	100	100	100	100	100	100	100

percentage of p-value less than 0·05. The percentages of p-value less than 0·05 of all methods are about 0·05.

In the second scenario for the first group, failure time follows $Exp(30)$ and censoring time follows $Exp(30)$. The censoring percentage is about 50%. For the second group, we use $Weibull(30, 20)$ to generate failure time, and $Exp(30)$ for censoring time with censoring percentage of about 50%. The power of the test at the α =0·05 level, i.e. the proportion of $p < 0{\cdot}05$ is shown in Table 3.5. In this scenario, the proposed fiducial test is as powerful as the sup log-rank tests.

In the third scenario, for the first group, let $Weibull(30, 20)$ be the distribution of failure time and $U(0, 80)$ be the distribution of censoring time. The censoring percentage is about 25%. For the second group, let $Weibull(20, 20)$ be the distribution of failure time and $U(0, 80)$ be the distribution of censoring time. The censoring percentage is about 20%. The power of the test at the $\alpha = 0{\cdot}05$ level, i.e. the proportion of $p < 0{\cdot}05$ is shown in Table 3.6. We see that only SGW, SPP, SMPP and the proposed fiducial test have power larger than half at $\alpha = 0{\cdot}05$ level.

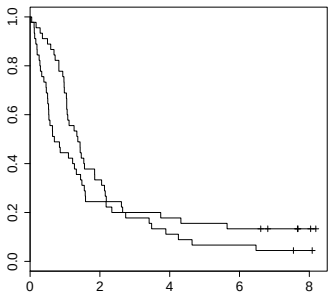
In the fourth scenario, for the first group, failure time follows $Exp(1)$, and censoring time follows $|N(0, 1)|$ with censoring percentage of about 50%. For the second group, failure time is $|N(0, 1)|$ censored by $Weibull(2, 1)$. The censoring percentage is about 40%. The power of the test at the $\alpha = 0{\cdot}05$ level, i.e. the proportion of $p < 0{\cdot}05$ is shown in Table 3.7. We see that only FH, SFH, and the proposed fiducial test have power larger than 0·1 at the $\alpha = 0{\cdot}05$ level. FH seems to use better weights than other log-rank tests, however, the proposed fiducial test doesn't need to specify any weight and is better than FH in this scenario.

Table 3.6: Percentage of p-value less than 0·05 (%)

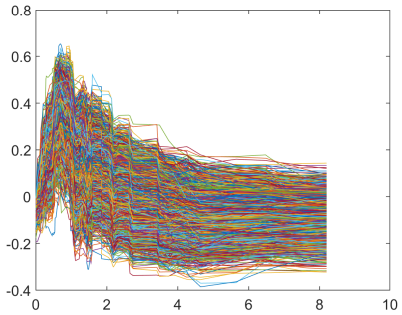
Fiducial	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
54·2	21·4	15·2	4·8	14·0	14·4	39·4	26·6	55·0	39·4	53·8	54·0	29·6

Table 3.7: Percentage of p-value less than 0·05 (%)

Fiducial	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
19·0	7·8	5·4	4·8	4·6	4·6	16·2	6·6	7·4	5·4	5·4	5·4	10·6



(a) Kaplan-Meier estimators for two treatment groups.



(b) Difference of two sample fiducial distributions using log-linear interpolation.

Figure 3.6

3.5 Gastric tumor study

In this section, we analyze the following dataset presented in [75]. A clinical trial of chemotherapy against chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer was conducted by the Gastrointestinal Tumor Study Group [108]. In this trial, forty-five patients were randomized to each of the two groups and followed for several years. We draw the Kaplan-Meier curves for these two datasets in Figure 3.6a.

By examining the plot in Figure 3.6a we notice that the two hazards appear to be crossing which could pose a problem for some log-rank tests. Table 3.8 reports p-values obtained using the same 13 tests described in Section 3.4.2.

The proposed fiducial test gives the smallest p-value of 0·002. To explain why the fiducial approach works on this dataset, we plot the sample of the difference of two fiducial distributions in Figure 3.6b. If these two datasets are from the same distribution, 0 should be well within the sample curves. However, from the picture, we could see that the majority of curves are very far

Table 3.8: p-value of different tests (in %)

	F	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
p	0·2	63·5	4·6	16·8	4·6	4·3	90·6	5·6	0·6	1·5	0·6	0·6	22·8

Table 3.9: Percentage of p-value less than 0·05 (%)

F	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
87·4	10·2	53·4	31·0	53·0	53·4	7·6	57·6	84·6	78·4	84·6	84·6	23·2

away from 0 on the interval $[0, 1]$.

In order to study the power of our test in this situation, we present a simulation study. We use the data to estimate the failure and censoring distribution for both datasets. Then we use these estimated distributions as truth to generate 500 synthetic datasets that mimic our data. On each dataset, we perform the proposed fiducial test with fiducial sample size 1000 and the 12 different types of log-rank tests. Table 3.9 shows the percentage of p-value less than 0·05. We see that the proposed fiducial test has the best power.

3.6 Discussion

In this paper we derived a nonparametric generalized fiducial distribution for right censored data. This GFD provided us with a unified framework for deriving statistical procedures such as pointwise and curvewise approximate confidence intervals and tests. This is to our knowledge the first time the fiducial distribution has been derived for a non-trivial nonparametric model. We proved a functional Bernstein-von Mises theorem which established the asymptotic correctness of the inference procedures based on our GFD. Additionally, our simulation studies suggest that our GFD inference procedures are as good and in some instances better than the many other statistical procedures proposed for the various aspects of this classical problem. Overall, we view generalized fiducial inference in a similar way as maximum likelihood, as a general purpose approach that provides good quality answers to many statistical problems. As we can see in the paper, the proposed point estimator of survival function is very similar to Kaplan-Meier estimator. However, the strength of the fiducial approach is in uncertainty quantification when the sample size is small. In particular, we recommend using proposed fiducial confidence intervals and tests in the small sample or heavy censoring cases.

We conclude by listing some open research problems:

1. We chose to use the sup-norm in the definition of the curvewise confidence intervals and tests. It could be possible to make the procedure somewhat more powerful by using a different (possibly weighted) norm [97]. Similarly, it might be also possible to use the choice of norm for tuning the GFD tests for use against specific alternatives.

2. The proposed fiducial test seems to be relatively powerful against a broad spectrum of alternatives. It would be interesting to implement it inside other statistical procedures where log-rank tests are recursively used, such as imputed survival random forests and their applications [141, 33, 32].

3. There seems to be an intriguing connection between GFD and empirical likelihood for semiparametric models [109, Chapter 11]. To investigate this connection should make for a fruitful avenue of future research.

CHAPTER 4

Tree based weighted learning for estimating ITRs with censored data

4.1 Introduction

An individualized treatment regime provides a personalized treatment strategy for each patient in the population based on their individual characteristics. A significant amount of work has been devoted to estimating optimal treatment rules [96, 104, 133, 136, 135]. While each of these approaches has strengths and weaknesses, we highlight the approach in [135] because of its robustness to model misspecification (this is similarly true of the approach in [133]) combined with its ability to incorporate support vector machines through the recognition that optimizing the treatment rule can be recast as a weighted classification problem. This approach is commonly referred to as outcome weighted learning. In clinical trials, right censored survival data are frequently observed as primary outcomes. Adapting outcome weighted learning to the censored setting, [138] proposed two new approaches, inverse censoring weighted outcome weighted learning and doubly robust outcome weighted learning, both of which require semiparametric estimation of the conditional censoring probability given the patient characteristics and treatment choice. The doubly robust estimator additionally involves semiparametric estimation of the conditional failure time expectation but only requires that one of the two models, for either the failure time or censoring time, be correct. Potential drawbacks of these methods are that either or both models may be misspecified and inverse censoring weighting estimation can be unstable numerically [104, 141].

In this chapter, we propose a nonparametric tree based approach for right censored outcome weighted learning which avoids both the inverse probability of censoring weighting and restrictive modeling assumptions for imputation through recursively imputed survival trees [141]. Since the true failure times T are only partially known, they cannot be used directly as weights in the outcome weighted learning [135] framework. However, recursively imputed survival trees [141] provide an alternative approach to weighting by using the conditional expectations of censored observations

without requiring inverse weighting. Tree-based methods [21, 19] are a broad class of nonparametric estimators which have become some of the most popular machine learning tools. Its adaptation to the survival setting has also drawn a lot of interests in the literature [81, 70, 73], and it has also been used for interpretable prediction modeling in personalized medicine [78]. The recursively imputed survival tree approach [141] combines extremely randomized trees with a recursive imputation method, which has been shown to improve performance and reduce prediction error while avoiding estimation of inverse censoring weights without making parametric or semiparametric assumptions on the conditional probability distribution of the failure time. Numerical studies demonstrate that the proposed method outperforms existing alternatives in a variety of settings.

The proposed method uses these recursively imputed survival trees to impute the survival times nonparametrically in a manner suitable for implementation within outcome weighted learning. We verify this novel approach both theoretically and in numerical examples. As part of this, we also present for the first time consistency and rate results for tree-based survival models in a more general setting than the categorical predictors considered in [72].

The remainder of the article is organized as follows. In section 4.2, we present the mathematical framework for individualized treatment rules for right censored survival outcomes. In section 4.3 we establish consistency and an excess value bound for the estimated treatment rules. Extensive simulation studies are presented in Section 4.4. We also illustrate our method using a phase III clinical trial on non-small cell lung cancer in Section 4.5. The article concludes with a discussion of future work in Section 4.6. Some needed technical results are provided in the Appendix.

4.2 Methodology

4.2.1 Individualized treatment regime framework

Before characterizing the individualized treatment regime, we first introduce some general notation and introduce the value function, and then extend the notation and ideas to the censored data setting. Let $X \in \mathcal{X}$ be the observed patient-level covariate vector, where $\mathcal{X}$ is a d dimensional vector space, and let $A \in \{-1, +1\}$ be the binary treatment indicator. $\widetilde{T}$ is the true survival time, however, we consider a truncated version at τ, i.e., $T = \min(\widetilde{T}, \tau)$, where the maximum follow-up time $\tau < \infty$ is a common practical restriction in clinical studies. The goal in this framework is to

maximize a reward R, which could represent any clinical outcome. Specifically, we wish to identify a treatment rule $\mathcal{D}$, which is a map from the patient-level covariate space $\mathcal{X}$ to the treatment space $\{+1, -1\}$ which maximizes the expected reward. In the survival outcome setting, we use $R = T$ or $\log(T)$ as done in [138].

To achieve this maximization, we define the value function as

$$V(\mathcal{D}) = E^{\mathcal{D}}(R) = E\big[RI\{A = \mathcal{D}(X)\}/\pi(A; X)\big],$$

where $I\{\cdot\}$ is an indicator function, $\pi(a; X) = \text{pr}(A = a \mid X) > M'$ $a.s.$ for some $M' > 0$ and each $a \in \{+1, -1\}$. The function π is the propensity score and is known in a randomized trial setting, which we assume is the case for this chapter, but needs to be estimated in a non-randomized, observational study setting. The individualized treatment regime we are most interested in is the optimal treatment rule $\mathcal{D}^*$ which maximizes the value function, i.e.

$$\mathcal{D}^* = \arg\max_{\mathcal{D}} E\big[RI\{A = \mathcal{D}(X)\}/\pi(A; X)\big]. \tag{4.1}$$

After rewriting the value function as

$$V(\mathcal{D}) = E\big[E(R \mid A = 1, X)I\{\mathcal{D}(X) = 1\} + E(R \mid A = -1, X)I\{\mathcal{D}(X) = -1\}\big],$$

it is easy to see that

$$\mathcal{D}^* = \text{sign}\big\{E(R \mid A = 1, X) - E(R \mid A = -1, X)\big\}.$$

Hence, the definition of $\mathcal{D}^*$ is equivalent to $\mathcal{D}^*(x) = \arg\max_a E(R \mid A = a, X = x)$. Instead of maximization the objective function in (4.1), the outcome weighted learning approach searches for the optimal decision rule $\mathcal{D}^*$ by minimizing the weighted misclassification error, i.e.,

$$\mathcal{D}^* = \arg\min_{\mathcal{D}} E\big[RI\{A \neq \mathcal{D}(X)\}/\pi(A; X)\big]. \tag{4.2}$$

In an ideal situation, we would replace R with T or $\log(T)$. However, this is not possible under right censoring.

4.2.2 Value function under right censoring

Consider a censoring time C that is independent of T given (X, A). We then have the observed time $Y = \min(T, C)$, and the censoring indicator $\delta = I(T \leq C)$. Assume that n independent and identically distributed copies, $\{Y_i, \delta_i, X_i, A_i\}_{i=1}^n$, are collected. Since T is not fully observed we seek for a sensible replacement which maintains as close as possible the same value function. We propose two approaches in the following, denoted as R_1 and R_2 respectively. The first approach is to obtain a nonparametric estimated conditional expectation $\widehat{E}(T \mid X, A)$. Letting $R_1 = E(T \mid X, A)$ and bringing the expectation of T inside, we have

$$E\big[TI\{A = \mathcal{D}(X)\}/\pi(A; X)\big] = E\big[R_1 I\{A = \mathcal{D}(X)\}/\pi(A; X)\big]. \tag{4.3}$$

Another approach is to replace only the censored observations conditioning on the observed data. It is interesting to observe that the conditional expectation of T, given Y and δ, can be written as

$$\begin{aligned} R_2 :=& E(T \mid X, A, Y, \delta) \\ =& I(\delta = 1)Y + I(\delta = 0)E(T \mid X, A, Y, \delta = 0) \\ =& I(\delta = 1)Y + I(\delta = 0)E(T \mid X, A, C = Y, T > Y, Y) \\ =& I(\delta = 1)Y + I(\delta = 0)E(T \mid X, A, T > Y, Y). \end{aligned} \tag{4.4}$$

An important property that we used in the last equality is the conditional independence between T and C. With the information of $Y = y$ given, and knowing that $\delta = 0$, the conditional distribution of T is defined on $(c, \tau]$ with density function proportional to the original density of T. In other words, the conditional survival function of T is $S(t \mid X, A)/S(c \mid X, A)$ for $t > c$, where $S(\cdot \mid X, A)$ is the conditional survival function of T. Hence, we can calculate the expectation of T accordingly. With the definition of R_2, it is easy to see that the corresponding value function is equivalent to the left side of equation (4.3) by further taking expectations with respect to Y and δ. Note that the above arguments remain unchanged if we replace T, C and Y with $\log(T)$, $\log(C)$, and

$\log(Y)$, respectively: this equivalence will be tacitly utilized throughout the chapter, except when the distinction is needed.

With our proposed two reward measures, the remaining challenge is to nonparametrically estimate the conditional expectations. To this end, we utilize the nonparametric tree based method proposed by [141]. It is worth noting that the conditional expectation of T defined in R_2 shares the same logical underpinnings as the imputation step in [141]. However, the goal of the imputation step is to replace the censored observations with a randomly generated conditional failure time which utilizes the same condition survival distribution of T given $T > C$. We will provide details of the estimation procedure in the next section. To conclude this section, we provide the empirical versions of the value function using the two rewards R_1 and R_2, respectively, which we solve for the optimal decision $\mathcal{D}^*$ by minimization:

$$n^{-1}\sum_{i=1}^{n}\frac{\widehat{E}(T_i \mid A_i, X_i)I\{A_i = \mathcal{D}(X_i)\}}{\pi(A_i; X_i)}, \tag{4.5}$$

$$\text{and } \; n^{-1}\sum_{i=1}^{n}\frac{\{\delta_i Y_i + (1-\delta_i)\widehat{E}(T_i \mid X_i, A_i, T_i > Y_i, Y_i)\}I\{A_i = \mathcal{D}(X_i)\}}{\pi(A_i; X_i)}. \tag{4.6}$$

4.2.3 Outcome weighted learning with survival trees

The recursively imputed survival trees method proposed by [141] is a powerful tool to estimate conditional survival functions for censored data. A brief outline of the algorithm is provided in the following. We refer interested readers to the original paper for details. To fit the model, we first generate extremely randomized survival trees for the training dataset. Secondly, we calculate conditional survival functions for each censored observation, which can be used for imputing the censored value to a random conditional failure time. Thirdly, we generate multiple copies of the imputed dataset, and one survival tree is fitted for each dataset. We repeat the last two steps recursively and the final nonparametric estimate of $\widehat{E}(T \mid X, A)$ is obtained by averaging the trees from the last step.

Following [135], we next use support vector machines to solve for the optimal treatment rule. A decision function $f(x)$ is learned by replacing $I\{A_i = \mathcal{D}(X_i)\}$ in Equations (4.5) or (4.6) with $\phi\{A_i f(X_i)\}$, where $\phi(x) = (1-x)^+$ is the hinge loss and $x^+ = \max(x, 0)$. Furthermore, to avoid

overfitting, a regularization term $\lambda_n \|f\|^2$ is added to penalize the complexity of the estimated decision function f. Here, $\|f\|$ is some norm of f , and λ_n is a tuning parameter. A high-level description of the proposed method is given in Algorithm 3 below. We consider both linear and nonlinear decision functions f when solving (4.7). For a linear decision function, $f(x) = \theta_0 + \theta^T x$ and we let $\|f\|$ be the Euclidean norm of θ. For nonlinear decision functions, we employ a universal kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such as the Gaussian kernel, which is continuous, symmetric and positive semidefinite. The optimization problem is then equivalent to a dual problem that maximizes

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j A_i A_j k(X_i, X_j),$$

subject to $0 \leq \alpha_i \leq \gamma W_i / \pi_i$ and $\sum_{i=1}^{n} \alpha_i A_i = 0$, where W_i is the numerator in either (4.5) or (4.6) and π_i is the respective denominator. Both settings can be efficiently solved by quadratic programming. For further details regarding solving weighted classification problems using support vector machines, we refer to [135, 138, 25].

Algorithm 3: Pseudo algorithm for the proposed method

Step 1. Use $\{(X_i^{\mathrm{T}}, A_i, A_i X_i^{\mathrm{T}})^{\mathrm{T}}, Y_i, \delta_i\}_{i=1}^n$ to fit recursively imputed survival trees. Obtain the estimation $\widehat{E}(T_i \mid A_i, X_i)$ for reward R_1 or the estimation $\widehat{E}(T_i \mid X_i, A_i, T_i > Y_i, Y_i)$ for reward R_2.

Step 2. Let the weights W_i be either $\widehat{E}(T_i \mid A_i, X_i)$ or $\delta_i Y_i + (1 - \delta_i)\widehat{E}(T_i \mid A_i, X_i, T_i > Y_i, Y_i)$, depending on which of the two proposed approaches is used. Minimize the following weighted misclassification error:

$$\widehat{f}(x) = \arg\min_{f} \sum_{i=1}^{n} W_i \frac{\phi\{A_i f(X_i)\}}{\pi(A_i; X_i)} + \lambda_n \|f\|^2. \tag{4.7}$$

Step 3. Output the estimated optimal treatment rule $\widehat{\mathcal{D}}(x) = \mathrm{sign}\{\widehat{f}(x)\}$.

4.3 Theoretical results

4.3.1 Preliminaries

The risk function is defined as

$$R(f) = E\Big[\frac{R}{\pi(A;X)} I\{A \neq \text{sign}(f(X))\}\Big],$$

where the reward $R = R_1 = E(T \mid X, A)$ for the first approach, or $R = R_2 = \delta Y + (1-\delta)E(T \mid X, A, T > Y, Y)$ for the second one. We define ϕ-risk for both the true and the working model as, respectively, $R_\phi(f) = E[R\phi\{Af(X)\}/\pi(A;X)]$ and $R'_\phi(f) = E[\widehat{R}\phi\{Af(X)\}/\pi(A;X)]$, where $\widehat{R}$ is the estimated value of R based on one of the two proposed methods. We also define the hinge loss function for the true and working models as $L_\phi(f) = R\phi\{Af(X)\}/\pi(A;X)$ and $L'_\phi(f) = \widehat{R}\phi\{Af(X)\}/\pi(A;X)$, respectively.

The proposed estimator $\widehat{\mathcal{D}} = \text{sign}(\widehat{f_n}(X))$, where $\widehat{f_n}$ is solved by one of the following optimization problems within some reproducible kernel Hilbert space $\mathcal{H}_k$:

$$\widehat{f_n} = \underset{f \in \mathcal{H}_k}{\arg\min}\ n^{-1} \sum_{i=1}^{n} \frac{\widehat{E}(T_i \mid X_i, A_i)}{\pi(A_i; X_i)} \phi\{f(X_i)A_i\} + \lambda_n ||f||_k^2,$$

or

$$\widehat{f_n} = \underset{f \in \mathcal{H}_k}{\arg\min}\ n^{-1} \sum_{i=1}^{n} \frac{\delta_i Y_i + (1-\delta_i)\widehat{E}(T_i \mid X_i, A_i, T_i > Y_i, Y_i)}{\pi(A_i; X_i)} \phi\{f(X_i)A_i\} + \lambda_n ||f||_k^2.$$

4.3.2 Consistency of tree-based survival models

In this section, we provide the convergence bound of a simplified tree-based survival model, which is very close to the original algorithm in [141]. The purpose of this section and its main result, Theorem 8, is to demonstrate the existence of an accurate estimator of the underlying hazard function when tree-based methods are used. An earlier result developed in [72] considers only categorical feature variables. To the best of our knowledge, what we present below is the first

consistency result for a tree-based survival model under general settings with restrictions only on the splitting rules, which is interesting in its own right.

For simplicity, we assume in this section that $\mathcal{Q}_n = \{(Y_i, \delta_i, X_i, A_i), i = 1, \ldots, n\}$ is the training sample, where X_i is independent uniformly distributed on $[0,1]^d$. The result can be easily generated to distributions with bounded support and density function bounded above and below. For any fixed X, our goal is to estimate the cumulative hazard function of failure time $r(\cdot, X, A) = \Lambda_T(\cdot \mid X, A)$; hereinafter, we write it as $\Lambda(\cdot \mid X, A)$.

A random forest is a collection of randomized regression trees $\{\widehat{r}_n(\cdot, X, A,$
$\Theta_j, \mathcal{Q}_n), 1 \leq j \leq m\}$, where m is the number of trees. The randomizing variable Θ is used to indicate how the successive cuts are performed when an individual tree is built. Hence the forest version of the survival tree model can be expressed as

$$\widehat{r}_n(\cdot, X, A, \mathcal{Q}_n) = \frac{1}{m} \sum_{j=1}^{m} \widehat{r}_n(\cdot, X, A, \Theta_j, \mathcal{Q}_n).$$

Here, we consider a simplified scenario in which the selection of the coordinate is completely random and independent from the training data [11]. We only consider the consistency of a single tree and denote our tree estimator as $\widehat{r}_n(\cdot, X, A)$. The result can be easily extended to the situation where m is finite.

A brief description of how each individual tree is constructed is provided in the appendix. Here we highlight some key assumptions and the main result. Our first assumption puts a lower bound on the probability of observing a failure at τ, and the second one assumes the smoothness of the hazard and cumulative hazard functions.

Assumption 10. *For some $M > 0$, $S_Y(\tau \mid X, A) > M$ almost surely.*

Assumption 11. *For any fixed time point t and treatment decision A, the cumulative hazard function $\Lambda(t \mid X, A)$ is L-Lipschitz continuous in terms of X, and the hazard function $\lambda(t \mid X, A)$ is L'-Lipschitz continuous in terms of X, i.e., $|\Lambda(t \mid X_1, A) - \Lambda(t \mid X_2, A)| \leq L||X_1 - X_2||$ and $|\lambda(t \mid X_1, A) - \lambda(t \mid X_2, A)| \leq L'||X_1 - X_2||$, respectively, where $||\cdot||$ is the Euclidean norm.*

The following theorem provides the bound of the proposed tree based survival model for each X. Details of the proof are collected in the Appendix.

Theorem 8. *Assume that Assumptions 10–11 and the construction of a tree-based survival model described in the Appendix. Further assume that $k_n \to \infty$ and $n/k_n \to \infty$ as $n \to \infty$, where k_n is a deterministic parameter which we can control (each individual tree has approximately k_n terminal nodes). We have for each X,*

$$pr\Big\{ \sup_{t<\tau} |\widehat{r}_n(t,X,A) - r(t,X,A)| \le C[d^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + b^{1/2}\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\}^{-1/2}]\Big\} \ge 1 - w_n,$$

where $r, u \in (0,1)$, $b \ge 1$, C is some universal constant and

$$w_n = \ 16[(1-u)n2^{-\lceil \log_2 k_n \rceil} + 2]e^{-b} + e^{-u^2 n 2^{-\lceil \log_2 k_n \rceil - 1}} + de^{-\lceil \log_2 k_n \rceil r^2/(2d)}.$$

The ideal balance happens when $k_n = n^{d/(d+2)}$. In this case, the optimal rate of the bound is close to $n^{-1/(d+2)}$. The following theorem proves consistency of the proposed tree based survival model. Details of the proof are collected in the Appendix.

Theorem 9. *Assume that Assumptions 10–11 and the construction of a tree-based survival model described in the Appendix. Further assume that $k_n = n^{\eta}$, where $0 < \eta < 1$. Then the estimator of the survival tree model is consistent. Moreover,*

$$\sup_{t<\tau} E_X|\widehat{r}_n(t,X,A) - r(t,X,A)| \le C[d^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + b^{1/2}\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\}^{-1/2} + w_n \ln(n)],$$

where $r, u \in (0,1)$, $b \ge 1$, C is some universal constant and

$$w_n = \ 16[(1-u)n2^{-\lceil \log_2 k_n \rceil} + 2]e^{-b} + e^{-u^2 n 2^{-\lceil \log_2 k_n \rceil - 1}} + de^{-\lceil \log_2 k_n \rceil r^2/(2d)}.$$

4.3.3 Consistency and excess value bound

Fisher consistency follows directly from Proposition 3.1 in [135], hence the proof is omitted. Here we restate the result as the following lemma. For the proposed method, we simply replace the

reward R in $R_\phi(f)$ with R_1 or R_2. Note that both versions are equivalent to the reward function $R_\phi(f) = E[T\phi\{Af(X)\}/\pi(A;X)]$:

Lemma 5 (Proposition 3.1 in [135]). *For any measurable function $\widetilde{f}$, if $\widetilde{f}$ minimizes $R_\phi(f)$, then $\mathcal{D}^*(x) = \text{sign}(\widetilde{f}(x))$.*

Provided the Assumptions in Section 4.3.2 hold, the following lemma ensures the convergence of the estimated conditional expectations. The proof is given in Appendix.

Lemma 6. *Based on Theorem 8, for each X the estimated conditional expectations converge in probability, i.e.,*

$$pr\Big\{\big|\widehat{E}(T \mid X, A) - E(T \mid X, A)\big| \\ \leq C_1[2^{-\{(1-r)\lceil \log_2 k_n\rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n\rceil}\})^{1/2}]\Big\} \geq 1 - w_n,$$

$$pr\Big\{\big|\widehat{E}(T \mid X, A, T > Y, Y) - E(T \mid X, A, T > Y, Y)\big| \\ \leq C_2[2^{-\{(1-r)\lceil \log_2 k_n\rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n\rceil}\})^{1/2}]\Big\} \geq 1 - 2w_n,$$

for some constant C_1, C_2 (depending on L, L', τ, M, d).

We will use the above lemmas to prove our main theorem based on the Gaussian kernel. Before we derive the convergence rate and excess value bound, we define the value function corresponding to the true and working model as $V(f) = E(RI[A = \text{sign}\{f(X)\}]/\pi(A;X))$, $V'(f) = E(\widehat{R}I[A = \text{sign}\{f(X)\}]/\pi(A;X))$, respectively. We further define the empirical L_2–norm, $\|f - g\|_{L_2(P_n)} = (\sum_{i=1}^n [f(X_i) - g(X_i)]^2/n)^{1/2}$, which also defines an ϵ-ball based on this norm. By Theorem 2.1 in [116], we restate the bound for covering numbers:

Lemma 7 (Theorem 2.1 in [116]). *For any $\beta > 0$, $0 < v \leq 2$, $\epsilon > 0$ we have $\sup_{P_n} \log N(B_{\mathcal{H}_k}, \epsilon, L_2(P_n)) \leq c_{v,\beta,d}\sigma_n^{(1-v/2)(1+\beta)d}\epsilon^{-v}$, where $B_{\mathcal{H}_k}$ is the closed unit ball of $\mathcal{H}_k$, σ_n is the kernel bandwidth, and d is the dimension of $\mathcal{X}$.*

Lastly, for $\widetilde{f} = \arg\min_{f\in\mathcal{F}} E\{L_\phi(f)\}$, we define the approximation error function

$$a(\lambda) = \inf_{f\in\mathcal{H}_k} [E\{L_\phi(f)\} + \lambda||f||_k^2 - E\{L_\phi(\widetilde{f})\}].$$

Then we have following theorem, the proof of which is given in Appendix.

Theorem 10. *Based on Theorem 9 and assuming that the sequence $\lambda_n > 0$ satisfies $\lambda_n \to 0$ and $\lambda_n \ln n \to \infty$, we have that*

$$pr(V(f^*) \le V(\widehat{f}_n) + \epsilon) \ge 1 - 2e^{-\rho},$$

*where f^* maximize the true value function V, $\epsilon = a(\lambda_n) + M_v(n\lambda_n/c_n)^{-2/(v+2)} + M_v\lambda_n^{-1/2}(c_n/n)^{2/(d+2)} + K\rho(n\lambda_n)^{-1} + 2K\rho n^{-1}\lambda_n^{-1/2} + C\lambda_n^{-1/2}\{2^{-(1-r)\lceil\log_2 k_n\rceil/d}$
$+ (b/\{(1-u)n2^{-\lceil\log_2 k_n\rceil}\})^{1/2} + w_n \ln n\}$, $c_n = c_{v,\beta,d}\sigma_n^{(1-v/2)(1+\beta)d}$ and $\rho > 0$ for both methods; also, M_v is a constant depending on v, K is a sufficiently large positive constant and C is a some large constant depending on d.*

The rate consists of two parts. The first part is from the approximation error using $\mathcal{H}_k$. The second part controls the approximation error due to using the proposed tree-based method to estimate the conditional expectation.

4.4 Simulation studies

We perform simulation studies to compare the proposed method with existing alternatives, including the Cox proportional hazards model with covariate-treatment interactions, inverse censoring weighted outcome weighted learning, and doubly robust learning, both proposed in [138]. We use survival time on the log scale $\log(T)$ as outcome. We also present for comparison an "oracle" approach which uses the true failure time on the log scale $\log(T)$ as the weight in outcome weighted learning, although this would not be implementable in practice. However, this approach is a representation of the best possible performance under the outcome weighted learning framework.

We generate X_i's independently from a uniform distribution. Treatments are generated from $\{+1, -1\}$ with equal probabilities. We present four scenarios in this simulation study. The failure

time T and censoring time C are generated differently in each scenario, including both linear and nonlinear decision rules. For each case, we learn the optimal treatment rule from a training dataset with sample size $n = 200$. A testing dataset with size 10000 is used to calculate the value function under the estimated rule. Each simulation is repeated 500 times.

Tuning parameters in the tree based methods need to be selected. We mostly use the default values. The number of variables considered at each split is the integer part of the square root of d as suggested by [73] and [55]. We set the total number of trees to be 50 as suggested by [141] and use one fold imputation. For the alternative approaches such as inverse censoring weighted outcome weighted learning and doubly robust learning, a Cox proportional hazards model with covariates (X, A, XA) is used to model T and C respectively. Note that when at least one of the two working models is correctly specified, the doubly robust method enjoys consistency. We implemented outcome weighted learning using a Matlab library for support vector machine [25]. Both linear and Gaussian kernels are considered for all methods except for the Cox model approach which could be directly inverted to obtain the decision rules. The parameter λ_n is chosen by ten-fold cross-validation.

4.4.1 Simulation settings

For all scenarios, we generate $\widetilde{T}$ and C independently. The failure time $T = \min(\tau, \widetilde{T})$. For all accelerated failure time models, ϵ is generated from a standard normal distribution. For all Cox proportional hazards models, the baseline hazard function $\lambda_0(t) = 2t$. For all simulation results presented in this section, we consider setting the censoring rates to approximately 45% for all scenarios. We also perform a sensitivity analysis for different censoring rates (30% and 60%) for each scenario. These additional results are presented in the Appendix.

Scenario 1. Both $\widetilde{T}$ and C are generated from the accelerated failure time model. $\tau = 2.5$ and $d = 10$. The optimal decision function is linear. The value of the optimal treatment rule is approximately 0.031:

$$\begin{aligned}
\log(\widetilde{T}) =& -0.2 - 0.5X_1 + 0.5X_2 + 0.3X_3 \\
&+ (0.5 - 0.1X_1 - 0.6X_2 + 0.1X_3)A + \epsilon, \\
\log(C) =& 0.1 - 0.8X_1 + 0.4X_2 + 0.4X_3 + (0.5 - 0.1X_1 - 0.6X_2 + 0.3X_3)A + \epsilon.
\end{aligned}$$

Scenario 2. $\widetilde{T}$ is generated from a Cox model and C is generated from the accelerated failure time model. The optimal decision function is nonlinear. $\tau = 8$ and $d = 10$. The value of the optimal treatment rule is approximately 0.181:

$$\begin{aligned}
\lambda_{\widetilde{T}}(t \mid A, X) =& \lambda_0(t) \exp\{-0.2 - 1.5X_1^{1.5} + 0.5X_2 + (0.8 - 0.7X_1^{0.5} - 1.2X_2^2)A\}, \\
\log(C) =& -0.5 + 0.7X_1 + X_2^2 + 0.6X_3 + 0.1X_4 \\
&+ (0.2 + X_1^{2.5} - 2X_2 + 0.5X_3)A + \epsilon.
\end{aligned}$$

Scenario 3. $\widetilde{T}$ is generated from an accelerated failure time model with tree structured effects. C is generated from a Cox model with nonlinear effects. $\tau = 8$ and $d = 5$. The value of the optimal treatment rule is approximately 1.079:

$$\begin{aligned}
\log(\widetilde{T}) =& X_1 + I(X_2 > 0.5)I(X_3 > 0.5) + (0.3 - X_1)A \\
&+ 2\{I(X_4 < 0.3)I(X_5 < 0.3) + I(X_4 > 0.7)I(X_5 > 0.7)\}A + \epsilon, \\
\lambda_C(t \mid A, X) =& \lambda_0(t) \exp\{-1.5 + X_1 + (1 + 0.6X_2^{1.5})A\}.
\end{aligned}$$

Scenario 4. $\widetilde{T}$ is is generated from an accelerated failure time model. C is generated from a Cox model. $\tau = 2$ and $d = 10$. The value of the optimal treatment rule is approximately -0.389:

$$\begin{aligned}
\log(\widetilde{T}) =& -0.5 - 0.8X_1 + 0.7X_2 + 0.2X_3 \\
&+ (0.6 - 0.4X_1 - 0.2X_2 - 0.4X_3)A + \epsilon, \\
\lambda_C(t \mid A, X) =& \lambda_0(t) \exp\{-0.5X_1 - 0.5X_2 + 0.2X_3 \\
&- (1 - 0.5X_1 + 0.3X_2 - 0.5X_3)A\}.
\end{aligned}$$

Table 4.1: Simulation results: Mean ($\times 10^3$) and (sd) ($\times 10^3$). Censoring rate: 45%. For each scenario, the theoretical optimal value ($\times 10^3$) is 31, 181, 1079, and -389, respectively.

	kernel	T	RIST-R_1	RIST-R_2	ICO	DR	Cox	
1	Linear	0 (26)	0 (31)	1 (30)	-20 (54)	-39 (76)	-29 (33)	
	Gaussian	-17 (44)	-11 (35)	-8 (36)	-25 (50)	-88 (79)		
2	Linear	22 (113)	-1 (112)	-24 (125)	-137 (131)	-232 (132)	53 (69)	T: using
	Gaussian	-39 (115)	-40 (103)	-72 (114)	-175 (120)	-311 (106)		
3	Linear	785 (52)	766 (59)	763 (51)	683 (113)	598 (120)	745 (64)	
	Gaussian	896 (61)	803 (56)	834 (71)	785 (105)	606 (115)		
4	Linear	-453 (37)	-469 (47)	-451 (27)	-469 (48)	-481 (59)	-464 (36)	
	Gaussian	-465 (35)	-482 (44)	-457 (28)	-487 (45)	-531 (43)		

true survival time as weight; RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning; Cox: Cox proportional hazards model using covariate-treatment interactions.

4.4.2 Simulation results

Figure 4.1 shows the boxplot of values based on the logarithm of T calculated from the test data. The mean and standard deviation of values are shown in Table 4.1. In scenario 1, since the model is not correctly specified for inverse probability of censoring outcome weighted learning, the doubly robust estimator, or Cox regression, our method performs better than all other competitors.

In scenario 2, we added some nonlinear terms into both the Cox and accelerated failure time models. The model assumptions for inverse censoring outcome weighted learning and the doubly robust estimator are not satisfied. Our estimated treatment rule performs much better than these two. Compared with inverse censoring outcome weighted learning and doubly robust learning, both our approaches improve more than 0.1 for the mean. Since the true model for the failure time is the Cox model, Cox regression performs better here. In this case, the Gaussian kernel performs less well than the linear kernel for most methods since the true model structure is linear and the Gaussian kernel is too flexible.

For scenario 3, which has a more complicated tree structure, the Gaussian kernel performs better than the linear kernel for all outcome weighted learning approaches. The performance of the Gaussian kernel is enhanced since it can better address the true nonlinear model structure. We can see that with either a linear or Gaussian kernel, our estimators perform better than Cox regression. Compared with doubly robust learning, our two approaches improve 0.2 for the mean.

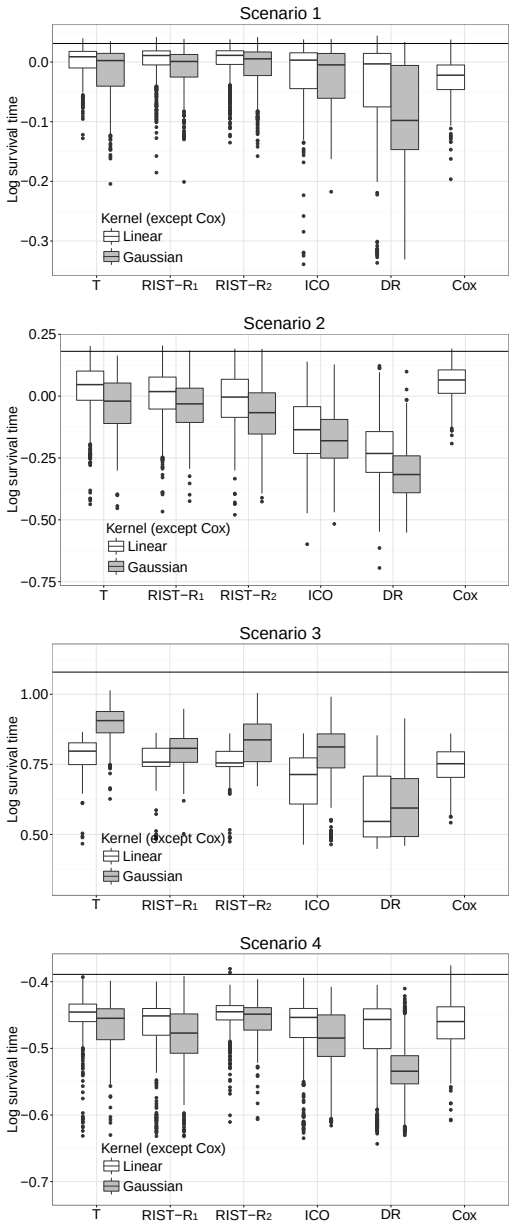


Figure 4.1: Boxplots of mean log survival time for different treatment regimes. Censoring rate: 45%. T: using true survival time as weight; RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning. The black horizontal line is the theoretical optimal value.

In scenario 4, we see that when the model is correctly specified for inverse probability of censoring outcome weighted learning and doubly robust learning, the performances of both approaches are satisfactory while our methods seem to be only a little better. The performances of our first approach, inverse probability of censoring outcome weighted learning and Cox regression are all similar. Our second approach has the best treatment effect among all estimators. Note that our second approach appears to perform as well as the first, oracle approach. Also, our two proposed methods have smaller standard errors in scenarios 1 and 3. The standard error is similar for all outcome weighted learning approaches in scenario 2 and 4. Overall, our proposed methods have generally lower variances.

Compared with results of censoring rates (30% and 60%) in the Appendix, we can observed a consistently pattern that lower censoring rate leads to higher performances in terms of both mean value and variance. The relative performances between the proposed and the competing methods remain similar across different censoring rates.

4.5 Data analysis

We apply the proposed method to a non-small-cell lung cancer randomized trial dataset described in [113]. 228 subjects with complete information are used in this analysis. Each treatment arm contains 114 subjects. Here we use five covariates: performance status (119 subjects ranging from 90% to 100% and 109 subjects ranging from 70% to 80%), cancer stage (31 subjects in stage 3 and 197 subjects in stage 4), race (167 white, 54 black and 7 others), gender (143 male and 85 female), age (ranging from 31 to 82 with median 63). The length of study is $\tau = 104$ weeks. We adopt the same tuning parameters used in the simulation study for this analysis. The value function is again calculated by using the logarithm of survival time $\log(T)$ (in weeks) as the reward.

We randomly divide the 228 patients into four equal proportions and use three parts as training data to estimate the optimal rule and calculate the empirical value based on the remaining part. We then permute the training and testing portions and average the four results. This procedure is then repeated 100 times and averaged to obtain the mean and standard deviation. To calculate the testing data performance, we consider two different measurements, both are calculated based on the formula $\sum_{i=1}^{n} R_i I\{A_i = \mathcal{D}(X_i)\} / \sum_{i=1}^{n} I\{A_i = \mathcal{D}(X_i)\}$ for the testing samples, where two

Table 4.2: Analysis of non-small-cell lung cancer data: Mean (sd) of value function

kernel	RIST-R_1	RIST-R_2	ICO	DR	Cox	
Linear	3.641 (0.144)	3.641 (0.138)	3.633 (0.158)	3.590 (0.174)	3.582 (0.158)	RIST-
Gaussian	3.611 (0.215)	3.615 (0.220)	3.302 (0.221)	3.470 (0.233)		

R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning; Cox: Cox proportional hazards model using covariate-treatment interactions.

versions of R_i's are used. We first consider the procedure proposed in [138], where R is defined as

$$\frac{\Delta Y}{\widehat{S}_C(Y \mid A, X)} - \int \widehat{E}_{\widetilde{T}}\{T \mid T > t, A, X\} \left\{ \frac{dN_C(t)}{\widehat{S}_C(t \mid A, X)} + I(Y_i \geq t) \frac{d\widehat{S}_C(t \mid A, X)}{\widehat{S}_C(t \mid A, X)^2} \right\}.$$

Here, $\widehat{S}_C(t \mid A, X)$ and $\widehat{E}_{\widetilde{T}}(T \mid T > t, A, X)$ are estimated from the Cox model for simplicity. We also consider a more direct clinical measurement without the double robustness correction, which can be interpreted in a similar way as the expected survival time or the restricted mean survival time [53, 89, 124]. To be specific, we consider a restricted mean (log) survival time truncated at τ defined as $\delta T + (1 - \delta)E(T)$, and use this as a plug-in quantity of R in the testing performance calculation. To estimate this quantity, we use a recursively imputed survival trees (RIST) method to produce the expected survival time $E(T)$.

The value function results are presented in Table 4.2 and Figure 4.2. Both proposed methods have higher values than the compared methods. Note that for the Gaussian kernel, our two new approaches are still better than Cox regression, however, inverse probability of censoring outcome weighted learning and doubly robust learning are not much different from Cox regression. The standard error is comparable among all four methods using the linear kernel. For the Gaussian kernel, the standard errors of the proposed methods and inverse probability of censoring weighted learning are similar. The standard error for the doubly robust method is slightly worse in this instance. Overall, the proposed methods seem to perform best.

The restricted log mean results are presented in Table 4.3 and Figure 4.3. Note for the linear kernel, the median of the proposed methods are higher than 3.6 and median of both inverse probability of censoring outcome weighted learning and doubly robust learning are lower. For the Gaussian kernel, the proposed methods are much better than inverse probability of censoring outcome weight-

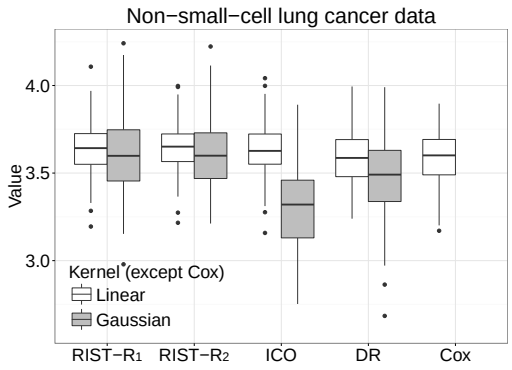


Figure 4.2: Boxplots of cross-validated value of survival weeks on the log scale. RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning.

Table 4.3: Analysis of non-small-cell lung cancer data: Mean (sd) of a clinical measure

kernel	RIST-R_1	RIST-R_2	ICO	DR	Cox
Linear	3.603 (0.040)	3.606 (0.037)	3.598 (0.037)	3.601 (0.042)	3.646 (0.039)
Gaussian	3.511 (0.064)	3.514 (0.068)	3.451 (0.062)	3.456 (0.052)	

RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning; Cox: Cox proportional hazards model using covariate-treatment interactions.

ed learning and doubly robust learning. Interestingly, under this measure, the performance of Cox regression is the best. A possible reason is that the true underlying model may not deviate much from the proportional hazard model, making the Cox model a better choice. This is also reflected by the fact that the results look similar to the simulation Scenario 2 plot, where the Cox model performs the best. Another possible reason is that the pseudo-outcome estimated from RIST may not be completely accurate and favors the Cox model in this particular dataset.

4.6 Discussion

We proposed a new method that redefines the reward function in a censored survival setting. The method works by replacing the censored observations (or all observations) by an estimated conditional expectation of the failure time. In practice, the failure time (or logarithm of the failure time) is commonly used in defining the reward function R, however, this choice could more flexible. For example, we may be interested in searching for a treatment rule that maximizes the median

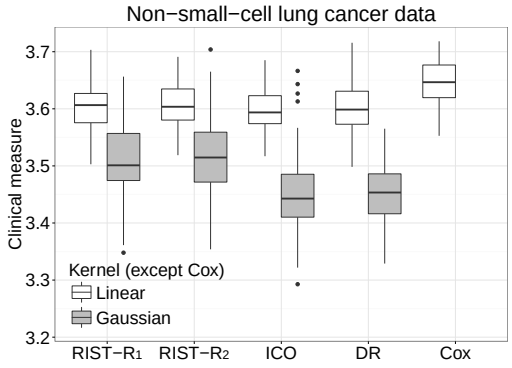


Figure 4.3: Boxplots of cross-validated value of survival weeks on the log scale. RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning.

survival time or a certain quantile. Under our framework, this is achievable by replacing the censored observations with a suitable estimate of the quantile. This part of the work is currently under investigation.

The proposed methods may be improved or extended in multiple ways. The estimated treatment rule may be affected by the shift of the outcome. A potential extension is to combine our methods with residual weighted learning [140], which has been shown to reduce the total variation of the weights and improve stability. Trials with multiple treatment arms occur frequently. Thus a potential extension of our method is in the direction of multicategory classification [18, 82]. It is also interesting to extend our method to dynamic treatment regimes where a sequence of decision rules [96, 136, 79, 137] need to be learned in a censored survival outcome setting [58].

APPENDIX A: SUPPLEMENTARY MATERIAL TO CHAPTER 2

The following table provides a summary of notations used in the proofs.

Basic Notations	
T	Failure time
C	Censoring time
Y	$= \min(T, C)$, observed time
δ	$= \mathbb{1}(T \leq C)$: censoring indicator
F_i, f_i	Survival distribution of i-th observation, $f_i = dF_i$
G_i	Censoring distribution of i-th observation
$\mathcal{A}$	A node, internal or terminal
$\boldsymbol{A}$	$= \{\mathcal{A}_u\}_{u \in \mathcal{U}}$, the collection of all terminal nodes in a single tree
$\Lambda(t\|x)$	Cumulative hazard function (CHF)
$\widehat{\Lambda}$, $\widehat{\Lambda}_{\mathcal{A}}$, $\widehat{\Lambda}_{\boldsymbol{A}}$	NA estimator on a set of samples, a node $\mathcal{A}$, or an entire tree $\boldsymbol{A}$
$\Lambda_n^*(t)$, $\Lambda_{\mathcal{A},n}^*$, $\Lambda_{\boldsymbol{A},n}^*$	Censoring contaminated averaged CHF on a set of samples, a node $\mathcal{A}$, or the entire tree $\boldsymbol{A}$
Λ^*, $\Lambda_{\mathcal{A}}^*$, $\Lambda_{\boldsymbol{A}}^*$	Population versions of Λ_n^*, $\Lambda_{\mathcal{A},n}^*$ and $\Lambda_{\boldsymbol{A},n}^*$, respectively
B	Number of trees in a forest
$d(d_0)$	Dimension of (important) covariates
τ	The positive constant as the upper bound of Y
Concentration Bounds	
k	Each terminal node contains at least k training examples, i.e., minimum leaf size.
α	Minimum proportion of observations contained in child node
$\mathcal{V}_{\alpha,k}(\mathcal{D})$	Set of all $\{\alpha, k\}$ valid partitions on the feature space $\mathcal{X}$
$\mathcal{H}_{\alpha,k}(\mathcal{D})$	Set of all $\{\alpha, k\}$ valid forests on the feature space $\mathcal{X}$
$\mathcal{R}$	Approximation node
$\boldsymbol{R}_{S,w,\epsilon}$, $\boldsymbol{R}$	The set of approximation nodes (rectangles)
$N(t)$	Counting process
$K(t)$	At-risk process
$\mu(\mathcal{R})$, $\mu(\mathcal{A})$	The expected fraction of training samples inside $\mathcal{R}, \mathcal{A}$
$\#\mathcal{R}$, $\#\mathcal{A}$	The number of training samples inside $\mathcal{R}, \mathcal{A}$
Consistency	
ψ_i	Proportion of length get of its parent node on the i-th dimension
L	Bound of the density function $f(t)$
L_1, L_2	Lipschitz constant of Λ and λ

Preliminary results

Proof of Lemma 1. For simplicity, we prove the results for the case when there are no ties in the failure time. The proof follows mostly [34]. Let $n_1 > n_2 > \ldots > n_k \geq 1$ be the sequence of counts of the at-risk sample size, i.e., $n_j = \sum_{i=1}^{n} \mathbb{1}(Y_i \geq t_j)$, where t_i is the ith ordered failure time. Then the Kaplan-Meier estimator at any observed failure time point t_j can be expressed as $\widehat{S}_{KM}(t_j) = \prod_{i=1}^{j}(n_i - 1)/n_i$, while the Nelson-Altshuler estimator at the same time point is $\widehat{S}_{NA}(t_j) = \exp\{-\sum_{i=1}^{j} 1/n_i\}$. We first apply the Taylor expansion of e^{-n_i} for $n_i \geq 1$:

$$1 - 1/n_i < e^{-n_i} < 1 - 1/n_i + 1/(2n_i^2) \leq 1 - 1/(n_i + 1).$$

Thus we can bound the Nelson-Altshuler estimator with

$$\widehat{S}_{KM}(t_j) < \widehat{S}_{NA}(t_j) < \prod_{i=1}^{j} n_i/(n_i + 1).$$

To bound the difference between the two estimators, note that for $n_j \geq 2$,

$$\begin{aligned}
\left|\widehat{S}_{KM}(t_j) - \widehat{S}_{NA}(t_j)\right| &< \left|\widehat{S}_{KM}(t_j) - \prod_{i=1}^{j} n_i/(n_i + 1)\right| \\
&= \widehat{S}_{KM}(t_j)\left|1 - \prod_{i=1}^{j} \tfrac{n_i/(n_i+1)}{(n_i-1)/n_i}\right| \\
&\leq \widehat{S}_{KM}(t_j) \sum_{i=1}^{j} (n_i^2 - 1)^{-1} \\
&\leq 2\widehat{S}_{KM}(t_j) \sum_{i=1}^{j} n_i^{-2} \\
&\leq 4\widehat{S}_{KM}(t_j)/n_j. \qquad (8)
\end{aligned}$$

Now note that both the Kaplan-Meier and the Nelson-Altshuler estimators stay constant within (t_i, t_{i+1}), and this bound applies to the entire interval $(0, t_k)$ for $n_k \geq 2$. □

Proof of Theorem 1. Recall the counting process

$$N(s) = \sum_{i=1}^{n} N_i(s) = \sum_{i=1}^{n} \mathbb{1}(Y_i \leq s, \delta_i = 1),$$

and the at risk process

$$K(s) = \sum_{i=1}^{n} K_i(s) = \sum_{i=1}^{n} \mathbb{1}(Y_i \geq s).$$

We prove the theorem based on the following key results.

Lemma 8. *Provided Assumption 1 holds, for arbitrary $\epsilon > 0$ and n such that $\frac{1}{n} \leq \frac{\epsilon^2}{8}$, we have*

$$\begin{aligned}
\text{pr}(\sup_{t\leq\tau} |\frac{1}{n}\sum_{i=1}^{n}\{K_i(s) - E[K_i(s)]\}| > \epsilon) &\leq 8(n+1)\exp\big\{\frac{-n\epsilon^2}{8}\big\}, \\
\text{pr}(\sup_{t\leq\tau} |\frac{1}{n}\sum_{i=1}^{n}\{N_i(t) - E[N_i(t)]\}| > \epsilon) &\leq 8(n+1)\exp\big\{\frac{-n\epsilon^2}{8}\big\}.
\end{aligned}$$

Lemma 9. *Provided Assumption 1 holds, for any $\epsilon > 0$, we have*

$$\text{pr}(\sup_{t\leq\tau} |\int_0^t (\frac{1}{K(s)} - \frac{1}{E[K(s)]})dN(s)| > \epsilon) \leq 8(n+2)\exp\Big\{-\frac{n\min(\epsilon^2 M^4, 16M^2)}{32}\Big\},$$

where n satisfies $\frac{1}{n} \leq \frac{\epsilon^2 M^4}{32}$.

Lemma 10. *Provided Assumption 1 holds, for any $\epsilon > 0$, we have*

$$\text{pr}(\sup_{t\leq\tau} |\int_0^t \frac{d\{N(s) - E[N(s)]\}}{E[K(s)]}| > \epsilon) \leq 8(n+1)\exp\Big\{-\frac{n\epsilon^2 M^2}{72}\Big\},$$

where n satisfies $\frac{1}{n} \leq \frac{\epsilon^2 M^2}{72}$.

The proof of Lemma 8 follows pages 14–16 in [102]. The proofs of Lemma 9 and 10 are presented in Appendix. Now we are ready to prove Theorem 1. Note that

$$\begin{aligned}
&\text{pr}\big(\sup_{t<\tau}|\widehat{\Lambda}(t) - \Lambda_n^*(t)| > \epsilon_1\big) \\
&= \text{pr}\Big(\sup_{t<\tau}|\widehat{\Lambda}(t) - \int_0^t \frac{dE[N(s)]}{E[K(s)]}| > \epsilon_1\Big) \\
&\leq \text{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t \Big[\frac{1}{K(s)} - \frac{1}{E[K(s)]}\Big]dN(s)\Big| > \frac{\epsilon_1}{2}\Big) \\
&\quad + \text{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t \frac{d\{N(s) - E[N(s)]\}}{E[K(s)]}\Big| > \frac{\epsilon_1}{2}\Big).
\end{aligned}$$

By Lemma 9, the first term is bounded by $8(n+2)\exp\{-\frac{n\min(\epsilon_1^2M^4,64M^2)}{128}\}$. By Lemma 10, the second term is bounded by $8(n+1)\exp\{-\frac{n\epsilon_1^2M^2}{288}\}$. The sum of these two terms is further bounded by $16(n+2)\exp\{-\frac{n\epsilon_1^2M^4}{288}\}$ for any $n \geq \frac{288}{\epsilon_1^2M^4}$. This completes the proof. □

Proof of Lemma 9. For any $t \leq \tau$,

$$\begin{aligned}
&\Big|\int_0^t (\frac{1}{K(s)} - \frac{1}{E[K(s)]})dN(s)\Big| \\
\leq &\int_0^t \frac{|E[K(s)] - K(s)|}{K(s)E[K(s)]}dN(s) \\
\leq &\int_0^t \frac{\sup_{0<r\leq\tau}|E[K(r)] - K(r)|}{K(s)E[K(s)]}dN(s). \qquad (9)
\end{aligned}$$

Thanks to Hoeffding's inequality, we have

$$\mathrm{pr}\Big(\big|K(\tau) - E[K(\tau)]\big| > \frac{nM}{2}\Big) < 2\exp\big\{-\frac{nM^2}{2}\big\}.$$

Then (9) is further bounded by

$$\frac{n}{(nM)^2/2}\sup_{0<t\leq\tau}\big|E[K(t)] - K(t)\big|.$$

Combining with Lemma 8, we have

$$\begin{aligned}
&\mathrm{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t (\frac{1}{K(s)} - \frac{1}{E[K(s)]})dN(s)\Big| > \epsilon\Big) \\
\leq &\ \mathrm{pr}\Big(\frac{2}{nM^2}\sup_{t\leq\tau}|E[K(t)] - K(t)| > \epsilon\Big) + 2\exp\big\{-\frac{nM^2}{2}\big\} \\
\leq &\ 8(n+2)\exp\Big\{-\frac{n\min(\epsilon^2M^4, 16M^2)}{32}\Big\},
\end{aligned}$$

for any n satisfying $\frac{1}{n} \leq \frac{\epsilon^2M^4}{32}$. This completes the proof. □

Proof of Lemma 10. For any $t \leq \tau$, we utilize integration by parts to obtain

$$
\begin{aligned}
&\Big| \int_0^t \frac{1}{EK(s)} d\{N(s) - E[N(s)]\} \Big| \\
&= \Big| \frac{N(s) - E[N(s)]}{E[K(s)]} \Big|_0^t - \int_0^t \{N(s) - E[N(s)]\} d\Big\{ \frac{1}{E[K(s)]} \Big\} \Big| \\
&\leq 2 \sup_{t \leq \tau} \big| N(t) - E[N(t)] \big| \frac{1}{E[K(\tau)]} + \sup_{t \leq \tau} \big| N(t) - E[N(t)] \big| \int_0^\tau d\Big\{ \frac{1}{E[K(s)]} \Big\} \\
&\leq \frac{3}{M} \sup_{t \leq \tau} \frac{1}{n} \big| N(t) - E[N(t)] \big|.
\end{aligned}
$$

Thanks to Lemma 8, we now have

$$
\begin{aligned}
&\operatorname{pr}\Big(\sup_{t \leq \tau} \Big| \int_0^t \frac{d\{N(s) - E[N(s)]\}}{E[K(s)]} \Big| > \epsilon \Big) \\
&\leq \operatorname{pr}\Big(\frac{3}{nM} \sup_{t \leq \tau} \big| N(t) - E[N(t)] \big| > \epsilon \Big) \\
&\leq 8(n+1) \exp\Big\{ - \frac{n \epsilon^2 M^2}{72} \Big\},
\end{aligned}
$$

where n satisfies $\frac{1}{n} \leq \frac{\epsilon^2 M^2}{72}$. This completes the proof. □

Adaptive concentration bound

The proof of Theorem 2 essentially relies on two main mechanics: the concentration bound results we established in Theorem 1 to bound the variations in each terminal node; and a construction of parsimonious set of rectangles, namely $\boldsymbol{R}$, so that any terminal node $\mathcal{A} \in \boldsymbol{A}$ can be approximated by a rectangle $\mathcal{R} \in \boldsymbol{R}$ [126]. We first introduce some notation and lemmas.

Preliminary. We denote the rectangles $\mathcal{R} \in [0,1]^d$ by

$$
\mathcal{R} = \bigotimes_{j=1}^{d} [r_j^-, r_j^+], \quad \text{where} \quad 0 \leq r_j^- < r_j^+ \leq 1 \quad \text{for all} \quad j = 1, \cdots, d.
$$

The Lebesgue measure of rectangle $\mathcal{R}$ is $\lambda(\mathcal{R}) = \prod_{j=1}^{d}(r_j^+ - r_j^-)$. Here we define the expected fraction of training samples and the number of training samples inside R, respectively, as follows:

$$\mu(\mathcal{R}) = \int_{\mathcal{R}} f(x)dx, \#\mathcal{R} = |\{i : X_i \in \mathcal{R}\}|.$$

We define the support of rectangle $\mathcal{R}$ as $S(\mathcal{R}) = \{j \in 1, \ldots, d : r_j^- \neq 0 \text{ or } r_j^+ \neq 1\}$.

The following lemmas are used in the proof. The construction of the approximation set is shown in [126]. Lemma 11 defines $\boldsymbol{R}_{S,w,\epsilon}$, a set of rectangles, and provides a bound of its cardinality.

Lemma 11. *(Theorem 7 and Corollary 8 in [126]) Let $S \in \{1, \ldots, d\}$ be a set of size $|S| = s$, and let $w, \epsilon \in (0, 1)$. There exists a set of rectangles $\boldsymbol{R}_{S,w,\epsilon}$ such that the following properties hold. Any rectangle $\mathcal{R}$ with support $S(\mathcal{R}) \subseteq S$ and Lebesgue measure $\lambda(\mathcal{R}) \geq w$ can be approximated by rectangles in $\boldsymbol{R}_{S,w,\epsilon}$. Specifically, there exist rectangles $\mathcal{R}_-, \mathcal{R}_+ \in \boldsymbol{R}_{S,w,\epsilon}$ such that*

$$\mathcal{R}_- \subseteq \mathcal{R} \subseteq \mathcal{R}_+, e^{-\epsilon}\lambda(\mathcal{R}_+) \leq \lambda(\mathcal{R}) \leq e^{\epsilon}\lambda(\mathcal{R}_-).$$

Moreover, the cardinality of the set $\boldsymbol{R}_{S,w,\epsilon}$ is bounded by

$$|\boldsymbol{R}_{S,w,\epsilon}| \leq \frac{1}{w}\Big(\frac{8s^2}{\epsilon^2}\big(1 + \log_2\lfloor\frac{1}{w}\rfloor\big)\Big)^s\big(1 + O(\epsilon)\big).$$

Furthermore, if we let $\mathcal{R}_{s,w,\epsilon} = \bigcup_{|S|=s}\mathcal{R}_{S,w,\epsilon}$ include all possible s-sparse rectangles, and let $w = \frac{k}{2\zeta n}$, $\epsilon = \frac{1}{\sqrt{k}}$ and $s = \lfloor\frac{\log(n/k)}{\log((1-\alpha)^{-1})}\rfloor + 1$, where $0 < \alpha < 0.5$ and $\zeta \geq 1$, we then have

$$\log(|\boldsymbol{R}_{S,w,\epsilon}|) \leq \frac{\log(n/k)(\log(dk) + 3\log\log(n))}{\log((1-\alpha)^{-1})} + O(\log(\max\{n, d\})).$$

Lemma 12 below shows that the number of training samples in the terminal node $\mathcal{A}$ is close to the approximation rectangle $\mathcal{R}$.

Lemma 12. *(Theorem 10 in [126]) Assume Assumption 2 and 3 hold. We set $w = \frac{k}{2\zeta n}$, $\epsilon = \frac{1}{\sqrt{k}}$. Then there exists an $n_0 \in N$, for every $n \geq n_0$, the following event holds with probability larger than $1 - n^{-1/2}$: For every possible terminal node $\mathcal{A} \in \boldsymbol{A}$, we can select a rectangle $\mathcal{R} \in \boldsymbol{R}_{s,w,\epsilon}$ such that*

$\mathcal{R} \subseteq \mathcal{A}$, $\lambda(\mathcal{A}) \leq e^{\epsilon}\lambda(\mathcal{R})$, *and*

$$\#\mathcal{A} - \#\mathcal{R} \leq 3\zeta^2\epsilon\#\mathcal{A} + 2\sqrt{3\log(|\boldsymbol{R}|)\#\mathcal{A}} + O(\log(|\boldsymbol{R}|)).$$

Lemma 13 below shows that with high probability there are enough observations larger than or equal to τ on the rectangle $\mathcal{R}$.

Lemma 13. *Provided Assumption 1 holds, the number of observations larger than or equal to τ on all $\mathcal{R} \in \boldsymbol{R}$ is larger than $\left(1 - \sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\right)kM$ with probability larger than $1 - 1/\sqrt{n}$.*

Proof. For one $\mathcal{R} \in \boldsymbol{R}$, by the Chernoff bound, with probability larger than $1 - \exp\big\{-\frac{c^2\#\mathcal{R}M}{2})\big\} \geq 1 - \exp\big\{-\frac{c^2kM}{4})\big\}$, the number of observations larger than or equal to τ on $\mathcal{R}$ is larger than $(1-c)kM$, where $0 < c < 1$ is a constant. Thus with probability larger than $1 - 1/\sqrt{n}$, the number of observations larger than or equal to τ on every $\mathcal{R} \in \boldsymbol{R}$ is larger than $\left(1 - \sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\right)kM$. □

Proof of Theorem 2. We first establish a triangle inequality by picking some element $\mathcal{R}$ in the set $\boldsymbol{R}$ such that it is a close approximation of $\mathcal{A}$. The following is satisfied with large probability, where the small probably event prevents us from finding a good enough $\mathcal{R}$.

$$\begin{aligned}
&\sup_{t<\tau,\, \mathcal{A}\in\boldsymbol{A},\, \boldsymbol{A}\in\mathcal{V}} \left|\widehat{\Lambda}_{\mathcal{A}}(t) - \Lambda^*_{\mathcal{A},n}(t)\right| \\
&\qquad \leq \sup_{t<\tau,\, \mathcal{A}\in\boldsymbol{A},\, \boldsymbol{A}\in\mathcal{V}} \inf_{\mathcal{R}\in\boldsymbol{R}_{s,w,\epsilon}} \left|\widehat{\Lambda}_{\mathcal{A}}(t) - \widehat{\Lambda}_{\mathcal{R}}(t)\right| \\
&\qquad + \sup_{t<\tau,\, \mathcal{R}\in\boldsymbol{R}_{s,w,\epsilon},\, \#\mathcal{R}\geq k/2} \left|\widehat{\Lambda}_{\mathcal{R}}(t) - \Lambda^*_{\mathcal{R},n}(t)\right| \\
&\qquad + \sup_{t<\tau,\, \mathcal{A}\in\boldsymbol{A},\, \boldsymbol{A}\in\mathcal{V}} \inf_{\mathcal{R}\in\boldsymbol{R}_{s,w,\epsilon}} \left|\Lambda^*_{\mathcal{R},n}(t) - \Lambda^*_{\mathcal{A},n}(t)\right|. \qquad (10)
\end{aligned}$$

Here, we have $\#\mathcal{R} \geq k/2$ in the sub-index of the second term because $\#\mathcal{A} \geq k$ and from Lemma 12, $\#\mathcal{A} - \#\mathcal{R} \leq 3\zeta^2\epsilon\#\mathcal{A} + 2\sqrt{3\log(|\boldsymbol{R}|)\#\mathcal{A}} + O(\log(|\boldsymbol{R}|)) = o(k)$ for any possible $\mathcal{A}$ with large probability. Hence the case that $\mathcal{R} < k/2$ is included in the small probability event.

We now bound each part of the right hand side of the above inequality. Noting that we always select a close approximation of $\mathcal{A}$ from the set $\boldsymbol{R}$, the first part is bounded by the following with the specific $\mathcal{R}$ constructed in Lemma 12. With a slight abuse of notation, we let the subject index i first run through the observations within $\mathcal{R}$ and then through the observations in $\mathcal{A}$ but not in

$\mathcal{R}$. This can always be done since $\mathcal{R} \subseteq \mathcal{A}$. Thus we have

$$
\begin{aligned}
&\sup_{t<\tau,\, \mathcal{A}\in\boldsymbol{A},\, \boldsymbol{A}\in\mathcal{V}} \big|\widehat{\Lambda}_{\mathcal{R}}(t) - \widehat{\Lambda}_{\mathcal{A}}(t)\big| \\
&\le \sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}} \Big| \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s)} - \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}} + [\Delta N(s)]_{\mathcal{A}\setminus\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s)} \Big| \\
&= \sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}} \Big| \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s)} \\
&\qquad\qquad - \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}} + [\Delta N(s)]_{\mathcal{A}\setminus\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s)} \Big| \\
&\le \sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}} \Big\{ \sum_{j=\#\mathcal{R}+1}^{\#\mathcal{A}} \frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)} \\
&\qquad + \sum_{j=1}^{\#\mathcal{R}} \Big[\frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j)} - \frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)} \Big] \Big\},
\end{aligned}
$$

where $N(s) = \sum_{i=1}^n N_i(s) = \sum_{i=1}^n \mathbb{1}(Y_i \le s, \delta_i = 1)$. By Lemma 13 the first term is bounded by

$$
\begin{aligned}
&\frac{\#\mathcal{A} - \#\mathcal{R}}{\Big(1 - \sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\Big) kM} \\
&\le \frac{1}{\Big(1 - \sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\Big) M} \Big[6\zeta^2\epsilon + 2\sqrt{\frac{6\log(|\boldsymbol{R}|)}{k}} + O\Big(\frac{\log(|\boldsymbol{R}|)}{k}\Big) \Big],
\end{aligned}
$$

and the second term is bounded by

$$
\begin{aligned}
&\sum_{j=1}^{\#\mathcal{R}} \Big[\frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j)} - \frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)} \Big] \\
&\le \sum_{j=1}^{\#\mathcal{R}} \frac{\Delta N(s_j) \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Z_i \ge s_j)}{\big[\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j)\big]\big[\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)\big]} \\
&\le \sum_{j=1}^{\#\mathcal{R}} \frac{\Delta N(s_j)(\#\mathcal{A} - \#\mathcal{R})}{\big[\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j)\big]\big[\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)\big]} \\
&\le \frac{\#\mathcal{R}(\#\mathcal{A} - \#\mathcal{R})}{\Big(1 - \sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\Big)^2 k^2 M^2} \\
&\le \frac{2}{\Big(1 - \sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\Big)^2 M^2} \Big[6\zeta^2\epsilon + 2\sqrt{\frac{6\log(|\boldsymbol{R}|)}{k}} + O\Big(\frac{\log(|\boldsymbol{R}|)}{k}\Big) \Big].
\end{aligned}
$$

Combining these two terms, the first part of Equation 10 is bounded by

$$\frac{3}{\Big(1-\sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\Big)^2 M^2}\Big[6\zeta^2\epsilon+2\sqrt{\frac{6\log(|\boldsymbol{R}|)}{k}}+O\Big(\frac{\log(|\boldsymbol{R}|)}{k}\Big)\Big], \tag{11}$$

with probability larger than $1-1/\sqrt{n}$. For the second part,

$$\sup_{t<\tau,\mathcal{R}\in\boldsymbol{R}_{s,w,\epsilon},\#\mathcal{R}\geq k/2}\big|\widehat{\Lambda}_{\mathcal{R}}(t)-\Lambda^*_{\mathcal{R},n}(t)\big|\leq\frac{\{288[1/2\log(n)+\log(16k+32)]\}^{1/2}}{k^{1/2}M^2}, \tag{12}$$

with probability larger than $1-1/\sqrt{n}$. The third part of Equation 10 is bounded by

$$\begin{aligned}
&\sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}}\big|\Lambda^*_{\mathcal{A},n}(t)-\Lambda^*_{\mathcal{R},n}(t)\big|\\
\leq&\sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}}\Big|\int_0^t\frac{\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}dF_i(s)}{\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}}-\int_0^t\frac{\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}dF_i(s)}{\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}}\Big|\\
\leq&\sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}}\int_0^t\Big|\frac{\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}dF_i(s)\big]\big[\sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}{\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}\big]\big[\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}\\
&\qquad-\frac{\big[\sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\{1-G_i(s)\}dF_i(s)\big]\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}{\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}\big]\big[\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}\Big|\\
\leq&\sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}}\tau\frac{\#\mathcal{A}(\#\mathcal{A}-\#\mathcal{R})}{\#\mathcal{R}\#\mathcal{A}M^4}\\
\leq&\frac{\tau}{M^4}\{3\zeta^2\epsilon+2\sqrt{\frac{3\log(|\boldsymbol{R}|)}{k}}+O(\frac{\log(|\boldsymbol{R}|)}{k})\}.
\end{aligned} \tag{13}$$

Combining inequalities (11), (12) and (13), we obtain the desired adaptive concentration bound. With probability $1-2/\sqrt{n}$, we have

$$\begin{aligned}
&\sup_{t<\tau,\mathcal{A}\in\boldsymbol{A},\boldsymbol{A}\in\mathcal{V}}|\widehat{\Lambda}_{\mathcal{A}}(t)-\Lambda^*_{\mathcal{A},n}(t)|\\
\leq&\frac{3}{\Big(1-\sqrt{\frac{4\log(|\boldsymbol{R}|\sqrt{n})}{kM}}\Big)^2M^2}\Big[6\zeta^2\epsilon+2\sqrt{\frac{6\log(|\boldsymbol{R}|)}{k}}+O\Big(\frac{\log(|\boldsymbol{R}|)}{k}\Big)\Big]\\
&+\frac{\{288[1/2\log(n)+\log(16k+32)]\}^{1/2}}{k^{1/2}M^2}+\frac{\tau}{M^4}\Big\{3\zeta^2\epsilon+2\sqrt{\frac{3\log(|\boldsymbol{R}|)}{k}}+O\Big(\frac{(\log(|\boldsymbol{R}|)}{k}\Big)\Big\}\\
\leq&M_1\Big[\sqrt{\frac{\log(|\boldsymbol{R}|)}{k}}+\sqrt{\frac{\log(n)}{k}}\Big]\leq M_1\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},
\end{aligned}$$

where M_1 is an universal constant. This completes the proof of Theorem 2. Furthermore, if $\liminf_{n\to\infty}(d/k) > 0$, the bound degenerates to $M_1\sqrt{\frac{\log(n)\log(d)}{k\log((1-\alpha)^{-1})}}$. This completes the proof of Corollary 1. □

Proof of Corollary 2. Since for any $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$ we have

$$\sup_{t<\tau,\, x\in[0,1]^d} \left|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda^*_{\boldsymbol{A},n}(t \mid x)\right| \leq M_1\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

and furthermore if $\liminf_{n\to\infty}(d/n) \to \infty$, for any $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$,

$$\sup_{t<\tau,\, x\in[0,1]^d} \left|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda^*_{\boldsymbol{A},n}(t \mid x)\right| \leq M_1\sqrt{\frac{\log(n)\log(d)}{k\log((1-\alpha)^{-1})}}.$$

By the definition of $\mathcal{H}_{\alpha,k}(\mathcal{D}_n)$, any $\{\boldsymbol{A}_{(b)}\}_1^B$ belonging to $\mathcal{H}_{\alpha,k}(\mathcal{D}_n)$ is an element of $\mathcal{V}_{\alpha,k}(\mathcal{D}_n)$. Hence we have,

$$\begin{aligned}\sup_{t<\tau,\, x\in[0,1]^d,\, \{\boldsymbol{A}_{(b)}\}_1^B\in\mathcal{H}_{\alpha,k}(\mathcal{D}_n)} &\left|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid x) - \Lambda^*_{\{\boldsymbol{A}_{(b)}\}_1^B,n}(t \mid x)\right| \\ &\leq M_1\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},\end{aligned}$$

for some universal constant M_1. Furthermore, if $\liminf_{n\to\infty}(d/n) \to \infty$,

$$\begin{aligned}\sup_{t<\tau,\, x\in[0,1]^d,\, \{\boldsymbol{A}_{(b)}\}_1^B\in\mathcal{H}_{\alpha,k}(\mathcal{D}_n)} &\left|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid x) - \Lambda^*_{\{\boldsymbol{A}_{(b)}\}_1^B,n}(t \mid x)\right| \\ &\leq M_1\sqrt{\frac{\log(n)\log(d)}{k\log((1-\alpha)^{-1})}},\end{aligned}$$

with probability larger than $1 - 2/\sqrt{n}$, for some universal constant M_1. □

Consistency when d is fixed

Proof of Theorem 3. In order to show consistency, we first show that each terminal node is small enough in all d dimensions. Let m be the lower bound of the number of splits on the terminal node $\mathcal{A}$ containing x, and m_i be the number of splits on the i-th dimension. Then we have

$$n\alpha^m = k, \quad m = \log_{1/\alpha}(n/k) = \frac{\log n - \log k}{\log(1/\alpha)} \quad \text{and} \quad \sum_{i=1}^{d} m_i = m.$$

The lower bound of the number of splits on i-th dimension m_i has distribution $Binomial(m, \frac{1}{d})$. By the Chernoff bound on each dimension,

$$\text{pr}\Big(m_i > \frac{(1-c_2)m}{d}\Big) > 1 - \exp\Big\{-\frac{c_2^2 m}{2d}\Big\}$$

with any $0 < c_2 < 1$. Then, by Bonferroni,

$$\text{pr}\Big(\min m_i > \frac{(1-c_2)m}{d}\Big) > 1 - d\exp\Big\{-\frac{c_2^2 m}{2d}\Big\}.$$

Suppose we are splitting at the i-th dimension on a specific internal node with ν observations. Recall the splitting rule is choosing the splitting point randomly between the $\max((k+1), \lceil \alpha\nu \rceil)$-th, and $\min((n-k-1), \lfloor (1-\alpha)\nu \rfloor)$-th observations. Without loss of generality, we consider splitting between $\lceil \alpha\nu \rceil$-th and $\lfloor (1-\alpha)\nu \rfloor$-th observations. The event that the splitting point is between α and $1-\alpha$ happens with probability larger than c_3, where $c_3 = (1-2\alpha)/8$ and is just a lower bound. Since with probability larger than $1/4$, the $\lfloor \frac{\alpha+0.5}{2}\nu \rfloor$-th order statistic is larger than α and the $\lceil \frac{1.5-\alpha}{2}\nu \rceil$-th order statistic is less than $1-\alpha$ for large enough ν, where ν is known to be larger than $2k$. So with probability larger than $c_3 = (1-2\alpha)/8$, the splitting point falls into the interval $[\alpha, 1-\alpha]$.

The number of splits which partition the parent node to two child nodes with proportion of length between both α and $1-\alpha$ on the i-th dimension of the terminal node $\mathcal{A}$ is denoted by m^* and is $Binomial(m_i, c_3)$. By the Chernoff bound, for any $0 < c_4 < 1$,

$$\text{pr}\big(m^* \geq (1-c_4)c_3 m_i\big) \geq 1 - \exp\Big\{-\frac{c_4^2 c_3 m_i}{2}\Big\}.$$

If we denote the length of the i-th dimension on the terminal node $\mathcal{A}$ as l_i,

$$\mathrm{pr}\big(l_i \le (1-\alpha)^{(1-c_4)c_3 m_i}\big) \ge 1 - \exp\Big\{-\frac{c_4^2 c_3 m_i}{2}\Big\}.$$

Furthermore, by combining the d dimensions together, we obtain

$$\mathrm{pr}\big(\max_i l_i \le (1-\alpha)^{(1-c_4)c_3 \min_i m_i}\big) \ge 1 - d\exp\Big\{-\frac{c_4^2 c_3 \min_i m_i}{2}\Big\},$$

and then

$$\max_{x_1,x_2\in\mathcal{A}} ||x_1 - x_2|| \le \sqrt{d}(1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

with probability larger than $1 - d\exp\big\{-\frac{c_2^2 m}{2d}\big\} - d\exp\big\{-\frac{(1-c_2)c_3c_4^2 m}{2d}\big\}$. Hence, for any observation x_j inside the node $\mathcal{A}$ containing x, by Assumption 4, we have

$$\begin{aligned}
\sup_{t<\tau} |F(t \mid x) - F(t \mid x_j)| &\le L_1\sqrt{d}(1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}}, \\
\sup_{t<\tau} |f(t \mid x) - f(t \mid x_j)| &\le (L_1^2 + L_2)\sqrt{d}(1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},
\end{aligned}$$

where $f(\cdot \mid x)$ and $F(\cdot \mid x)$, respectively, denote the true density function and distribution function at $x \in \mathcal{A}$. Then $\Lambda^*_{\mathcal{A},n}(t)$ has the upper and lower bounds

$$\int_0^t \frac{f(s \mid x) + b_1}{1 - F(s \mid x) - b_2}\, ds \quad \text{and} \quad \int_0^t \frac{f(s \mid x) - b_1}{1 - F(s \mid x) + b_2}\, ds,$$

respectively, where

$$b_1 = (L_1^2 + L_2)\sqrt{d}(1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}}, \text{ and } b_2 = L_1\sqrt{d}(1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}}.$$

Hence, $|\Lambda^*_{\mathcal{A},n}(t) - \Lambda(t \mid x)|$ has the bound

$$\int_0^t \frac{b_1(1 - F(s \mid x)) + b_2 f(s \mid x)}{(1 - F(s \mid x) - b_2)(1 - F(s \mid x))} ds \le M_2 \tau \sqrt{d}(1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

for any $t < \tau$, where M_2 is some constant depending on L_1 and L_2. Hence, for the terminal node $\mathcal{A}$ containing x, we bound the bias by

$$\sup_{t<\tau} |\Lambda^*_{\mathcal{A},n}(t) - \Lambda(t \mid x)| \leq M_2 \tau \sqrt{d} (1-\alpha)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

with probability larger than $1 - d\exp\{-\frac{c_2^2 m}{2d}\} - d\exp\{-\frac{(1-c_2)c_3c_4^2 m}{2d}\}$. Combining this with the adaptive concentration bound result from Theorem 2, for each x, we further have

$$\sup_{t<\tau} |\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda(t \mid x)| = O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big),$$

with probability larger than $1 - w_n$, where

$$w_n = \frac{2}{\sqrt{n}} + d\exp\Big\{-\frac{c_2^2 \log_{1/\alpha}(n/k)}{2d}\Big\} + d\exp\Big\{-\frac{(1-c_2)c_3c_4^2\log_{1/\alpha}(n/k)}{2d}\Big\},$$

and $c_1 = \frac{c_3(1-c_2)(1-c_4)}{\log_{1-\alpha}(\alpha)}$. This completes the proof of point-wise consistency.

Lastly, we need to establish the bound of $|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda(t \mid x)|$ under the event with small probability w_n. Noticing that $\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x)$ is simply the Nelson-Aalen estimator of the cumulative hazard function with at most k terms, for any $t < \tau$, we have

$$\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) \leq \frac{1}{k} + \ldots + \frac{1}{1} = O(\log(k)),$$

which implies that

$$|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda(t \mid x)| \leq O(\log(k)).$$

This completes the proof. □

Proof of Theorem 4. From Theorem 3, we have

$$\begin{aligned}
&\sup_{t<\tau} E_X\big|\widehat{\Lambda}(t \mid X) - \Lambda(t \mid X)\big| \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \big(\frac{k}{n}\big)^{\frac{c_1}{d}} + \log(k)w_n\Big),
\end{aligned}$$

which leads to the following bounds:

$$\begin{aligned}
&\sup_{t<\tau} E_X|\widehat{\Lambda}_{\{\boldsymbol{A}_{(b)}\}_1^B}(t \mid X) - \Lambda(t \mid X)| \\
&= \lim_{B\to\infty} \sup_{t<\tau} E_X|\frac{1}{B}\sum_{b=1}^{B} \widehat{\Lambda}_{\boldsymbol{A}_{(b)}}(t \mid X) - \frac{1}{B}\sum_{b=1}^{B} \Lambda(t \mid X)| \\
&\le \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \sup_{t<\tau} E_X|\widehat{\Lambda}_{\boldsymbol{A}_{(b)}}(t \mid X) - \Lambda(t \mid X)| \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \big(\frac{k}{n}\big)^{\frac{c_1}{d}} + \log(k)w_n\Big). \ \square
\end{aligned}$$

Consistency when d is infinite

Lemma 14. *Assume that the density function of the failure time $f_i(t) = dF_i(t)$ is bounded above by L for each i. The difference between $\Lambda^*_{\mathcal{A},n}(t)$ and $\Lambda^*_{\mathcal{A}}(t)$ is bounded by*

$$\sup_{t<\tau} \big|\Lambda^*_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)\big| \le \sqrt{\frac{4\tau^2 L^2 \log(4\sqrt{n})}{M^2 n}},$$

with probability larger than $1 - 1/\sqrt{n}$.

Proof. By Hoeffding's inequality, we have for each $s \le t$,

$$\begin{aligned}
\mathrm{pr}\Big(\Big|\frac{1}{n}\sum_{X_i\in\mathcal{A}} [1 - G_i(s)]f_i(s) - E_X\{[1 - G(s \mid X)]f(s \mid X)\}\Big| \\
\ge \sqrt{\frac{L^2\log(4\sqrt{n})}{2n}}\ \Big) \le \frac{1}{2\sqrt{n}},
\end{aligned}$$

and

$$\mathrm{pr}\Big(\Big| \frac{1}{n} \sum_{X_i \in \mathcal{A}} [1 - G_i(s)][1 - F_i(s \mid X)] - E_X\{[1 - G(s \mid X)][1 - F(s \mid X)]\} \Big| \geq \sqrt{\frac{\log(4\sqrt{n})}{2n}} \Big) \leq \frac{1}{2\sqrt{n}}.$$

After combining the above two inequalities, we have

$$\sup_{t<\tau} \big|\Lambda^*_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)\big| \leq \sqrt{\frac{4\tau^2 L^2 \log(4\sqrt{n})}{M^2 n}},$$

with probability larger than $1 - 1/\sqrt{n}$. □

Proof of Lemma 2. In a similar way as done for Lemma 14, for each $s \leq t$,

$$\mathrm{pr}\Big(\Big| \frac{1}{n} \sum_{X_i \in \mathcal{A}} [1 - G_i(s)] f_i(s) - E_X\{[1 - G(s \mid X)] f(s \mid X)\} \Big| \geq \sqrt{\frac{L^2 \log(4\sqrt{n}|\mathcal{R}|)}{2n}} \Big) \leq \frac{1}{2\sqrt{n}},$$

and

$$\mathrm{pr}\Big(\Big| \frac{1}{n} \sum_{X_i \in \mathcal{A}} [1 - G_i(s)][1 - F_i(s \mid X)] - E_X\{[1 - G(s \mid X)][1 - F(s \mid X)]\} \Big| \geq \sqrt{\frac{\log(4\sqrt{n}|\mathcal{R}|)}{2n}} \Big) \leq \frac{1}{2\sqrt{n}}.$$

Thus, with probability larger than $1/\sqrt{n}$,

$$\begin{aligned} |\Lambda^*_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)| &\leq \sqrt{\frac{4\tau^2 L^2 \log(4\sqrt{n}|\mathcal{R}|)}{M^2 n}} \\ &\leq M_2 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}}, \end{aligned}$$

for all $t < \tau$ and all $\mathcal{A} \in \boldsymbol{A}, \boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$, where M_2 is some universal constant depending on L and M. □

Proof of Lemma 3. We start with defining $\Delta^*(x) = \int_0^\tau \big|\Lambda^*_{\mathcal{A}_j^+(x)}(t) - \Lambda^*_{\mathcal{A}_j^-(x)}(t)\big|\mathrm{d}t$. Then for any noise variable j,

$$\begin{aligned}
\Delta^*(x) = & \int_0^\tau \big|\Lambda^*_{\mathcal{A}_j^+(x)}(t) - \Lambda^*_{\mathcal{A}_j^-(x)}(t)\big|\mathrm{d}t \\
= & \int_0^\tau \Big| \int_0^t \frac{E_{X\in\mathcal{A}_j^+(x)}[1-G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^+(x)}[1-G(s \mid X)][1-F(s \mid X)]} \\
& \qquad - \int_0^t \frac{E_{X\in\mathcal{A}_j^-(x)}[1-G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^-(x)}[1-G(s \mid X)][1-F(s \mid X)]}\Big|\mathrm{d}t \\
= & \, 0.
\end{aligned}$$

From the adaptive concentration bound result and Lemma 2, we have, for an arbitrary $x \in [0,1]^d$ and a valid partition $\boldsymbol{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$,

$$\int_0^\tau \big|\widehat{\Lambda}_{\boldsymbol{A}}(t \mid x) - \Lambda^*_{\boldsymbol{A}}(t \mid x)\big|\mathrm{d}t \le M_3\tau\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

with probability $1-3/\sqrt{n}$, where $M_3 = \max(M_1, M_2)$. Hence

$$|\Delta(x) - \Delta^*(x)| \le \Delta(x) \le 2M_3\tau\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

with probability $1-3/\sqrt{n}$ uniformly over all possible nodes with at least $2k$ observations and all noise variables. Thus only with probability $3/\sqrt{n}$ will the proposed survival tree split on a noise variable. □

Proof of Lemma 4. Suppose $\mathcal{A}$ is the current node and $X^{(j)}$ is an important variable. We show with high probability that a split happens at $x = 1/2$. Since we choose the splitting point $\tilde{x}$ which

maximizes $\Delta^2(x)$, the signal is more significant at $\tilde{x}$ than $1/2$. Hence we are interested in

$$\begin{aligned}\Delta^*(1/2) &= \int_0^\tau \left|\Lambda^*_{\mathcal{A}_j^+(1/2)}(t) - \Lambda^*_{\mathcal{A}_j^-(1/2)}(t)\right| \mathrm{d}t \\ &= \int_0^\tau \left| \int_0^t \frac{E_{X\in\mathcal{A}_j^+(1/2)}[1-G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^+(1/2)}[1-G(s \mid X)][1-F(s \mid X)]} \right. \\ &\qquad \left. - \int_0^t \frac{E_{X\in\mathcal{A}_j^-(1/2)}[1-G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^-(1/2)}[1-G(s \mid X)][1-F(s \mid X)]} \right| \mathrm{d}t.\end{aligned}$$

Since $1-G(\tau)$ is bounded away from 0 by our assumption with $1-G(\tau) \geq \tilde{M}$, the above expression can be further bounded below by

$$\begin{aligned}&\Delta^*(1/2) \\ &\geq \int_0^\tau \left| \tilde{M} \int_0^t \frac{E_{X\in\mathcal{A}_j^+(1/2)}\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^+(1/2)}[1-F(s \mid X)]} - \frac{1}{\tilde{M}} \int_0^t \frac{E_{X\in\mathcal{A}_j^-(1/2)}\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^-(1/2)}[1-F(s \mid X)]} \right| \mathrm{d}t. \\ &\geq \ell.\end{aligned}$$

Then, by the adaptive concentration bound results above, $\Delta^*(1/2)$ has to be close enough to $\Delta(1/2)$. Thus we have

$$\Delta(1/2) \geq \ell - M_3\tau\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

with probability $1-3/\sqrt{n}$ uniformly over all possible nodes and all signal variables. $\square$

APPENDIX B: SUPPLEMENTARY MATERIAL TO CHAPTER 3

Derivation of generalized fiducial distribution under dependence

We derive GFD for situations when censoring distribution might depend on the failure time. In particular, consider the following data generating equation:

$$Y_i = F^{-1}(U_i) \wedge R_i^{-1}\{V_i \mid F^{-1}(U_i)\}, \quad \delta_i = I[F^{-1}(U_i) \leq R_i^{-1}\{V_i \mid F^{-1}(U_i)\}]. \tag{14}$$

Here, $R_i^{-1}(v \mid t)$ is the inverse of the conditional distribution function of the censoring time given failure time t specific to the i-th subject. Equation (14) allows for any within subject dependence between failure and censoring times.

The corresponding inverse map for a single observation is: If $\delta_i = 1$,

$$Q_1^{F,R_i}(y_i, u_i, v_i) = \{F : F(y_i) \geq u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0\} \times \{R_i : R_i^{-1}(v_i \mid y_i) \geq y_i\}.$$

If $\delta_i = 0$, the inverse map for this datum is

$$Q_0^{F,R_i}(y_i, u_i, v_i) = \{F, R_i : \ F(y_i) < u_i, \\ R_i\{y_i \mid F^{-1}(u_i)\} \geq v_i, R_i\{y_i - \epsilon \mid F^{-1}(u_i)\} < v_i \text{ for any } \epsilon > 0\}.$$

Unlike in (3.6), the inverse $Q^{F,R}(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v}) = \bigcap_i Q_{\delta_i}^{F,R_i}(y_i, u_i, v_i)$ does not factorize into a Cartesian product. However, the projection of $Q^{F,R}(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v})$ onto the failure time distribution margin remains the same as in (3.7), and $Q^{F,R}(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v}) \neq \emptyset$ if and only if $Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}) \neq \emptyset$. Consequently, the marginal fiducial distribution $Q^F(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}) \mid Q^{F,R}(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v}) \neq \emptyset$ is the same as (3.8).

Remarkably, the data generating equation (14) leads to the same fiducial distribution for failure times as in the independent case given by (3.5). The difference is that unlike in the fully independent case, (14) does not provide any useful information about the censoring times and can be viewed as allocating all information in the data to the estimation of failure times.

Proofs

In this section we collect proofs from Section 3.3.

Proof of Theorem 6. For simplicity, in this proof, we denote $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}$ as pr. By the definition of S^U and $\hat{S}$,

$$\sup_{s\leq t}|S^U(s)-\hat{S}(s)| = \sup_{s\leq t}\left|\prod_{i=1}^{\bar{N}(s)}(1-B_i) - \prod_{i=1}^{\bar{N}(s)}(1-\frac{1}{1+\bar{K}(s_i)})\right|, \tag{15}$$

where $B_i \sim Beta(1,\bar{K}(s_i))$, $E(B_i)=\{1+\bar{K}(s_i)\}^{-1}$.

In order to deal with supremum in Equation (15), we use a coupling idea to get

$$\text{pr}(\sum_{i=1}^{\bar{N}(t)} B_i^2 \leq \frac{\epsilon^2}{n^{1/2}}) \geq 1-\bar{N}(t)(1-\frac{\epsilon}{n^{3/4}})^{\bar{K}(t)}. \tag{16}$$

In particular, define $\tilde{B}_i \sim Beta(1,\bar{K}(t))$ generated by the same uniform random variable as B_i, so $\tilde{B}_i \geq B_i$. We have

$$\text{pr}(\max_{1\leq i\leq \bar{N}(t)} B_i \geq \frac{\epsilon}{n^{3/4}}) \leq \bar{N}(t)\bar{K}(t)\int_0^{1-\frac{\epsilon}{n^{3/4}}}\xi^{\bar{K}(t)-1}d\xi = \bar{N}(t)(1-\frac{\epsilon}{n^{3/4}})^{\bar{K}(t)}. \tag{17}$$

Since $\sum_{i=1}^{\bar{N}(t)} B_i^2 \leq \bar{N}(t)\max_{1\leq i\leq\bar{N}(t)} B_i^2$, further we have

$$\text{pr}(\sum_{i=1}^{\bar{N}(t)} B_i^2 \geq \frac{\epsilon^2}{n^{1/2}}) \leq \text{pr}(\max_{1\leq i\leq\bar{N}(t)} B_i \geq \frac{\epsilon}{n^{1/4}\{\bar{N}(t)\}^{1/2}}) \leq \text{pr}(\max_{1\leq i\leq\bar{N}(t)} B_i \geq \frac{\epsilon}{n^{3/4}}).$$

So Equation (16) follows.

In order to bound Equation (15), recall the following facts: $E(B_i)=\{1+\bar{K}(s_i)\}^{-1}\leq 0{\cdot}6$,

$$\text{pr}(\max_{1\leq i\leq\bar{N}(t)} B_i \leq 0{\cdot}6) = 1-\text{pr}(\max_{1\leq i\leq\bar{N}(t)} B_i > 0{\cdot}6) \geq 1-\bar{N}(t)0{\cdot}4^{\bar{K}(t)},$$

and for any $x \leq 0{\cdot}6$, $-x - x^2 \leq \log(1-x) \leq -x$. Equation (15) is bounded by

$$
\begin{aligned}
&\sup_{s\leq t} |\exp\{\sum_{i=1}^{\bar{N}(s)} \log(1-B_i)\} - \exp\{\sum_{i=1}^{\bar{N}(s)} \log(1-E(B_i))\}| \\
\leq &\sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} B_i\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} [E(B_i) + \{E(B_i)\}^2]\}| \\
+ &\sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} (B_i + B_i^2)\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\}| \\
\leq &\sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} B_i\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\}| \\
+ &\sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} [E(B_i) + \{E(B_i)\}^2]\}| \\
+ &\sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} (B_i + B_i^2)\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} B_i\}| + \sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} B_i\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\}| \\
\leq &2\sup_{s\leq t} |\exp\{-\sum_{i=1}^{\bar{N}(s)} B_i + E(B_i) - E(B_i)\} - \exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\}| + \sum_{i=1}^{\bar{N}(t)} \{\bar{K}(t)+1\}^{-2} + \sum_{i=1}^{\bar{N}(t)} B_i^2 \\
\leq &2\sup_{s\leq t} |\sum_{i=1}^{\bar{N}(s)} \{B_i - E(B_i)\}| \exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\} + \bar{N}(t)/\bar{K}(t)^{-2} + \sum_{i=1}^{\bar{N}(t)} B_i^2, \qquad (18)
\end{aligned}
$$

with probability larger than $1 - \bar{N}(t)0{\cdot}4^{\bar{K}(t)}$. Since $\exp\{-\sum_{i=1}^{\bar{N}(s)} E(B_i)\}$ is bounded by 1 for any $s \leq t$, to complete the proof we only need to bound $\sup_{s\leq t} |\sum_{i=1}^{\bar{N}(s)} \{B_i - E(B_i)\}|$.

Let $T_m = \sum_{i=1}^{m} \{B_i - E(B_i)\}$. Then we have $E(T_m) = 0$, $\mathrm{var}(T_m) \leq m/\bar{K}(t)^2 \to 0$. By Kolmogorov's inequality $\mathrm{pr}(\max_{1\leq m\leq n} |T_m| \geq x) \leq x^{-2}\mathrm{var}(T_n)$ [39], we know

$$
\mathrm{pr}(\sup_{s\leq t} |\sum_{i=1}^{\bar{N}(s)} \{B_i - E(B_i)\}| \geq \epsilon^2/n^{1/2}) \leq n\bar{N}(t)/\{\epsilon^2 \bar{K}(t)\}^2. \qquad (19)
$$

Combine (16), (18) and (19), we have

$$
\mathrm{pr}\{\sup_{s\leq t} |S^U(s) - \hat{S}(s)| \geq 3\epsilon^2/n^{1/2} + \bar{N}(t)/\bar{K}(t)^{-2}\} \leq \bar{N}(t)[(1-\epsilon/n^{3/4})^{\bar{K}(t)} + 0{\cdot}4^{\bar{K}(t)} + n/\{\epsilon^2\bar{K}(t)\}^2].
$$

This completes the proof.

□

For the proof of the next Theorem, we will construct a martingale, and check the two conditions similar to Theorem 5.1.1 in [50].

Proof of Theorem 7. For any t with $\pi(t) > 0$, consider a fixed growing sequence of data $(\boldsymbol{y}, \boldsymbol{\delta})$ for which the statement of Assumption 6 and 9 are valid. The set of all such sequences is assumed to have probability one.

For convenience, we denote $\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}$ as pr in the rest of this section. Additionally, in this proof only, we denote S^U, F^L as $\tilde{S}, \tilde{F}$, and define $u(x) = \sum_{i=1}^{\bar{N}(t)} I\{s_{i-1} < x \leq s_i\}B_i$, where s_i are ordered failure times, s_0 is assumed to be 0, and B_i are independent $Beta(1, \bar{K}(s_i))$. Let $\tilde{\Lambda}(s) = \int_0^s u(x)d\bar{N}(x) = \sum_{i=1}^{\bar{N}(t)} B_i$. For fixed $t \in \mathcal{I}$, suppose $0 \leq s \leq t$, we could rewrite $\tilde{S}$ recursively as

$$\tilde{S}(s) = 1 - \int_0^s \tilde{S}(x-)d\tilde{\Lambda}(x).$$

Then we have

$$\tilde{S}(s-) - \tilde{S}(s) = -\Delta\tilde{S}(s) = \tilde{S}(s-)\Delta\bar{N}(s)u(s),$$

$$\tilde{S}(s) = \tilde{S}(s-)\{1 - \Delta\bar{N}(s)u(s)\},$$

This is the same as Equation (3.11).

We know $\hat{S}(s) > 0$, therefore

$$\begin{aligned}\frac{\tilde{S}(s)}{\hat{S}(s)} &= \frac{\tilde{S}(0)}{\hat{S}(0)} + \int_0^s \tilde{S}(x-)[-\{\hat{S}(x)\hat{S}(x-)\}^{-1}d\hat{S}(x)] + \int_0^s \frac{1}{\hat{S}(x)}d\tilde{S}(x) \\ &= 1 - \int_0^s \frac{\tilde{S}(x-)}{\hat{S}(x)}\{d\bar{N}(x)u(x) - \frac{d\bar{N}(x)}{1+\bar{K}(x)}\},\end{aligned}$$

so

$$\tilde{S}(s) - \hat{S}(s) = -\hat{S}(s)\int_0^s \frac{\tilde{S}(x-)}{\hat{S}(x)}\{d\bar{N}(x)u(x) - \frac{d\bar{N}(x)}{1+\bar{K}(x)}\},$$

and

$$n^{1/2}\{\tilde{F}(s) - \hat{F}(s)\} = \hat{S}(s)\int_0^s n^{1/2}\frac{\tilde{S}(x-)}{\hat{S}(x)}\{d\bar{N}(x)u(x) - \frac{d\bar{N}(x)}{1+\bar{K}(x)}\}. \tag{20}$$

Now we want to find the asymptotic distribution of right-hand-side of (20). First, notice that for our fixed sequence of data, $\hat{S}(s) \to 1 - F_0(s)$. Next, let

$$U(s) = \int_0^s n^{1/2} \frac{\tilde{S}(x-)}{\hat{S}(x)} \{d\bar{N}(x)u(x) - \frac{d\bar{N}(x)}{1+\bar{K}(x)}\}.$$

We need to construct a martingale $M(s)$ to use the martingale central limit theorem. Let

$$M(s) = \sum_{x \le s} [u(x)\{1+\bar{K}(x)\}\{2+\bar{K}(x)\}^{1/2} - \{2+\bar{K}(x)\}^{1/2}]\Delta\bar{N}(x).$$

It is easy to see that $M(s)$ is a martingale,

$$dM(s) = 0 \text{ if } \Delta\bar{N}(s) = 0,$$
$$dM(s) = u(s)\{1+\bar{K}(s)\}\{2+\bar{K}(s)\}^{1/2} - \{2+\bar{K}(s)\}^{1/2} \text{ if } \Delta\bar{N}(s) = 1.$$

From here

$$dM(s) = d\bar{N}(s)[u(s)\{1+\bar{K}(s)\}\{2+\bar{K}(s)\}^{1/2} - \{2+\bar{K}(s)\}^{1/2}].$$

Let

$$H(s) = n^{1/2} \frac{\tilde{S}(s-)}{\hat{S}(s)\{1+\bar{K}(s)\}\{2+\bar{K}(s)\}^{1/2}},$$

then

$$U(s) = \int_0^s H(x)dM(x).$$

In order to obtain desired convergence, we need to establish the two conditions of Theorem 5.1.1 in [50].

First, we need to check the first condition

$$< U, U > (s) \xrightarrow{pr} \int_0^s f^2(x)dx, \text{ where } f(x) = \{\lambda(x)/\pi(x)\}^{1/2}. \tag{21}$$

We have

$$d < M, M > (x) = \text{var}(dM(x)|\mathcal{F}_{x-}) = \bar{K}(x)d\bar{N}(x),$$

and

$$< U, U > (s) = \int_0^s n \frac{\tilde{S}^2(x-)\bar{K}(x)d\bar{N}(x)}{\hat{S}^2(x)\{1+\bar{K}(x)\}^2\{2+\bar{K}(x)\}}.$$

By Assumption 6, we have

$$\forall n \geq n_0, \quad \text{pr}\left(\sup_{x\leq s}\left|\frac{\tilde{S}^2(x-)}{\hat{S}^2(x)}\left\{\frac{n}{\bar{K}(x)}-\frac{1}{\pi(x)}\right\}\right| > \epsilon/3\right) < \epsilon/2.$$

By the consistency of $\tilde{S}$, we have

$$\forall n \geq n_1, \quad \text{pr}\left(\sup_{x\leq s}\left|\frac{1}{\pi(x)}\left\{\frac{\tilde{S}^2(x-)}{\hat{S}^2(x)}-1\right\}\right| > \epsilon/2\right) < \epsilon/2.$$

So for $\forall \epsilon > 0, \forall n \geq \max(n_0, n_1)$,

$$\text{pr}\left(\sup_{x\leq s}\left|\frac{\tilde{S}^2(x-)n}{\hat{S}^2(x)\bar{K}(x)}-\frac{1}{\pi(x)}\right| > \epsilon\right) < \epsilon.$$

Then by Assumption 8, the condition (21) is satisfied.

Then we need to check the second condition, i.e., $< U_\epsilon, U_\epsilon > (s) \xrightarrow{pr} 0$. For any $\epsilon > 0$,

$$< U_\epsilon, U_\epsilon > (s) = \int_0^s n \frac{\tilde{S}^2(x-)\bar{K}(x)d\bar{N}(x)}{\hat{S}^2(x)\{1+\bar{K}(x)\}^2\{2+\bar{K}(x)\}} I\{\frac{n^{1/2}\tilde{S}(x-)}{\hat{S}(x)\{1+\bar{K}(x)\}\{2+\bar{K}(x)\}^{1/2}} \geq \epsilon\}.$$

Consistency and Assumption 6 implies

$$\begin{aligned}
&\sup_{x\leq s}\left|H^2(x)\{1+\bar{K}(x)\}\{2+\bar{K}(x)\}-\frac{1}{\pi(x)}\right| \\
=&\sup_{x\leq s}\left|n\frac{\tilde{S}^2(x-)}{\hat{S}^2(x)\{1+\bar{K}(x)\}}+\frac{\tilde{S}^2(x-)}{\hat{S}^2(x)\pi(x)}-\frac{\tilde{S}^2(x-)}{\hat{S}^2(x)\pi(x)}-\frac{1}{\pi(x)}\right| \\
\leq&\frac{1}{\hat{S}^2(s)}\sup_{x\leq s}\left|\frac{n}{1+\bar{K}(x)}-\frac{1}{\pi(x)}\right|+\frac{1}{\hat{S}(s)\pi(s)}\sup_{x\leq s}\left|\tilde{S}(x-)-\hat{S}(x)\right| \xrightarrow{pr} 0. \qquad (22)
\end{aligned}$$

From $\bar{K}(x) \xrightarrow{pr} \infty$ and monotonicity of $\bar{K}$ we have

$$\inf_{x\leq s}|\bar{K}(x)| \xrightarrow{pr} \infty.$$

Combined with Equation (22), we have

$$\sup_{x\leq s}|H(x)| \xrightarrow{pr} 0,$$

which is equivalent to

$$\sup_{x\leq s} I\{\frac{n^{1/2}\tilde{S}(x-)}{\hat{S}(x)\{1+\bar{K}(x)\}\{2+\bar{K}(x)\}^{1/2}} \geq \epsilon\} \xrightarrow{pr} 0.$$

Then

$$\int_0^s n\frac{\tilde{S}^2(x-)\bar{K}(x)d\bar{N}(x)}{\hat{S}^2(x)\{1+\bar{K}(x)\}^2\{2+\bar{K}(x)\}} I\{\frac{n^{1/2}\tilde{S}(x-)}{\hat{S}(x)\{1+\bar{K}(x)\}\{2+\bar{K}(x)\}^{1/2}} \geq \epsilon\} \xrightarrow{pr} 0,$$

and the second condition is satisfied.

By replicating the proof in Theorem 5.1.1 in [50] for our martingale, we get $U(s) \Rightarrow U_\infty(s) = \int_0^s \{\lambda(x)/\pi(x)\}^{1/2} dW(x)$. We know

$$\text{cov}(U_\infty(s_1), U_\infty(s_2)) = \int_0^{s_1} \frac{\lambda(x)}{\pi(x)} ds = \gamma(s_1) \quad \text{for} \quad s_1 < s_2,$$

and

$$\text{cov}(W\{\gamma(s_1)\}, W\{\gamma(s_2)\}) = \gamma(s_1) \quad \text{for} \quad s_1 < s_2.$$

So $U_\infty(\cdot)$ is the same as $W\{\gamma(\cdot)\}$. The conclusion of the Theorem 7 follows. □

We conclude this section by proving the corollary.

Proof of Corollary 3. We know $n^{1/2}\{\hat{F}(\cdot) - F_0(\cdot)\} \to \{1 - F_0(\cdot)\}W\{\gamma(\cdot)\}$ on $D[0,t]$ and $n^{1/2}\{F^L(\cdot) - \hat{F}(\cdot)\} \to \{1 - F_0(\cdot)\}W\{\gamma(\cdot)\}$ in distribution on $D[0,t]$ almost surely from Theorem 7.

From the properties in (3.14) we have that the fiducial probability

$$\begin{aligned} 1-\alpha &= \mathrm{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(\{F : \Psi\{F(\cdot)-\hat{F}(\cdot)\} \le \epsilon_{n,\alpha}\}) \\ &= \mathrm{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(\{F : \Psi[n^{1/2}\{F(\cdot)-\hat{F}(\cdot)\}] \le \psi(n^{1/2})\epsilon_{n,\alpha}\}). \end{aligned} \tag{23}$$

By continuous mapping theorem and the fact that $\Psi[\{1-F_0(\cdot)\}W\{\gamma(\cdot)\}]$ is continuous and has unique $(1-\alpha)$-th quantile, Equation (23) converges to

$$\mathrm{pr}(\Psi[\{1-F_0(\cdot)\}W\{\gamma(\cdot)\}] \le \epsilon_\infty),$$

where ϵ_∞ is the unique limit of $\psi(n^{1/2})\epsilon_{n,\alpha}$, and pr is the sampling distribution of the data.

Then we have

$$\begin{aligned} \mathrm{pr}(F_0 \in \{F : \Psi\{F(\cdot)-\hat{F}(\cdot)\} \le \epsilon_{n,\alpha}\}) &= \mathrm{pr}(\Psi\{F_0(\cdot)-\hat{F}(\cdot)\} \le \epsilon_{n,\alpha}) \\ &= \mathrm{pr}(\Psi[n^{1/2}\{F_0(\cdot)-\hat{F}(\cdot)\}] \le \psi(n^{1/2})\epsilon_{n,\alpha}) \\ &\to \mathrm{pr}(\Psi[\{1-F_0(\cdot)\}W\{\gamma(\cdot)\}] \le \epsilon_\infty) \\ &= 1-\alpha. \end{aligned}$$

This completes the proof. □

Results for alternative selection schemes

Lemma 15. *The following modification of Theorem 6 is valid for S^L:*

$$\begin{aligned} pr^*_{\boldsymbol{y},\boldsymbol{\delta}}\{\sup_{s\le t}|S^L(s)-\hat{S}(s)| \ge \epsilon/n^{3/4} + 3\epsilon^2/n^{1/2} + \bar{N}(t)/\bar{K}(t)^{-2}\} \\ \le \{\bar{N}(t)+1\}(1-\epsilon/n^{3/4})^{\bar{K}(t)} + \bar{N}(t)[0{\cdot}4^{\bar{K}(t)} + n/\{\epsilon^2\bar{K}(t)\}^2]. \end{aligned} \tag{24}$$

The same bound also holds for S^I. Moreover, Theorem 7 holds for S^L and S^I.

Proof. Recall that $S^L(s) \ge S^U(s^+)$ and $S^L(s) \le S^U(s)$ hold for any $s \le t$, where s^+ denotes the next failure time right after s. Furthermore, the difference between $S^U(s)$ and $S^U(s^+)$ is bounded

by

$$|S^U(s) - S^U(s^+)| = |\prod_{i=1}^{\bar{N}(s)}\{1-B_i\} - \prod_{i=1}^{\bar{N}(s)}\{1-B_i\}(1-B_{\bar{N}(s)+1})|$$
$$= |\prod_{i=1}^{\bar{N}(s)}\{1-B_i\}B_{\bar{N}(s)+1}| \leq \max_{1\leq i\leq \bar{N}(s)+1} B_i,$$

where B_i follows $Beta(1, \bar{K}(s_i))$ and s_i are ordered failure times before or at time s. By Equation (17) in the previous section, we have

$$\text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(|S^U(s^+) - S^U(s)| > \epsilon/n^{3/4}) \leq \text{pr}^*_{\boldsymbol{y},\boldsymbol{\delta}}(\max_{1\leq i\leq \bar{N}(s)+1} B_i > \epsilon/n^{3/4})$$
$$\leq \{\bar{N}(t)+1\}(1-\frac{\epsilon}{n^{3/4}})^{\bar{K}(t)}.$$

Notice that $S^L(s) - \hat{S}(s) = \{S^L(s) - S^U(s)\} + \{S^U(s) - \hat{S}(s)\}$ and $S^U(s^+) - S^U(s) \leq S^L(s) - S^U(s) \leq 0$. This implies (24). In addition, Theorem 7 holds for S^L and S^I by Slutsky's theorem. □

Lemma 16. *For any failure time t, $E^*_{\boldsymbol{y},\boldsymbol{\delta}}[S^L(t)] \leq \tilde{S}(t) \leq E^*_{\boldsymbol{y},\boldsymbol{\delta}}[S^U(t)]$, where $E^*_{\boldsymbol{y},\boldsymbol{\delta}}$ is the expectation with respect to $pr^*_{\boldsymbol{y},\boldsymbol{\delta}}$, and $\tilde{S}(t)$ is the Kaplan-Meier estimator.*

Proof. For any failure time t, we have $S^U(t) = \prod_{i=1}^{\bar{N}(t)}\{1-B_i\}$. From here

$$E^*_{\boldsymbol{y},\boldsymbol{\delta}}[S^U(t)] = \prod_{i=1}^{\bar{N}(t)}\left\{1 - \frac{1}{1+\bar{K}(s_i)}\right\} \geq \prod_{i=1}^{\bar{N}(t)}\left\{1 - \frac{1}{\bar{K}(s_i)}\right\},$$

where s_i are ordered failure times. Similarly, $S^L(t) = S^U(t)(1-B)$, where B follows $Beta(1, \bar{K}(t)-1)$ and is independent of B_i for $i \leq \bar{N}(t)$. Thus

$$E^*_{\boldsymbol{y},\boldsymbol{\delta}}[S^L(t)] = \prod_{i=1}^{\bar{N}(t)}\left\{1 - \frac{1}{1+\bar{K}(s_i)}\right\}(1 - \frac{1}{\bar{K}(t)}) \leq \prod_{i=1}^{\bar{N}(t)}\left\{1 - \frac{1}{\bar{K}(s_i)}\right\}.$$

This completes the proof. □

Algorithm for sampling from the fiducial distribution

1. Generate $U = (u_1, \ldots, u_n)$ from $U(0,1)$ and sort them. Denote sorted values as $preU$.

2. Sort the data. Denote sorted data as $(y_1, \ldots, y_n)$ and $(\delta_1, \ldots, \delta_n)$.

3. Initialize $LowerFid = (0)_{n+1}$, $UpperFid = (1)_{n+1}$.

4. For $i = 1$ to n:

Let $UpperFid(i) = preU(1)$, where $preU(1)$ is the smallest element left in $preU$.

If $\delta = 1$, set $LowerFid(i+1) = preU(1)$, and delete $preU(1)$;

If $\delta = 0$, randomly pick one u from $preU$, set $LowerFid(i+1) = LowerFid(i)$, and delete the selected u from $preU$.

5. We output 3 survival functions that are needed for the conservative and log-linear interpolation methods.

5.1. Lower fiducial bound: using $LowerFid$ as a fiducial curve.

5.2. Upper fiducial bound: using $UpperFid$ as a fiducial curve.

5.3. Log-linear interpolation: Fit a continuous fiducial distribution by linear interpolation based on failure observations as described in Section 3.2.1. Then correct the linear interpolation at the censoring observations so that the upper fiducial bound on continuous distribution function (lower fiducial bound for survival function) is satisfied. Let y_{n-k} $(k = 0, 1, \ldots, n-1)$ denotes the last failure observation. We fit a single line after last uncensored observation and take the maximum of $s_0, s_1, \ldots, s_k$ as slope, where s_1 is the slope between $(y_{n-k}, \log u_{n-k})$ and $(y_{n-k+1}, \log u_{n-k+1})$, $\ldots$, s_k is the slope between $(y_{n-k}, \log u_{n-k})$ and $(y_n, \log u_n)$, s_0 is the slope between $(\tilde{y}, \log \tilde{u})$ and $(y_{n-k}, \log u_{n-k})$, $\tilde{y}$ is the second last uncensored observation. If there is only one failure time, $\tilde{y}$ and $\log \tilde{u}$ are 0.

6. From step 1–5 we get one curve of fiducial distribution. Repeat step 1–5 to get one fiducial sample with m curves.

APPENDIX C: SUPPLEMENTARY MATERIAL TO CHAPTER 4

A simplified tree-based survival model used in Theorem 8

We consider a simplified version of a tree-based survival model. Starting from the root node $[0,1]^d$, at each internal node, we randomly chose the j-th feature of X to split the node, while the splitting point is always at the midpoint of the range of the chosen feature. We repeat splitting $\lceil \log_2 k_n \rceil$ times, where k_n is a deterministic parameter which we can control. Hence, each individual tree has exactly $2^{\lceil \log_2 k_n \rceil}$ terminal nodes, which is approximately k_n. In practice, we always chose k_n to go to infinity as n goes to infinity.

After we build an individual tree, let B_i $(i = 1, 2, \ldots, 2^{\lceil \log_2 k_n \rceil})$ be the rectangular cell of the random partition. We treat observations inside each leaf node as a group of homogeneous subjects and compute the Nelson-Aalen estimator $\widehat{\Lambda}(\cdot \mid B_i)$ for each leaf node B_i. Hence, our estimator is essentially

$$\widehat{r}_n(\cdot, X, A) = \sum_{i=1}^{2^{\lceil \log_2 k_n \rceil}} I\{(X, A) \in B_i\} \widehat{\Lambda}(\cdot \mid B_i).$$

Proof of Theorem 8

Proof. Since we always assume that the treatment variable A is important, and A has only two categories, we force a split on A at the root node. This is equivalent to fitting trees for $A = 1$ and $A = -1$ separately. In a balanced design, the problem reduces to estimating $r(\cdot, X, 1)$ or $r(\cdot, X, -1)$ with sample size $n/2$. Without the risk of ambiguities, the following results are developed for $\widehat{r}_n(\cdot, X)$ with sample size n, where the results can be applied to either $A = 1$ or -1. Our proof utilizes two facts from [11]:

Fact 1 Let $K_{nj}\{B_i\}$ be the number of times the j-th coordinate $(j = 1, \ldots, d)$ is split on to reach the terminal node $B_i, (i = 1, 2, \ldots, 2^{\lceil \log_2 k_n \rceil})$. Conditionally on X, $K_{nj}\{B_i\}$ is $Binomial(\lceil \log_2 k_n \rceil, 1/d)$. Moreover, $\sum_{j=1}^{d} K_{nj}\{B_i\} = \lceil \log_2 k_n \rceil$.

Fact 2 Let $N_n(B_i)$ be the number of data points falling in the cell $B_i, (i = 1, 2, \ldots, 2^{\lceil \log_2 k_n \rceil})$. Conditionally on Θ, $N_n(B_i)$ follows $Binomial(n, 2^{-\lceil \log_2 k_n \rceil})$.

The following lemma, for later reference, provides the deterministic limit of the Nelson-Aalen estimator in the independent non-identically distributed case. This lemma can be found in an unpublished technical report by Mai Zhou at the University of Kentucky.

Lemma 17 (Theorem 1 in Chapter 2)**.** *Suppose we have two sets of non-negative random variables: $T_1, T_2, \ldots, T_n$ which are survival times, independent but non-identically distributed with continuous distribution $F_1(t), F_2(t), \ldots, F_n(t)$; $C_1, C_2, \ldots, C_n$ which are censoring times, independent but non-identically distributed with continuous distribution $G_1(t), G_2(t), \ldots, G_n(t)$. We also assume the $T_i's$ and $C_i's$ are independent. The Nelson-Aalen estimator of data $Y_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i)$ is $\widehat{\Lambda}(t)$. Provided Assumption 10, we have*

$$pr(\sup_{t<\tau} |\widehat{\Lambda}(t) - \int_0^t \frac{\sum_i \{1 - G_i(s)\} dF_i(s)}{\sum_i \{1 - G_i(s)\}\{1 - F_i(s)\}}| > \frac{(288b)^{1/2}}{n^{1/2} M^2}) < 16(n+2)e^{-b}, \tag{25}$$

where $b \geq 1$.

Now we start the proof of Theorem 8. Let the limit of the Nelson-Aalen estimator inside the cell B_i $(i = 1, 2, \ldots, 2^{\lceil \log_2 k_n \rceil})$ be

$$\Lambda^*(t \mid B_i) = \int_0^t \frac{[\sum_{X_j \in B_i} \{1 - G_j(s)\} dF_j(s)]}{[\sum_{X_j \in B_i} \{1 - G_j(s)\}\{1 - F_j(s)\}]}.$$

For any $t < \tau$, in order to bound the $|\widehat{r}_n(t, X) - r(t, X)|$, we define

$$r_n^*(t, X) = \sum_{i=1}^{2^{\lceil \log_2 k_n \rceil}} I\{X \in B_i\} \Lambda^*(t \mid B_i).$$

Then $|\widehat{r}_n(t, X) - r(t, X)|$ can be decomposed as

$$|\widehat{r}_n(t, X) - r(t, X)| = |\widehat{r}_n(t, X) - r_n^*(t, X)| + |r_n^*(t, X) - r(t, X)|. \tag{26}$$

We start with the first term in Equation (26). From Fact 2, we know the number of observations in each terminal node is $Binomial(n, 2^{-\lceil \log_2 k_n \rceil})$. By the Chernoff bound, with probability larger than $1 - e^{-u^2 n 2^{-\lceil \log_2 k_n \rceil - 1}}$, in one terminal node we have at least $(1-u)n2^{-\lceil \log_2 k_n \rceil}$ observations for some $0 < u < 1$.

Combining Equation (25), the following equation holds:

$$
\begin{aligned}
&|\widehat{r}_n(t,X) - r_n^*(t,X)| \\
&\qquad \leq \sum_{i=1}^{2^{\lceil \log_2 k_n \rceil}} I\{X \in B_i\}(288b)^{1/2}\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\}^{-1/2}M^{-2} \\
&\qquad = (288b)^{1/2}\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\}^{-1/2}M^{-2}, \qquad (27)
\end{aligned}
$$

with probability $1 - 16[(1-u)n2^{-\lceil \log_2 k_n \rceil} + 2]e^{-b} - e^{-u^2 n 2^{-\lceil \log_2 k_n \rceil - 1}}$, where $b \geq 1$.

Before we bound the second term in Equation (26). We first show the bound for the difference between the true cumulative hazard function and aggregated estimator inside the cell B_i ($i = 1, 2, \ldots, 2^{\lceil \log_2 k_n \rceil}$), i.e. $|I\{X \in B_i\}\{\Lambda^*(t \mid B_i) - \Lambda(t \mid X)\}|$.

From Fact 1, we know the number of times the terminal node B_i is split on the j-th coordinate ($j = 1, \cdots, d$) $K_{nj}\{B_i\}$ is $Binomial(\lceil \log_2 k_n \rceil, 1/d)$. By the Chernoff bound, $P(K_{nj}\{B_i\} \leq (1-r)\lceil \log_2 k_n \rceil / d) \leq e^{-\lceil \log_2 k_n \rceil r^2/(2d)}$ for some $0 < r < 1$. So with probability $(1 - e^{-\lceil \log_2 k_n \rceil r^2/(2d)})^d \geq 1 - de^{-\lceil \log_2 k_n \rceil r^2/(2d)}$, every dimension of B_i is less than $2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d}$. Then with probability larger than $1 - de^{-\lceil \log_2 k_n \rceil r^2/(2d)}$, we have

$$
\max_{X_1, X_2 \in B_i} ||X_1 - X_2|| \leq d^{1/2} 2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d}.
$$

So for all the observations X_j inside the same cell as X, by Assumption 11, we have

$$
|F_X(\cdot) - F_j(\cdot)| \leq L d^{1/2} 2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d},
$$

$$
|f_X(\cdot) - f_j(\cdot)| \leq (L' + L^2) d^{1/2} 2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d},
$$

where $f_X(\cdot)$ and $F_X(\cdot)$ denote the true density function and distribution function at X, respectively. Then $\Lambda^*(t \mid B_i)$ has the upper bound and lower bound

$$
\int_0^t [f_X(s) + b_1]/[1 - F_X(s) - b_2] ds \quad \text{and} \quad \int_0^t [f_X(s) - b_1]/[1 - F_X(s) + b_2] ds,
$$

respectively, where

$$b_1 = (L' + L^2)d^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} \quad \text{and} \quad b_2 = Ld^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d}.$$

Hence, $|I\{X \in B_i\}\{\Lambda^*(t \mid B_i) - \Lambda(t \mid X)\}|$ has the bound

$$\int_0^t \frac{b_1(1-F(s)) + b_2 f(s)}{(1-F(s)-b_2)(1-F(s))} ds \leq C\tau d^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d},$$

where C is some constant depending on L and L'. We then bound the second term of Equation (26) as follows:

$$\begin{aligned} |r_n^*(t,X) - r(t,X)| &\leq \sum_{i=1}^{2^{\lceil \log_2 k_n \rceil}} I\{X \in B_i\}|\Lambda^*(t \mid B_i) - \Lambda(t \mid X)| \\ &\leq C\tau d^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d}. \end{aligned} \tag{28}$$

Combining Equation (27) and (28), For each X, we have

$$\begin{aligned} \text{pr}[\sup_{t<\tau} |\widehat{r}_n(t,X) - r(t,X)| \leq C[&\tau d^{1/2}2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} \\ &+ (288b)^{1/2}\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\}^{-1/2}M^{-2}] \geq 1 - w_n, \end{aligned}$$

where

$$w_n = \ 16[(1-u)n2^{-\lceil \log_2 k_n \rceil} + 2]e^{-b} + e^{-u^2 n 2^{-\lceil \log_2 k_n \rceil - 1}} + de^{-\lceil \log_2 k_n \rceil r^2/(2d)}.$$

This completes the proof. □

Proof of Theorem 9

Proof. Based on Theorem 8, we now only need to establish the bound of $|\widehat{r}_n(t,X,A) - r(t,X,A)|$ under the event with small probability w_n. Noticing that $\widehat{r}_n(t,X,A)$ is simply the Nelson-Aalen estimator of the cumulative hazard function with at most n terms, for any

$t < \tau$ we have

$$\widehat{r}_n(t, X, A) \leq \frac{1}{n} + \ldots + \frac{1}{1} = O(\ln(n)),$$

which implies that

$$|\widehat{r}_n(t, X, A) - r(t, X, A)| \leq O(\ln(n)).$$

Combining this with Theorem 8 completes the proof. □

Proof of Lemma 6

Proof. Our survival function estimator is $\widehat{S}(t) = e^{-\widehat{\Lambda}(t)}$. From Theorem 8, we know that for any $t < \tau$,

$$\begin{aligned} &pr(|\widehat{S}(t \mid X, A) - S(t \mid X, A)| \leq C[2^{-(1-r)\lceil \log_2 k_n \rceil/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2}]) \\ &\geq 1 - 16[(1-u)n2^{-\lceil \log_2 k_n \rceil} + 2]e^{-b} - e^{-u^2 n 2^{-\lceil \log_2 k_n \rceil - 1}} - de^{-\lceil \log_2 k_n \rceil r^2/(2d)}. \end{aligned}$$

It is then easy to see that for R_1,

$$\begin{aligned} &\left|\widehat{E}(T \mid X, A) - E(T \mid X, A)\right| \\ &= \left| \int_0^\tau \widehat{S}(t \mid X, A)dt - \int_0^\tau S(t \mid X, A)dt \right| \\ &\leq \int_0^\tau |\widehat{S}(t \mid X, A) - S(t \mid X, A)|dt \\ &\leq \tau C[2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2}], \end{aligned}$$

with probability larger than $1 - w_n$. And for reward R_2, we have

$$\begin{aligned}
&\left|\widehat{E}(T \mid X, A, T > Y, Y) - E(T \mid X, A, T > Y, Y)\right| \\
&= \Big| \int_Y^\tau \{\widehat{S}(t \mid X, A)/\widehat{S}(Y \mid X, A)\}dt - \int_Y^\tau \{S(t \mid X, A)/S(Y \mid X, A)\}dt \Big| \\
&\leq \Big| \int_Y^\tau \{\widehat{S}(t \mid X, A)/\widehat{S}(Y \mid X, A)\}dt - \int_Y^\tau \{\widehat{S}(t \mid X, A)/S(Y \mid X, A)\}dt \Big| \\
&+ \Big| \int_Y^\tau \{\widehat{S}(t \mid X, A)/S(Y \mid X, A)\}dt - \int_Y^\tau \{S(t \mid X, A)/S(Y \mid X, A)\}dt \Big|.
\end{aligned}$$

Note that we can bound the distance between $\widehat{S}(Y \mid X, A)$ and $S(Y \mid X, A)$ with probability no less than $1 - w_n$, which is further bounded above by

$$\begin{aligned}
&(1/M^2 + 1/M) \int_Y^\tau |\widehat{S}(Y \mid X, A) - S(Y \mid X, A)|dt \\
&\qquad \leq C_2[2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2}],
\end{aligned}$$

for some constant C_2 with probability larger than $1 - 2w_n$. □

Proof of Theorem 10

Proof. We restate the value function corresponding to the true and working model as

$$\begin{aligned}
V(f) &= E(RI[A = \text{sign}\{f(X)\}]/\pi(A; X)) \\
\text{and } V'(f) &= E(\widehat{R}I[A = \text{sign}\{f(X)\}]/\pi(A; X)),
\end{aligned}$$

respectively. Then we have

$$\begin{aligned}
V(f^*) - V(\widehat{f}_n) &\leq V(f^*) - \sup_{f \in \mathcal{F}} V'(f) + \sup_{f \in \mathcal{F}} V'(f) - V'(\widehat{f}_n) + V'(\widehat{f}_n) - V(\widehat{f}_n) \\
&\leq V(f^*) - V'(f^*) + \sup_{f \in \mathcal{F}} V'(f) - V'(\widehat{f}_n) + V'(\widehat{f}_n) - V(\widehat{f}_n) \\
&\leq \sup_{f \in \mathcal{F}} V'(f) - V'(\widehat{f}_n) + 2 \sup_{f \in \mathcal{F}} |V(f) - V'(f)|. \qquad (29)
\end{aligned}$$

We start with the first term in Equation (29). From Lemma 5, we know that $\sup_{f\in\mathcal{F}} V'(f) - V'(\widehat{f}_n) = V'(\widetilde{f}) - V'(\widehat{f}_n)$, where $\widetilde{f} = \arg\min_{f\in\mathcal{F}} E\{L_\phi(f)\}$.

Let $\widetilde{f}_{\lambda_n} = \arg\min_{f\in\mathcal{H}_k}[E\{R\phi\{Af(X)\}/\pi(A;X)\} + \lambda_n\|f\|_k^2]$, then

$$n^{-1}\sum_{i=1}^{n}\frac{\widehat{R}\phi\{A_i\widehat{f}_n(X_i)\}}{\pi(A_i;X_i)} + \lambda_n\|\widehat{f}_n\|_k^2 \le n^{-1}\sum_{i=1}^{n}\frac{\widehat{R}\phi\{A_i\widetilde{f}_{\lambda_n}(X_i)\}}{\pi(A_i;X_i)} + \lambda_n\|\widetilde{f}_{\lambda_n}\|_k^2. \tag{30}$$

By the definition of $a(\lambda)$, we have

$$a(\lambda_n) = [E\{L_\phi(\widetilde{f}_{\lambda_n})\} + \lambda_n||\widetilde{f}_{\lambda_n}||_k^2 - E\{L_\phi(\widetilde{f})\}],$$

and by Theorem 3.2 in [135], we further have

$$\begin{aligned}
V'(\widetilde{f}) - V'(\widehat{f}_n) &\le E\{L_\phi(\widehat{f}_n)\} - E\{L_\phi(\widetilde{f})\} \\
&\le E\{L_\phi(\widehat{f}_n)\} - E\{L_\phi(\widetilde{f}_{\lambda_n})\} - \lambda_n||\widetilde{f}_{\lambda_n}||_k^2 \\
&\quad + E\{L_\phi(\widetilde{f}_{\lambda_n})\} - E\{L_\phi(\widetilde{f})\} + \lambda_n||\widetilde{f}_{\lambda_n}||_k^2 \\
&\le E\{L_\phi(\widehat{f}_n)\} - E\{L_\phi(\widetilde{f}_{\lambda_n})\} - \lambda_n||\widetilde{f}_{\lambda_n}||_k^2 + \lambda_n||\widehat{f}_n||_k^2 + a(\lambda_n).
\end{aligned}$$

Combined with (30),

$$\begin{aligned}
V'(\widetilde{f}) - V'(\widehat{f}_n) \le\ & a(\lambda_n) + E\left[\frac{R\phi\{A\widehat{f}_n(X)\}}{\pi(A;X)} - \frac{\widehat{R}\phi\{A\widehat{f}_n(X)\}}{\pi(A;X)}\right] \\
& + E\left[\frac{\widehat{R}\phi\{A\widetilde{f}_{\lambda_n}(X)\}}{\pi(A;X)} - \frac{R\phi\{A\widetilde{f}_{\lambda_n}(X)\}}{\pi(A;X)}\right] \\
& + \Bigg(- n^{-1}\sum_{i=1}^{n}\Big[\lambda_n\|\widehat{f}_n\|_k^2 + \frac{\widehat{R}\phi\{A_i\widehat{f}_n(X_i)\}}{\pi(A_i;X_i)} - \lambda_n\|\widetilde{f}_{\lambda_n}\|_k^2 - \frac{\widehat{R}\phi\{A_i\widetilde{f}_{\lambda_n}(X_i)\}}{\pi(A_i;X_i)}\Big] \\
& + E\Big[\lambda_n\|\widehat{f}_n\|_k^2 + \frac{\widehat{R}\phi\{A\widehat{f}_n(X)\}}{\pi(A;X)} - \lambda_n\|\widetilde{f}_{\lambda_n}\|_k^2 - \frac{\widehat{R}\phi\{A\widetilde{f}_{\lambda_n}(X)\}}{\pi(A;X)}\Big]\Bigg) \\
=\ & a(\lambda_n) + (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).
\end{aligned}$$

Since

$$n^{-1}\sum_{i=1}^{n}\frac{\widehat{R}\phi\{A_i\widehat{f}_n(X_i)\}}{\pi(A_i;X_i)} + \lambda_n\|\widehat{f}_n\|_k^2 \le n^{-1}\sum_{i=1}^{n}\frac{\widehat{R}\phi(0)}{\pi(A_i;X_i)} = n^{-1}\sum_{i=1}^{n}\frac{\widehat{R}}{\pi(A_i;X_i)},$$

and the estimated value function $\widehat{R}$ is bounded by τ, we know that $\|\widehat{f}_n\|_k \leq \tau^{1/2}\lambda_n^{-1/2}$. Furthermore, since

$$\lambda_n \|\widetilde{f}_{\lambda_n}\|_k^2 \leq \inf_{f \in \mathcal{H}_k} \left\{ \lambda_n \|f\|_k^2 + E\left[\frac{R\phi\{Af(X)\}}{\pi(A;X)} \right] \right\} \leq E\left[\frac{R\phi(0)}{\pi(A;X)} \right],$$

we have $\|\widetilde{f}_{\lambda_n}\|_k \leq \tau^{1/2}\lambda_n^{-1/2}$. Combining with Lemma 6, |I| and |II| are bounded by $C_1\lambda_n^{-1/2}\{2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2} + w_n \ln n\}$ for both R_1 and R_2, where C_1 is some constant. Following the results in [138], |III| is bounded by $M_v(n\lambda_n/c_n)^{-2/(v+2)} + M_v\lambda_n^{-1/2}(c_n/n)^{2/(d+2)} + K\rho(n\lambda_n)^{-1} + 2K\rho n^{-1}\lambda_n^{-1/2}$ with probability larger than $1 - 2e^{-\rho}$, where M_v is a constant depending on v and K is a sufficiently large positive constant. Finally, combining (I), (II) and (III), we have

$$\operatorname{pr}(\sup_{f \in \mathcal{F}} V'(f) \leq V'(\widehat{f}_n) + \epsilon_1) \geq 1 - 2e^{-\rho}, \tag{31}$$

where $\epsilon_1 = a(\lambda_n) + M_v(n\lambda_n/c_n)^{-2/(v+2)} + M_v\lambda_n^{-1/2}(c_n/n)^{2/(d+2)} + K\rho(n\lambda_n)^{-1} + 2K\rho n^{-1}\lambda_n^{-1/2} + C_1\lambda_n^{-1/2}\{2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2} + w_n \ln n\}$.

For the second part in Equation (29),

$$\begin{aligned} V(f) - V'(f) &= E\Big(\frac{RI[A = \operatorname{sign}\{f(X)\}]}{\pi(A;X)}\Big) - E\Big(\frac{\widehat{R}I[A = \operatorname{sign}\{f(X)\}]}{\pi(A;X)}\Big) \\ &= E\Big(\{E(T \mid X, A) - \widehat{E}(T \mid X, A)\}\frac{I[A = \operatorname{sign}\{f(X)\}]}{\pi(A;X)}\Big) \end{aligned}$$

if $R = R_1$. For $R = R_2$, we have

$$\begin{aligned} &V(f) - V'(f) \\ &= E\Big((1-\delta)\{E(T \mid X, A, T > Y, Y) - \widehat{E}(T \mid X, A, T > Y, Y)\}\frac{I[A = \operatorname{sign}\{f(X)\}]}{\pi(A;X)}\Big). \end{aligned}$$

By Lemma 6,

$$\begin{aligned} &\sup_{f \in \mathcal{F}} |V(f) - V'(f)| \\ \leq &C_2\lambda_n^{-1/2}\{2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2} + w_n \ln n\}, \end{aligned} \tag{32}$$

Table A.1: Simulation results: Mean ($\times 10^3$) and (sd) ($\times 10^3$). Censoring rate: 30%. For each scenario, the theoretical optimal value ($\times 10^3$) is 31, 181, 1079, and -389, respectively.

	kernel	T	RIST-R_1	RIST-R_2	ICO	DR	Cox	
1	Linear	0 (26)	1 (31)	2 (28)	-10 (40)	-20 (63)	-26 (33)	
	Gaussian	-17 (44)	-10 (34)	-7 (37)	-18 (45)	-48 (65)		
2	Linear	22 (113)	17 (105)	-14 (126)	-110 (136)	-193 (133)	65 (63)	T: using
	Gaussian	-39 (115)	-25 (101)	-62 (113)	-164 (119)	-285 (112)		
3	Linear	785 (52)	768 (53)	771 (52)	737 (95)	667 (124)	763 (61)	
	Gaussian	896 (61)	810 (54)	854 (69)	817 (124)	679 (123)		
4	Linear	-453 (37)	-465 (46)	-448 (27)	-461 (42)	-471 (54)	-457 (32)	
	Gaussian	-465 (35)	-477 (42)	-456 (27)	-474 (41)	-505 (48)		

true survival time as weight; RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning; Cox: Cox proportional hazards model using covariate-treatment interactions.

where C_2 is some constant. Now, combining (31) and (32) we have

$$\mathrm{pr}(V(f^*) \leq V(\widehat{f}_n) + \epsilon) \geq 1 - 2e^{-\rho},$$

where

$$\begin{aligned}\epsilon = \; & a(\lambda_n) + M_v(n\lambda_n/c_n)^{-2/(v+2)} + M_v\lambda_n^{-1/2}(c_n/n)^{2/(d+2)} + K\rho(n\lambda_n)^{-1} \\ & + 2K\rho n^{-1}\lambda_n^{-1/2} + C\lambda_n^{-1/2}\{2^{-\{(1-r)\lceil \log_2 k_n \rceil\}/d} + (b/\{(1-u)n2^{-\lceil \log_2 k_n \rceil}\})^{1/2} \\ & + w_n \ln n\}.\end{aligned}$$

This completes the proof. □

Additional simulation results for different censoring rates

We summarize the additional simulation results in this section. For each simulation scenario considered in Section 4.4, we alter the first constant term in the censoring distribution to achieve 30% (Table A.1 and Figure A.1), and 60% (Table A.2 and Figure A.2) censoring rates.

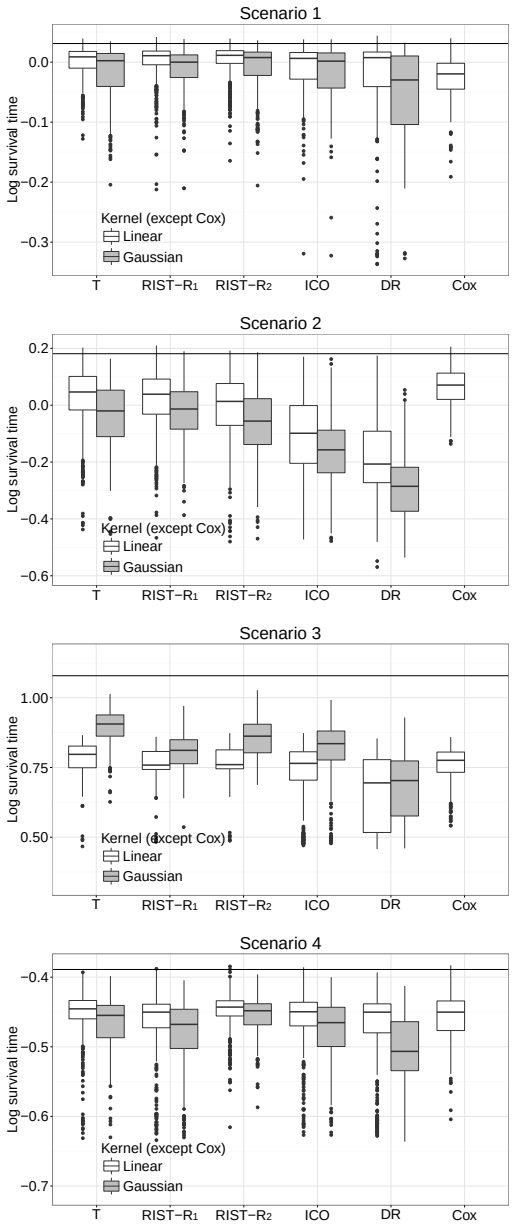


Figure A.1: Boxplots of mean log survival time for different treatment regimes. Censoring rate: 30%. T: using true survival time as weight; RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning. The black horizontal line is the theoretical optimal value.

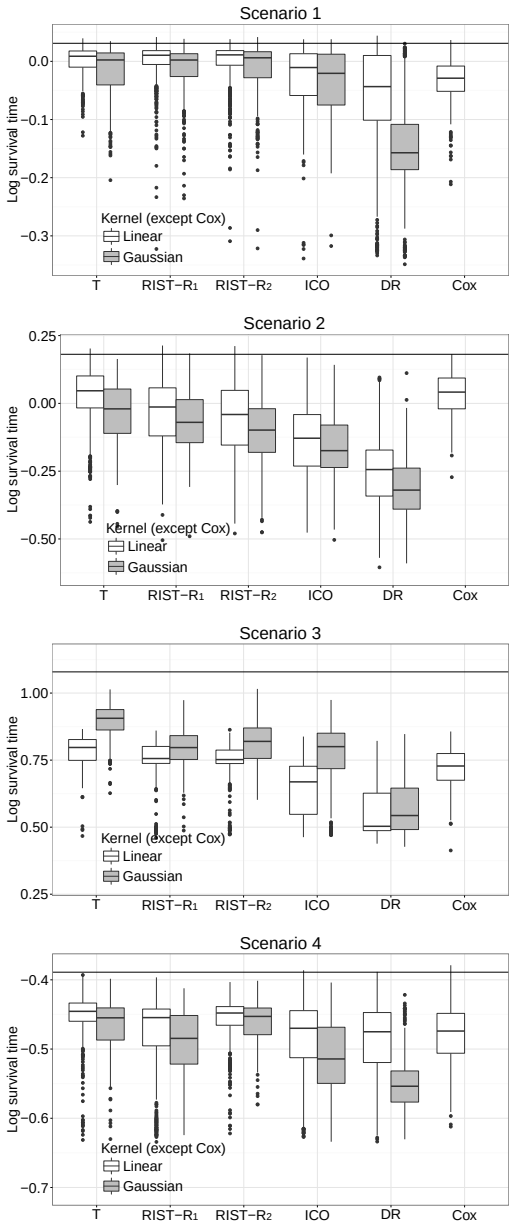


Figure A.2: Boxplots of mean log survival time for different treatment regimes. Censoring rate: 60%. T: using true survival time as weight; RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning. The black horizontal line is the theoretical optimal value.

Table A.2: Simulation results: Mean ($\times 10^3$) and (sd) ($\times 10^3$). Censoring rate: 60%. For each scenario, the theoretical optimal value ($\times 10^3$) is 31, 181, 1079, and -389, respectively.

	kernel	T	RIST-R_1	RIST-R_2	ICO	DR	Cox	
1	Linear	0 (26)	-2 (39)	-5 (43)	-29 (57)	-64 (92)	-34 (36)	
	Gaussian	-17 (44)	-12 (40)	-12 (45)	-35 (55)	-144 (78)		
2	Linear	22 (113)	-36 (123)	-61 (135)	-138 (133)	-248 (129)	31 (79)	T: using
	Gaussian	-39 (115)	-69 (108)	-102 (115)	-165 (117)	-313 (101)		
3	Linear	785 (52)	753 (77)	748 (69)	646 (104)	556 (94)	721 (70)	
	Gaussian	896 (61)	796 (63)	819 (67)	775 (106)	573 (93)		
4	Linear	-453 (37)	-478 (55)	-458 (33)	-486 (55)	-492 (59)	-480 (43)	
	Gaussian	-465 (35)	-492 (48)	-461 (29)	-513 (53)	-551 (38)		

true survival time as weight; RIST-R_1 and RIST-R_2: using the estimated R_1 and R_2 respectively as weights, while the conditional expectations are estimated using recursively imputed survival trees; ICO: inverse probability of censoring weighted learning; DR: doubly robust outcome weighted learning; Cox: Cox proportional hazards model using covariate-treatment interactions.

BIBLIOGRAPHY

[1] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

[2] Ahn, H. and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics*, pages 471–485.

[3] Akritas, M. G. (1986). Bootstrapping the kaplan-meier estimator. *Journal of the American Statistical Association*, 81(396):1032–1038.

[4] Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11.

[5] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120.

[6] Arlot, S. and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

[7] Barber, S. and Jennison, C. (1999). Symmetric tests and confidence intervals for survival probabilities and quantiles of censored survival data. *Biometrics*, 55(2):430–436.

[8] Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40:1550–1577.

[9] Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37:905–938.

[10] Berger, J. O., Bernardo, J. M., and Sun, D. (2012). Objective Priors for Discrete Parameter Spaces. *Journal of the American Statistical Association*, 107:636–648.

[11] Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095.

[12] Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

[13] Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

[14] Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2010). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367.

[15] Borkowf, C. B. (2005). A simple hybrid variance estimator for the kaplan–meier survival function. *Statistics in medicine*, 24(6):827–851.

[16] Bou-Hamad, I., Larocque, D., Ben-Ameur, H., et al. (2011). A review of survival trees. *Statistics Surveys*, 5:44–71.

[17] Bouliotis, G. and Billingham, L. (2011). Crossing survival curves: alternatives to the log-rank test. *Trials*, 12(1):1.

[18] Bredensteiner, E. J. and Bennett, K. P. (1999). Multicategory classification by support vector machines. In *Computational Optimization*, pages 53–79. Springer.

[19] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[20] Breiman, L. (2004). Consistency for a simple model of random forests.

[21] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

[22] Breslow, N., Crowley, J., et al. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453.

[23] Brillinger, D. R. (1962). Examples bearing on the definition of fiducial probability with a bibliography. *Ann. Math. Statist.*, 33(4):1349–1355.

[24] Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 2nd edition.

[25] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[26] Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

[27] Ciampi, A., Chang, C.-H., Hogg, S., and McKinney, S. (1987). Recursive partition: A versatile method for exploratory-data analysis in biostatistics. In *Biostatistics*, pages 23–50. Springer.

[28] Ciampi, A., Thiffault, J., Nakache, J.-P., and 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.

[29] Ciampi, A., Thiffault, J., and Sagman, U. (1989). Recpam: a computer program for recursive partition amalgamation for censored survival data and other situations frequently occurring in biostatistics. ii. applications to data on small cell carcinoma of the lung (sccl). *Computer methods and programs in biomedicine*, 30(4):283–296.

[30] Cisewski, J. and Hannig, J. (2012). Generalized fiducial inference for normal linear mixed models. *The Annals of Statistics*, 40:2102–2127.

[31] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

[32] Cui, Y., Zhu, R., and Kosorok, M. (2017a). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electron. J. Statist.*, 11(2):3927–3953.

[33] Cui, Y., Zhu, R., Zhou, M., and Kosorok, M. (2017b). Some asymptotic results of survival tree and forest models. *arXiv preprint:1707.09631*.

[34] Cuzick, J. (1985). Asymptotic properties of censored linear rank tests. *The Annals of Statistics*, pages 133–141.

[35] Dardis, C. (2016). *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.4.

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

[37] Dempster, A. P. (2008). The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48:365–377.

[38] Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

[39] Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

[40] Edlefsen, P. T., Liu, C., and Dempster, A. P. (2009). Estimating limits from Poisson counting data using Dempster–Shafer analysis. *The Annals of Applied Statistics*, 3:764–790.

[41] Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319.

[42] Eng, K. H. and Kosorok, M. R. (2005). A sample size formula for the supremum log-rank statistic. *Biometrics*, 61(1):86–91.

[43] Fan, J., Su, X.-G., Levine, R. A., Nunn, M. E., and LeBlanc, M. (2006). Trees for correlated survival data by goodness of split, with applications to tooth prognosis. *Journal of the American Statistical Association*, 101(475):959–967.

[44] Fay, M. P. and Brittain, E. H. (2016). Finite sample pointwise confidence intervals for a survival distribution with right-censored data. *Statistics in medicine*.

[45] Fay, M. P., Brittain, E. H., and Proschan, M. A. (2013). Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics*, page kxt016.

[46] Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700 – 725.

[47] Fisher, R. A. (1930). Inverse Probability. *Proceedings of the Cambridge Philosophical Society*, xxvi:528–535.

[48] Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London series A*, 139:343–348.

[49] Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. *The Annals of Eugenics*, VI:91–98.

[50] Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

[51] Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association*, 82(397):312–320.

[52] Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–223.

[53] Geng, Y., Zhang, H. H., and Lu, W. (2015). On optimal treatment regimes selection for mean survival time. *Statistics in medicine*, 34(7):1169–1184.

[54] Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562.

[55] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

[56] Gill, R. D. (1980). Censoring and stochastic integrals. *Statistica Neerlandica*, 34(2):124–124.

[57] Gill, R. D. (1994). Glivenko-cantelli for kaplan-meier. *Mathematical Methods of Statistics*, 3:76–87.

[58] Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *Annals of statistics*, 40(1):529.

[59] Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer treatment reports*, 69(10):1065–1069.

[60] Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511.

[61] Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19:491–544.

[62] Hannig, J. (2013). Generalized Fiducial Inference via Discretization. *Statistica Sinica*, 23:489–514.

[63] Hannig, J., Feng, Q., Iyer, H. K., Wang, J. C.-M., and Liu, X. (2018). Fusion learning for inter-laboratory comparisons. *Journal of Statistical Planning and Inference*, page to appear.

[64] Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, (just-accepted).

[65] Hannig, J., Iyer, H. K., and Wang, J. C.-M. (2007). Fiducial approach to uncertainty assessment: accounting for error due to instrument resolution. *Metrologia*, 44:476–483.

[66] Hannig, J. and Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*, 96:847 – 860.

[67] Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, pages 553–566.

[68] Hjort, N. L. and Schweder, T. (2018). Confidence distributions and related themes. *Journal of Statistical Planning and Inference*, 195:1–13.

[69] Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

[70] Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in medicine*, 23(1):77–91.

[71] Huang, L., Jin, Y., Gao, Y., Thung, K.-H., Shen, D., Initiative, A. D. N., et al. (2016). Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of aging*, 46:180–191.

[72] Ishwaran, H. and Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & probability letters*, 80(13):1056–1064.

[73] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, pages 841–860.

[74] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[75] Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

[Krętowska] Krętowska, Małgorzata, b. p. y. o. Random forest of dipolar trees for survival prediction.

[77] Krętowska, M. (2004). Dipolar regression trees in survival analysis. *Biocybernetics and biomedical engineering*, 24:25–33.

[78] Laber, E. and Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514.

[79] Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225.

[80] Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015). Generalized fiducial inference for ultra-high dimensional regression. *Journal of American Statistical Association*. To appear.

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

[82] Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81.

[83] Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

[84] Linero, A. R. (2016). Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association*, (just-accepted).

[85] Liu, K.-H. and Huang, D.-S. (2008). Cancer classification using rotation forest. *Computers in biology and medicine*, 38(5):601–610.

[86] Liu, Y. and Hannig, J. (2017). Generalized fiducial inference for logistic graded response models. *Psychometrika*, page to appear.

[87] Loh, W.-Y. (1991). Survival modeling through recursive stratification. *Computational statistics & data analysis*, 12(3):295–313.

[88] Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386.

[89] Ma, J., Hobbs, B. P., and Stingo, F. C. (2015). Statistical methods for establishing personalized treatment rules in oncology. *BioMed research international*, 2015.

[90] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50(3):163–170.

[91] Martin, R. and Liu, C. (2015). *Inferential Models: Reasoning with Uncertainty*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

[92] Mentch, L. and Hooker, G. (2014). Ensemble trees and clts: Statistical inference for supervised learning. *stat*, 1050:25.

[93] Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer.

[94] 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(1):154–177.

[95] Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

[96] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

[97] Nair, V. N. (1984). Confidence bands for survival functions with censored data: a comparative study. *Technometrics*, 26(3):265–275.

[98] Nash, R. A., McSweeney, P. A., Crofford, L. J., Abidi, M., Chen, C.-S., Godwin, J. D., Gooley, T. A., Holmberg, L., Henstorf, G., LeMaistre, C. F., et al. (2007). High-dose immunosuppressive therapy and autologous hematopoietic cell transplantation for severe systemic sclerosis: long-term follow-up of the us multicenter pilot study. *Blood*, 110(4):1388–1396.

[99] Nelson, W. (1969). Hazard plotting for incomplete failure data(multiply censored data plotting on various type hazard papers for engineering information on time to failure distribution). *Journal of Quality Technology*, 1:27–52.

[100] Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, J., Portier, K., and Greer, G. (2013). Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In *Social Intelligence and Technology (SOCIETY), 2013 International Conference on*, pages 109–113. IEEE.

[101] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207.

[102] Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

[103] Praestgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability*, 21(4):2053–2086.

[104] Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

[105] Rice, T. W., Rusch, V. W., Ishwaran, H., and Blackstone, E. H. (2010). Cancer of the esophagus and esophagogastric junction. *Cancer*, 116(16):3763–3773.

[106] Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630.

[107] Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, 9(1):130–134.

[108] Schein, P. S. (1982). A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. *Cancer*, 49(9):1771–1777.

[109] Schweder, T. and Hjort, N. L. (2016). *Confidence, likelihood, probability*, volume 41. Cambridge University Press.

[110] Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

[111] Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, pages 35–47.

[112] Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press Princeton.

[113] Socinski, M. A., Schell, M. J., Peterman, A., Bakri, K., Yates, S., Gitten, R., Unger, P., Lee, J., Lee, J.-H., Tynan, M., et al. (2002). Phase iii trial comparing a defined duration of therapy versus continuous therapy followed by second-line therapy in advanced-stage iiib/iv non–small-cell lung cancer. *Journal of Clinical Oncology*, 20(5):1335–1343.

[114] Starling, R. C., Moazami, N., Silvestry, S. C., Ewald, G., Rogers, J. G., Milano, C. A., Rame, J. E., Acker, M. A., Blackstone, E. H., Ehrlinger, J., et al. (2014). Unexpected abrupt increase in left ventricular assist device thrombosis. *New England Journal of Medicine*, 370(1):33–40.

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

[116] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, pages 575–607.

[117] Strawderman, R. L., Parzen, M. I., and Wells, M. T. (1997). Accurate confidence limits for quantiles under random censoring. *Biometrics*, pages 1399–1415.

[118] Strawderman, R. L. and Wells, M. T. (1997). Accurate bootstrap confidence limits for the cumulative hazard and survivor functions under random censoring. *Journal of the American Statistical Association*, 92(440):1356–1374.

[119] Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *The Annals of Statistics*, 21(3):1591–1607.

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

[121] Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, pages 156–160.

[122] Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

[123] Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70(352):865–871.

[124] Tian, L., Zhao, L., and Wei, L. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2):222–233.

[125] Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625–1651.

[126] Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

[127] Wandler, D. V. and Hannig, J. (2012). Generalized Fiducial Confidence Intervals for Extremes. *Extremes*, 15:67–87.

[128] Wang, J. C.-M., Hannig, J., and Iyer, H. K. (2012). Pivotal methods in the propagation of distributions. *Metrologia*, 49:382–389.

[129] Wang, J. C.-M. and Iyer, H. K. (2005). Propagation of uncertainties in measurements using generalized inference. *Metrologia*, 42:145–153.

[130] Wang, J. C.-M. and Iyer, H. K. (2006a). A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. *Measurement*, 39:856–863.

[131] Wang, J. C.-M. and Iyer, H. K. (2006b). Uncertainty analysis for vector measurands using fiducial inference. *Metrologia*, 43:486–494.

[132] Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81:3 – 39.

[133] Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

[134] Zhang, H. (1995). Splitting criteria in survival trees. In *Statistical Modelling*, pages 305–313. Springer.

[135] Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

[136] Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433.

[137] Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015a). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598.

[138] Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015b). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

[139] Zhou, M. (1991). Some properties of the kaplan-meier estimator for independent nonidentically distributed random variables. *The Annals of Statistics*, pages 2266–2274.

[140] Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, (just-accepted):00–00.

[141] Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340.

[142] Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.