Nuisance Parameter Estimation in Survival Models

A Dissertation submitted to the faculty of the University of Minnesota by

Andrew Wey

In partial fulfillment of the requirements for the degree of Doctor of Philosophy

Kyle Rudser and John Connett

August 2014

© Andrew Wey, 2014

Acknowledgements

I want to thank the Biostatistical Design and Analysis Center (BDAC), and all associated individuals, for the opportunity to work on many interesting projects. I enjoyed the constant challange of clincial collaborations that regularly extend past statistics. In addition, this thesis was inspired, in part, by the difficulties encountered in day-to-day collaborations with clinicians. Kyle Rudser deserves specific thanks for many helpful suggestions and ensuring I stayed on course throughout the past several years. Lastly, I want to acknowledge individuals who helped prior to graduate school: Steve Keppen at Brookings High School; Engin Sungur and Peh Ng at the Unversity of Minnesota, Morris.

"But I'll tell you a terrible secret. Are you listening to me? There isn't anyone out there who isn't Seymour's Fat Lady... There isn't anyone anywhere that isn't Seymour's Fat Lady. Don't you know that? Don't you know that goddamn secret yet? And don't you know - listen to me, now - don't you know who that Fat Lady really is? Ah, buddy. Ah, buddy. It's Christ Himself. Christ Himself, buddy."

-Zooey Glass

UNIVERSITY OF MINNESOTA

Abstract

Kyle Rudser and John Connett
School of Public Health

Doctor of Philosophy

Nuisance Parameter Estimation in Survival Models

by Andrew Wey

The Cox proportional hazards model is the most commonly used method for right-censored data. Unfortunately, the interpretation of the hazard ratio is non-intuitive and it is commonly misinterpreted as a relative risk. This motivates alternative methods with more intuitive interpretations that avoid a reliance on the proportional hazards assumption. We explore censored quantile regression and restricted mean treatment effects as potential alternatives to the Cox proportional hazards model. However, both alternatives require a conditional survival function as a nusiance parameter.

This thesis focuses on the impact of the conditional survival function on the estimation of censored quantile regression and restricted means. In particular, we illustrate that a non-parametric estimator of the conditional survival function improves the estimation of censored quantile regression when the semi-parametric assumptions of current methods are badly violated. Unfortunately, the non-parametric estimator is inefficient when the semi-parametric assumptions are satisfied. Rather than rely on either parametric assumptions or a non-parametric model, we instead pursue an estimator that performs well in both situations.

We propose estimating the conditional survival function with stacked survival models. By minimizing prediction error, stacked survival models estimate an optimally weighted combination of several survival models. This allows stacking to span parametric, semi-parametric, and non-parametric models to estimate the conditional survival function. As such, stacking can give weight to approximately correct parametric models, but shifts weight to non-parametric models when assumptions are badly violated. We demonstrate that the stacked survival model improves estimation of conditional survival functions and found it to always outperform the model selected by cross-validation. In addition, we illustrate that stacked survival models improve the estimation of restricted mean treatment effects in a wide variety of situations, while maintaining the efficiency of current methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistics plays a large role in collaborative clinical research. Yet the translation of statistical results into a clinically meaningful language remains a common challenge. One approach that can help communication when comparing groups is the careful consideration of an appropriate summary measure. Unfortunately, summary measures that possess nice mathematical properties do not always possess intuitive interpretations. For example, the odds ratio is commonly misinterpreted by clinicians as the more intuitive relative risk, which can be misleading when the event under investigation is not rare. A statistician could instead use relative risk regression, but the estimation procedures are fraught with mathematical and, in turn, computational difficulties [Fitzmaurice et al., 2014]. This illustrates an unfortunate dichotomy in that a preferred summary measure faces mathematical or computational difficulties, while the easy-to-estimate summary measure possesses a difficult interpretation.

A similar situation is encountered in survival analysis: the Cox proportional hazards model [Cox, 1972] possesses nice mathematical properties under censoring that allows relatively straightforward extensions to more complicated situations (e.g., competing risks). However, many authors have pointed out that the hazard ratio is difficult to interpret. In fact, the hazard ratio is commonly interpreted by clinicians as a relative risk even though the hazard function does not possess a probabilistic interpretation. Although, a more significant problem with the hazard ratio is that the interpretation does not directly relate back to survival time. In particular, in the absence of censoring, we would likely estimate a multiplicative or additive change in mean survival time. Yet the hazard ratio, and the hazard function, have a complicated relationship with average survival time. For more discussion, Rudser et al. [2012] and Royston and Parmar [2013] provide a comprehensive overview of the difficulties associated with interpreting the hazard ratio.

This thesis explores alternatives to the Cox model and the corresponding hazard ratio for comparing survival between groups. In particular, we investigate censored quantile regression and restricted mean treatment effects. Quantile regression estimates the multiplicative or additive effect of covariates on a quantile of interest, e.g., median survival, while a restricted mean modifies the traditional summary measure of mean differences in a manner that is estimable under censoring. For example, the 10-year restricted mean is the average survival over 10 years. Both of these methods help translate statistical results by directly incorporating survival time into the interpretation of the covariate effect.

Despite potential advantages in interpretation, censored quantile regression and restricted mean treatment effects face significant computational difficulties. In particular, both approaches either implicitly or explicitly estimate a conditional survival function as a nuisance parameter. As a result, the estimator of the conditional survival function may affect estimation of the parameters of interest, e.g., a restricted mean treatment effect. As such, this thesis focuses on two related objectives:

1. Improving conditional survival function estimation.
2. The impact of conditional survival function estimation on censored quantile regression and restricted mean treatment effects.

The idea is that improving estimation of the conditional survival function will, in turn, improve estimation of censored quantile regression and restricted mean treatment effects.

Most current censored quantile regression approaches use a recursive estimation procedure that implicitly estimates the conditional survival function, while restricted mean treatment effects are estimated with a Cox proportional hazards model. Both of these approaches are broadly classified as semi-parametric estimators of the conditional survival function. Despite some flexibility, semi-parametric estimators can perform poorly when their assumptions are badly violated.

Non-parametric estimation is one approach to alleviating problems associated with violated parametric or semi-parametric assumptions. In particular, we propose bagged survival trees for estimating censored quantile regression in Chapter 2. However, a lack of efficiency poses a significant problem for bagged survival trees and, in general, non-parametric estimators. This means that bagged survival trees require major violations of assumptions to ensure better performance than parametric and semi-parametric estimators. This creates a practical issue as it is difficult to determine the underlying validity, or violations, of parametric and semi-parametric assumptions.

Chapter 3 proposes stacked survival models to effectively estimate the conditional survival function in a wide variety of situations. By minimizing predicted error, stacked survival models estimate an optimally weighted combination of models that can span parametric, semi-parametric, and non-parametric survival models. As such, stacking can exploit the low variance of approximately correct parametric models, while maintaining the robustness of non-parametric models. Chapter 3 shows that stacking parametric, semi-parametric, and non-parametric models always performs better than the model selected through cross-validation.

Chapter 4 applies stacked survival models to the estimation of restricted mean treatment effects. This represents a promising situation to evaluate stacked survival models within an inference setting due to the direct relationship between the restricted mean and conditional survival function. Chapter 4 shows that stacked survival models achieve as good, or better, mean-squared error as current methods for estimating restricted mean treatment effects.

1.1 Notational Notes

There are some important notes regarding notation throughout the thesis. Most importantly, there are, despite our best attempts, slight notational differences between chapters. The meaning of τ possesses the most significant notational difference. In particular, Chapter 2 defines τ as a specific quantile, while Chapters 3 and 4 define τ as a specific time point. In addition, random variables and observed variables are distinguished by capital and lower case letters, respectively. However, throughout the thesis, we suppress the random variable notation when conditioning on covariates. For example, we write $E\{T|\boldsymbol{x}\}$ rather than $E\{T|\boldsymbol{X} = \boldsymbol{x}\}$. To try to alleviate confusion, each chapter reintroduces the necessary notation.

Chapter 2

Censored Quantile Regression

Censored quantile regression is a useful alternative to the Cox model that has recently received considerable attention. Uncensored quantile regression methods have been extensively studied within the econometrics literature since the seminal work of Koenker and Bassett [1978]; see Koenker [2005] for a comprehensive introduction. Quantile regression models the relationship between the event time and the covariates using the quantile function:

$$Q_T(\tau|\boldsymbol{x}) \quad = \quad \boldsymbol{x}\boldsymbol{\beta}(\tau), \tag{2.1}$$

where $\tau \, \epsilon \, (0,1)$ is the quantile of interest, $\boldsymbol{\beta}(\tau)$ is the vector of τ^{th} quantile effects, and T is the event time random variable. This enables researchers to model not only measures of central tendency, such as the median, but also other aspects of the conditional distribution such as the tails. An advantage of quantile regression is its invariance under monotonically increasing transformations, i.e., $Q_{h(T)}(\tau|\boldsymbol{x}) = h(Q_T(\tau|\boldsymbol{x}))$ where h is a monotonically increasing function [Koenker, 2005].

Censored quantile regression was first investigated in the econometrics literature for fixed censoring, i.e., all the censoring times are known regardless of whether the event occurs, see Powell [1986]. This assumption is almost never met within applied health research. Ying et al. [1995] and Yang [1999] both proposed median estimators (presumably generalizable to any quantile) that assumed unconditional independence between event and censoring times (i.e., $T \perp C$, where C is the censoring time random variable and $\perp$ denotes statistical independence).

Portnoy [2003] adopted the more relaxed assumption of conditionally independent censoring (i.e., $T \perp C \mid \boldsymbol{x}$). He proposed a novel method of recursively estimating a series of quantile regression functions defined on a grid along $(0, \tau_o)$, where τ_o is the quantile

of interest. However, this recursive estimation relies on the assumption that the conditional quantile function is linear for all $\tau \epsilon (0, \tau_o)$. Wang and Wang [2009] refer to this assumption as the "global linearity assumption", and observed that noticeable bias can occur when this assumption is violated.

Peng and Huang [2008] proposed an estimator, referred to in this chapter as 'PH', that uses a martingale estimating equation which exploits the relationship between the quantiles and cumulative hazard function. Similar to Portnoy's approach, the PH estimator assumes both conditionally independent censoring and linearity in all quantiles by estimating a series of regression quantiles along a grid. Although it has not been investigated in the literature, it is anticipated that the performance of the PH estimator is influenced when the global linearity assumption is violated, as reflected in simulation results presented later in this Chapter.

Wang and Wang [2009] proposed a locally weighted censored quantile regression approach that adopts the redistribution-of-mass idea of Efron [1967] and employs a local reweighting scheme. Its validity only requires the conditional independence of the survival time and the censoring variable given the covariates, and linearity at the quantile of interest. However, their locally weighted estimator suffers from two notable drawbacks in real data analysis. First, kernel smoothing becomes impractical with only a moderate number of covariates ($p > 2$), i.e., the curse of dimensionality. Second, kernel theory was developed for continuous covariates, so the presence of categorical variables causes the method to become ill-defined.

We propose a procedure that uses survival trees with Kaplan-Meier estimates [Kaplan and Meier, 1958] as the basis for the locally weighted estimator. By avoiding the use of a kernel, the approach is more flexible in handling moderate to high dimensions and discrete covariates while avoiding the global linearity assumption. We establish that the procedure leads to consistent estimation of the quantile regression coefficients.

The next section introduces the estimator, certain important aspects of survival trees, and censored quantile regression. Section 2.2 shows the consistency and discusses the difficulties in showing asymptotic normality of the estimator. Section 2.3 presents a series of simulations analyzing the finite sample performance of the proposed estimator, which is illustrated in Section 2.4 with an analysis of data on primary biliary cirrhosis. Finally, a summary is presented in Section 2.5.

2.1 Proposed Estimator

We start by making some distinctions and introducing some notation: capitalized letters with no subscripts indicate a random variable, while lower case letters with subscripts indicate an observed variable. The conditional distribution of the event time is $F_T(t|\boldsymbol{x}) = 1 - S(t|\boldsymbol{x}) = P(T \leq t|\boldsymbol{x})$, and the conditional distribution of the censoring time is $F_C(t|\boldsymbol{x}) = 1 - G(t|\boldsymbol{x}) = P(C \leq t|\boldsymbol{x})$. The covariates measured at the beginning of the study are denoted by the vector $\boldsymbol{x}_i$, and the event and censoring random variables are assumed to be conditionally independent (i.e., $T \perp C|\boldsymbol{x}$). Hence a sample of right censored survival data of size n consists of triplets $\{y_i, \delta_i, \boldsymbol{x}_i\}$, $i = 1, ..., n$, where $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i < c_i)$.

2.1.1 Estimation

When there is no censoring (i.e., $y_i = t_i$ for all $i = 1, ..., n$), the τ^{th} conditional quantile $\beta(\tau)$ can be estimated by minimizing the following quantile objective function [Koenker, 2005]

$$S_n(\boldsymbol{\beta}(\tau)) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i\boldsymbol{\beta}(\tau)), \tag{2.2}$$

where $\rho_\tau(z) = z \cdot \{\tau - I(z < 0)\}$ is the quantile loss function and $I(u)$ is the indicator function (i.e., $I(A)$ is 1 if the event A is true, and 0 otherwise). When the survival time is subject to random right censoring, Wang and Wang [2009] proposed estimating $\boldsymbol{\beta}(\tau)$ by minimizing the weighted quantile objective function

$$\begin{aligned} R_n(\boldsymbol{\beta}(\tau), F_T) \quad = \quad & \frac{1}{n}\sum_{i=1}^{n} \{w_i(F_T)\rho_\tau(y_i - \boldsymbol{x}_i\boldsymbol{\beta}(\tau)) + \\ & (1 - w_i(F_T))\rho_\tau(y^{+\infty} - \boldsymbol{x}_i\boldsymbol{\beta}(\tau))\}, \end{aligned} \tag{2.3}$$

where $y^{+\infty}$ represents a number large enough to be effectively infinity, and

$$w_i(F_T) = \begin{cases} 1 & \text{if } \delta_i = 1 \text{ or } F_T(c_i|\boldsymbol{x}_i) > \tau \\ \dfrac{\tau - F_T(c_i|\boldsymbol{x}_i)}{1 - F_T(c_i|\boldsymbol{x}_i)} & \text{if } \delta_i = 0 \text{ and } F_T(c_i|\boldsymbol{x}_i) < \tau \end{cases}$$

with $F_T(t|\boldsymbol{x})$ being the conditional distribution function of T given $\boldsymbol{x}$.

The motivation for the weighted quantile objective function in (2.3) is that the contribution of each point for the estimation of $\boldsymbol{\beta}(\tau)$ depends only on the sign of the residual, where the residual is defined as $t_i - \boldsymbol{x}_i\boldsymbol{\beta}(\tau)$. For the uncensored observations, the sign of

the residual is directly observed for a given $\boldsymbol{\beta}(\tau)$. For the censored observations, there are two possibilities.

1. If $c_i > \boldsymbol{x}_i\boldsymbol{\beta}(\tau)$, then $t_i - \boldsymbol{x}_i\boldsymbol{\beta}(\tau) > 0$. That is, if the censored time is larger than the predicted quantile of the survival time, then the sign of the residual is known since $t_i > c_i$.

2. If $c_i < \boldsymbol{x}_i\boldsymbol{\beta}(\tau)$, then the sign of the residual is not determined. In this case, given $(\boldsymbol{x}_i, c_i)$, the conditional probability of obtaining a negative residual is

$$\begin{aligned} E[I(T - \boldsymbol{x}_i\boldsymbol{\beta}(\tau) < 0)|T > c_i] &= P(T < \boldsymbol{x}_i\boldsymbol{\beta}(\tau)|T > c_i) \\ &= \frac{P(c_i < T < \boldsymbol{x}_i\boldsymbol{\beta}(\tau))}{P(T > c_i)} \\ &= \frac{\tau - F_T(c_i|\boldsymbol{x}_i)}{1 - F_T(c_i|\boldsymbol{x}_i)}. \end{aligned} \tag{2.4}$$

In this ambiguous case, adopting the redistribution-of-mass idea of Efron [1967], we assign weight $w_i(F_T)$ to the observation at $(\boldsymbol{x}_i, c_i)$ and redistribute the complimentary weight $1 - w_i(F_T)$ to $(\boldsymbol{x}_i, y^{+\infty})$ without altering the quantile.

To estimate the weights, it is essential to estimate the conditional distribution of the survival time. In Section 2.1.2, we propose an approach for estimating the weights that enjoy some appealing properties. It is worthwhile to note that the weighting scheme reduces to ordinary quantile regression in the presence of no censoring or when no censored observations are reweighted (i.e., extremely late censoring relative to the quantile of interest). Also, the censoring distribution can have a direct impact beyond the marginal level of censoring. Depending on the timing, e.g., early vs. late censoring, more or less of the censored observations would be re-weighted. As an example, across a range from early to late censoring, with the same marginal level of 35% censoring, the proportion of censored observations that were re-weighted ranged from 20% to 87% using Portnoy's approach (more details are presented in Section 2.3). Furthermore, the subset of censored observations that are re-weighted would often differ between methods in addition to differences in ascribed weight (e.g., due to differences in the estimates of $\hat{F}(t|\boldsymbol{x})$).

2.1.2 Survival Trees

The proposed estimator uses survival trees, or recursive partitioning, as described by LeBlanc and Crowley [1993] and Butler et al. [1989] to estimate the weights of censored observations described by (2.4) for the estimating equation (2.3). The goal is not to fully describe recursive partitioning or survival trees in detail so some familiarity is assumed. The interested readers are referred to Breiman et al. [1984] for a comprehensive treatment

of recursive partitioning and Bou-Hamad et al. [2011] for a review of recent survival tree literature. Briefly, there is a need to introduce two concepts: splitting and stopping rules.

Splitting rules determine where and how to split a node. The trees used in this paper only consider splits on one variable at a time, resulting in binary trees. We use a splitting criteria that is the maximum of four $G^{\rho,\gamma}$ statistics:

$$G^{\rho,\gamma} = \frac{M_1 + M_0}{M_1 M_0} \sum_{t \epsilon \mathcal{F}} \frac{n_{1t} n_{0t}}{n_{1t} + n_{0t}} \hat{S}(t-)^{\rho} [1 - \hat{S}(t-)]^{\gamma} [\hat{\lambda}_1(t) - \hat{\lambda}_0(t)], \tag{2.5}$$

where M_j is the number of subjects initially at risk in group j, $\mathcal{F}$ is the set of unique failure times, n_{jt} is the number of subjects at risk in group j at time t and $\hat{\lambda}_j(t)$ is the estimated hazard of group j at time t [Rudser et al., 2012]. The four $G^{\rho,\gamma}$ statistics used are: $(\rho, \gamma) = \{(0,0), (1,0), (0,1), (1,1)\}$. Note that $(0,0)$ and $(1,0)$ correspond to the log-rank and weighted Wilcoxon form of the logrank test, respectively (the other two do not have common names). This cocktail of $G^{\rho,\gamma}$ statistics is used to increase the power to detect a variety of differences between survival functions [Lee, 1996]. While this collection of $G^{\rho,\gamma}$ statistics is designed to find several different types of differences in survival functions, one may choose fewer or only one $G^{\rho,\gamma}$ statistic (e.g., only the log rank statistic).

Stopping rules are used to indicate when to stop splitting at a particular node. These are used to prevent any particular node from not having enough information (e.g., small sample size, lack of events, etc.) to effectively estimate the probabilities of interest. This naturally leads to two 'tuning parameters' that need to be specified:

1. "Minimum at Risk": Each node is required to have a minimum number of subjects at risk for an event.

2. "Minimum Events": Each node is required to have a minimum number of events.

For censored quantile regression, we are interested in the conditional probabilities required for properly weighting censored observations [see equation (2.4)]. By letting the minimum events depend upon the number at risk within a particular node and the quantile being estimated, we can ensure that each terminal node (i.e., a node that did not split further) has enough information to effectively estimate the quantile of interest using a Kaplan-Meier estimator. While the Kaplan-Meier estimator is used here, it can be replaced by any cumulative distribution estimator for censored data.

Sensitivity to small changes in the data is a common criticism of trees. Breiman [1996a] suggested "bagging" as one effective way to alleviate this problem. Bagging requires taking a prespecified number of bootstrapped data sets that are sampled with replacement, then uses the average of the estimand over the bootstrapped datasets as the 'bagged' estimate. In terms of trees, this means bootstrapping the data set a number of times, say $bagN$, and obtaining $\tilde{F}_{bag_b}(t|\boldsymbol{x})$ for the b^{th} bootstrapped data set. Then the final conditional distribution estimate for subject i is defined as

$$\hat{F}(t|\boldsymbol{x}_i) \quad = \quad \frac{1}{bagN} \sum_{b=1}^{bagN} \tilde{F}_{bag_b}(t|\boldsymbol{x}_i). \tag{2.6}$$

This should have a stabilizing effect on the tree-based estimate of $F(t|\boldsymbol{x}_i)$.

2.1.3 Implementation

To implement the proposed method, a researcher needs to specify three aspects of the survival trees: the splitting and stopping rules, and how many bags to use. After using (bagged) survival trees to determine the weights, reweighted censored observations are split with weight $w_i(F_T)$ at $(y_i, \boldsymbol{x}_i)$ and weight $1 - w_i(F_T)$ at $(y_i^*, \boldsymbol{x}_i)$, where y_i^* is a large enough number to ensure a positive residual (e.g., $1000 \times (\max_i\{y_i\} + 1)$). After splitting the appropriate observations between y_i and y_i^*, the estimating equation (2.3) can be solved in R [R Development Core Team, 2013] using the function rq() from the 'quantreg' package [Koenker, 2011] with user-defined weights.

2.2 Asymptotics

The proposed tree based censored quantile regression estimator is consistent given certain regularity conditions. The following theorem summarizes this property.

Theorem 2.1. *Assume that* $\{y_i, \delta_i, \boldsymbol{x}_i\}$, $i = 1, ..., n$, *are independent and identically distributed with* T *independent of* C *conditional on* $\boldsymbol{x}$, *and that assumptions (A1) through (A5) in Appendix A hold. Let* $\hat{\boldsymbol{\beta}}(\tau)$ *be the minimizer of (2.3) with* $\hat{F}_T(\cdot|\boldsymbol{x})$ *computed using a survival tree. Then*

$$\hat{\boldsymbol{\beta}}(\tau) \quad \rightarrow \quad \boldsymbol{\beta}(\tau), \tag{2.7}$$

in probability, as $n \rightarrow \infty$.

The proof is in Appendix A and relies on the theory of Chen et al. [2003] for nonsmooth estimating equations with an infinite-dimensional nuisance parameter that requires the

survival tree estimate to be uniformly consistent for the conditional survival function. This is shown using recursive partitioning theory developed by Gordon and Olshen [1984] and Butler et al. [1989] that require the size of every terminal node to become arbitrarily small in every covariate. This suggests that the tree size, i.e., number of terminal nodes, needs to grow at a slower rate than the sample size within each terminal node with both tending to infinity or, practically, that the minimum number of events increases with the sample size.

Showing asymptotic normality is not straightforward. The sufficient conditions outlined by Chen et al. [2003] for asymptotic normality require substantial additions to the recursive partitioning asymptotic literature for censored data: a more accurate limit on the rate of convergence of survival trees, and a linear representation of survival trees into mean 0 and finite variance random variables. To our knowledge, there is little to no survival tree literature on these specific topics. Most recursive partitioning asymptotic results focus on showing the consistency of estimated summary measures of conditional distribution functions while avoiding the discussion on rates of convergence and linear representations. We do not pursue these topics further.

Inference is an important matter in statistics, which helps motivate showing the asymptotic distribution of an estimator. With any conditional quantile regression method the covariance matrix of $\hat{\boldsymbol{\beta}}(\tau)$ depends upon an unknown conditional density [Koenker, 2005]. The unknown density function makes accessible variance solutions extremely difficult to obtain. Portnoy [2003] proposed to sample the observed triplets $\{y_i, \delta_i, \boldsymbol{x}_i\}$ with replacement (i.e., non-parametric bootstrap). After drawing a sufficient number of bootstraps, confidence intervals can be constructed based on sample quantiles or normal approximations of the bootstrap distribution. The tree-based method presented here uses the 2.5^{th} and 97.5^{th} sample quantiles of the bootstrap distribution to construct an approximate 95% confidence interval.

2.3 Simulations

We assess the finite sample performance of the tree-based estimator (TW) compared to the Portnoy and Peng, Huang (PH) estimators through two simulation scenarios. When analyzing the effectiveness of tree-based weights, only bagged survival trees are included with $bagN = 10$. The minimum number at risk is 60 and the minimum number of events is $N_{TN} \cdot \tau$, where τ is the quantile being estimated and N_{TN} is the number of observations within a node. All simulations were performed using R version 2.12.2 with the `quantreg` package used to fit the Portnoy and PH estimators. Approaches are compared based on operating characteristics of bias, mean squared error (MSE), coverage of 95% confidence

intervals (Cov.), average confidence interval lengths (ACL), and power for a variety of simulations scenarios at the median ($\tau = 0.5$) and $\tau = 0.25$ quantile. The Wang and Wang estimator was left out due to the computational difficulties associated with moderate to high dimensional kernel estimation.

The simulation scenarios are categorized by two sets of covariate distributions (i.e., number of covariates) with varying levels of non-linearity (i.e., specification of the error distribution). The scenarios are formed from subsets of

$$\begin{aligned}
\boldsymbol{x}_i\beta &= 2 + x_{i,1} - 2 \cdot x_{i,2} + x_{i,3}, \\
x_{i,1} &\sim Unif(-2, 2), \\
x_{i,2} &\sim N(0, 1), \\
x_{i,3} &\sim P(X_3 = m) = \frac{1}{6}, \text{ for } m = 1, 4, \text{ and } P(X_3 = m) = \frac{1}{3}, \text{ for } m = 2, 3.
\end{aligned}$$

The first and second simulation scenarios consist of, respectively, $\Omega_1 = \{x_{i,1}, x_{i,2}\}$ and $\Omega_2 = \{x_{i,1}, x_{i,2}, x_{i,3}\}$, where Ω_k is the set of covariates for simulation k. The error structures are defined as $E_l \times (N(0,1) - \Phi^{-1}(\tau))$, where E_l are the equations that induce non-linearity, τ is the quantile of interest and Φ^{-1} is the inverse c.d.f. of the standard normal.

The error distribution with linearity in all quantiles is $E_1 = 3$. Mild non-linearity and severe non-linearity correspond to, respectively, $E_2 = \frac{3}{2} + 2 \cdot (x_{i,1} - \frac{1}{2})^2$ and $E_3 = \frac{3}{2} + 6 \cdot (x_{i,1} - \frac{1}{2})^2$. The censoring distributions are chosen depending upon the error structure with linear, mild non-linearity, and severe non-linearity represented by, respectively $Unif(-3, a(\Omega_k, E_l))$, $(\frac{3}{10} + (x_{i,1} - \frac{1}{2})^2) \times Unif(-3.75, a(\Omega_k, E_l))$, and $(\frac{3}{10} + (x_{i,1} - \frac{1}{2})^2) \times Unif(-3, a(\Omega_k, E_l))$, where $a(\Omega_k, E_l)$ is chosen to ensure 25% censoring for the median scenarios and 45% censoring for $\tau = 0.25$ scenarios. These censoring distributions lead to fairly even censoring across time and $x_{i,1}$. Each simulation scenario and error structure combination is evaluated over 2500 simulation iterations, and each combination has a sample size of 400 with 300 bootstrap replicates for confidence intervals.

The first error structure, E_1, possesses linearity in all quantiles for all variables. Due to their implicit assumption of linearity in all quantiles, it is expected that the Portnoy and PH estimators will perform better than the tree-based approach. The second and third error structures impose non-linearity in all quantiles for $x_{i,1}$ except the quantile of interest. These scenarios, particularly the severe non-linearity scenario, are likely more favorable for the tree-based approach compared to Portnoy and PH. Note that $x_{i,1}$ is the only covariate that possesses non-linearity in all quantiles except the quantile of interest.

The potential advantage of the proposed tree-based estimator is improved performance in multivariate scenarios with non-linearity in some quantile. As such, we have two

primary interests: whether the tree-based estimators are competitive in scenarios with linearity through all quantiles and, second, the degree to which the tree-based estimators outperform the Portnoy and PH estimators in the presence of non-linearity. Tables 2.1 and 2.2 illustrate that the tree-based estimator accomplishes the former at some cost of bias for $\tau = 0.25$, but are similar to the Portnoy and PH estimators for the median (i.e., the 'No Non-Linearity' columns in Tables 2.1 and 2.2). For the latter question (i.e., the 'Mild Non-Linearity' and 'Severe Non-Linearity' columns), the tree-based estimator - for severe non-linearity - possesses less bias and MSE (particularly for the covariates that are linear in all quantiles) when estimating the median and $\tau = 0.25$. The advantages of the tree-based estimator are attenuated for the mild non-linearity scenarios.

The advantage of the tree-based estimator appears to depend upon the level of censoring. In particular, the tree-based estimator shows less improvement for bias when the percent of censoring increases with respect to the quantile of interest. This may be due to our strict stopping rule that forces the number of events to be proportional to the quantile of interest. This stopping rule is increasingly restrictive when the marginal censoring rate is closer to the quantile of interest, but is necessary to guarantee coherent estimation of the weights, i.e., for the Kaplan-Meier estimate to reach the quantile of interest. This could be relaxed by making distributional assumptions (e.g., assuming a log-Normal distribution). However, the performance would then depend on the extent to which the data follow the distributional assumptions.

Additionally, the performance of all censored quantile regression estimators can vary wildly depending on the location of the censored observations even while keeping the overall marginal level of censoring constant. As an illustration, we designed a small univariate simulation study that was similar to the above. The bias was unaffected when all the covariates possessed linearity in all quantiles, but - in the presence of non-linearity - we observed that the bias ranged from 0.17 to 0.26 for 'late' to 'early' censoring, respectively. Due to the large variations in performance and percent of reweighted observations, it is important for simulation studies to specify the censoring used when evaluating censored quantile regression methods, and ensure resulting patterns of censoring are realistic. Explicitly stating the censoring distributions and the percent of observations reweighted (Table 2.3) when presenting simulation results would be helpful as well.

TABLE 2.1: *First simulation scenario:* $N = 400$, $N_{SIM} = 2500$, *censoring is* 45% *and* 25% *for* $\tau = 0.25$ *and* $\tau = 0.5$, *respectively,* $\beta_0 = 2$, $\beta_1 = 1$, $\beta_2 = -2$, 300 *bootstrap replicates,* 95% *nominal coverage with ACL representing the average CI width.*

			No Non-Linearity					Mild Non-Linearity					Severe Non-Linearity				
Quantile	Variable	Method	Bias	MSE	Cov.	ACL	Power	Bias	MSE	Cov.	ACL	Power	Bias	MSE	Cov.	ACL	Power
0.25	Variable 1	Portnoy	0.00	0.04	0.97	0.83	1.00	0.07	0.15	0.96	1.51	0.81	0.17	0.70	0.95	3.26	0.28
	$\beta_1 = 1$	PH	0.01	0.04	0.97	0.83	1.00	0.00	0.14	0.96	1.50	0.76	-0.04	0.67	0.96	3.24	0.21
		TW	-0.06	0.04	0.96	0.81	1.00	-0.03	0.13	0.96	1.47	0.76	0.01	0.59	0.96	3.09	0.25
	Variable 2	Portnoy	0.01	0.06	0.96	0.99	1.00	-0.07	0.09	0.96	1.20	1.00	-0.20	0.28	0.96	2.10	1.00
	$\beta_2 = -2$	PH	-0.01	0.06	0.96	0.99	1.00	-0.09	0.09	0.96	1.20	1.00	-0.26	0.31	0.95	2.13	1.00
		TW	0.10	0.06	0.95	0.97	1.00	0.03	0.09	0.97	1.21	1.00	0.06	0.21	0.97	2.02	1.00
0.5	Variable 1	Portnoy	0.01	0.03	0.96	0.71	1.00	0.05	0.11	0.96	1.32	0.89	0.10	0.52	0.95	2.85	0.34
	$\beta_1 = 1$	PH	0.00	0.03	0.96	0.72	1.00	-0.01	0.11	0.96	1.33	0.84	-0.08	0.54	0.95	2.90	0.23
		TW	-0.01	0.03	0.97	0.71	1.00	0.02	0.11	0.96	1.33	0.86	0.04	0.52	0.96	2.90	0.31
	Variable 2	Portnoy	0.00	0.04	0.96	0.82	1.00	-0.05	0.05	0.97	0.95	1.00	-0.13	0.15	0.95	1.56	1.00
	$\beta_2 = -2$	PH	0.00	0.04	0.97	0.84	1.00	-0.06	0.06	0.97	0.97	1.00	-0.15	0.16	0.95	1.62	1.00
		TW	0.02	0.04	0.97	0.84	1.00	-0.02	0.06	0.97	0.98	1.00	-0.03	0.13	0.97	1.60	1.00

TABLE 2.2: *Second simulation scenario:* $N = 400$, $N_{SIM} = 2500$, *censoring is* 45% *and* 25% *for* $\tau = 0.25$ *and* $\tau = 0.5$, *respectively,* $\beta_0 = 2$, $\beta_1 = 1$, $\beta_2 = -2$, $\beta_3 = 1$, 300 *bootstrap replicates,* 95% *nominal coverage with ACL representing the average CI width.*

			No Non-Linearity					Mild Non-Linearity					Severe Non-Linearity				
Quantile	Variable	Method	Bias	MSE	Cov.	ACL	Power	Bias	MSE	Cov.	ACL	Power	Bias	MSE	Cov.	ACL	Power
0.25	Variable 1	Portnoy	-0.01	0.04	0.97	0.86	1.00	0.06	0.15	0.95	1.54	0.78	0.16	0.75	0.95	3.37	0.26
	$\beta_1 = 1$	PH	0.00	0.04	0.96	0.86	1.00	-0.01	0.15	0.95	1.53	0.73	-0.04	0.73	0.95	3.35	0.20
		TW	-0.06	0.04	0.97	0.86	1.00	-0.05	0.14	0.96	1.51	0.72	-0.01	0.68	0.95	3.28	0.22
	Variable 2	Portnoy	-0.01	0.06	0.97	1.01	1.00	-0.06	0.09	0.97	1.26	1.00	-0.19	0.29	0.97	2.27	0.99
	$\beta_2 = -2$	PH	-0.02	0.06	0.97	1.01	1.00	-0.08	0.09	0.97	1.26	1.00	-0.24	0.33	0.96	2.29	0.99
		TW	0.06	0.06	0.97	1.02	1.00	0.02	0.09	0.97	1.29	1.00	0.02	0.25	0.97	2.21	0.98
	Variable 3	Portnoy	0.00	0.06	0.97	1.04	0.98	0.03	0.09	0.97	1.27	0.91	0.10	0.27	0.96	2.24	0.53
	$\beta_3 = 1$	PH	0.01	0.06	0.97	1.03	0.98	0.04	0.09	0.96	1.27	0.92	0.12	0.28	0.96	2.26	0.54
		TW	-0.09	0.07	0.96	0.99	0.97	-0.09	0.09	0.97	1.23	0.88	-0.11	0.21	0.97	2.01	0.44
0.5	Variable 1	Portnoy	-0.01	0.03	0.96	0.73	1.00	0.05	0.11	0.95	1.33	0.88	0.11	0.56	0.95	2.93	0.32
	$\beta_1 = 1$	PH	-0.01	0.03	0.96	0.74	1.00	-0.02	0.11	0.96	1.34	0.82	-0.07	0.56	0.95	2.96	0.23
		TW	-0.01	0.03	0.97	0.74	1.00	0.01	0.11	0.96	1.34	0.86	0.03	0.56	0.95	2.98	0.28
	Variable 2	Portnoy	-0.01	0.05	0.96	0.85	1.00	-0.04	0.06	0.95	0.98	1.00	-0.12	0.17	0.95	1.66	1.00
	$\beta_2 = -2$	PH	-0.01	0.05	0.95	0.86	1.00	-0.05	0.06	0.95	1.00	1.00	-0.15	0.19	0.95	1.72	1.00
		TW	0.00	0.05	0.96	0.86	1.00	-0.02	0.06	0.96	1.02	1.00	-0.04	0.16	0.96	1.71	1.00
	Variable 3	Portnoy	0.00	0.05	0.97	0.88	0.99	0.01	0.06	0.97	1.00	0.98	0.05	0.15	0.97	1.68	0.72
	$\beta_3 = 1$	PH	0.00	0.05	0.97	0.89	1.00	0.01	0.06	0.97	1.03	0.98	0.06	0.16	0.97	1.74	0.70
		TW	0.00	0.05	0.97	0.89	1.00	0.01	0.06	0.97	1.03	0.98	0.02	0.15	0.97	1.71	0.69

TABLE 2.3: *Percent of total observations reweighted by the simulation scenario (i.e., number of covariates) and the degree of non-linearity (NL). The marginal censoring for all simulation scenarios was* 45% *and* 25% *for* $\tau = 0.25$ *and* $\tau = 0.5$*, respectively.*

		Scenario 1			Scenario 2		
Quantile	Method	No NL	Mild NL	Severe NL	No NL	Mild NL	Severe NL
0.25	Portnoy	26.8%	30.1%	29.1%	21.3%	28.9%	31.1%
	TW	31.2%	32.8%	29.5%	32.3%	33.3%	30.8%
0.5	Portnoy	18.3%	20.2%	16.9%	17.5%	20.7%	19.2%
	TW	19.5%	21.0%	17.0%	21.2%	21.9%	19.2%

2.4 Analysis of Primary Biliary Cirrhosis Dataset

As an illustration, we apply the proposed method to the well-recognized primary biliary cirrhosis (PBC) data set described by Fleming and Harrington [1991] from a clinical trial investigating the effect of the drug D-penicillamine conducted at the Mayo Clinic in Rochester, Minnesota. The data set is readily available in the R package `survival` as the `pbc` object (Therneau, 2013), and is widely considered a benchmark dataset for survival analysis. We are interested in evaluating the association of the treatment, age, bilirubin and prothrombin time with the log time till death or transplant. Since bilirubin and prothrombin time appear to violate the global linearity assumption (see Figure 2.1), this is a scenario suited for the proposed tree-based estimator.

Considering only complete cases, this results in 312 patients with approximately 53.8% censoring. Portnoy's approach is compared to the proposed estimator with 10 bags. The minimum number at risk is set to 60, and the minimum number of events is $N_{TN} \cdot \tau$, where τ is the quantile being estimated and N_{TN} is the number of observations within a node. Both approaches use bootstrap re-sampling for confidence intervals: the 2.5^{th} and 97.5^{th} quantiles were used to construct the 95% confidence intervals using 1000 bootstraps for both estimators.

Figure 2.2 displays the estimated covariate effects on quantiles from $\tau = 0.05$ to $\tau = 0.50$. Of the four variables of interest, the treatment appears to have no effect on time till transplant or death, while bilirubin appears to have a substantial constant effect on time till transplant or death. Longer prothrombin times appear to have a significant negative effect on survival time that attenuates for quantiles closer to the median. The estimated effects of bilirubin and age are different between the tree and Portnoy approaches. In particular, the tree-based weights have estimates closer to the null relative to Portnoy's estimator. Take the 25th quantile as an example, the Portnoy estimator displays about 30% and 18% larger absolute effect estimates [for $\log(T)$] compared to the tree-based estimator for the effect of age and bilirubin, respectively. This direction and relative

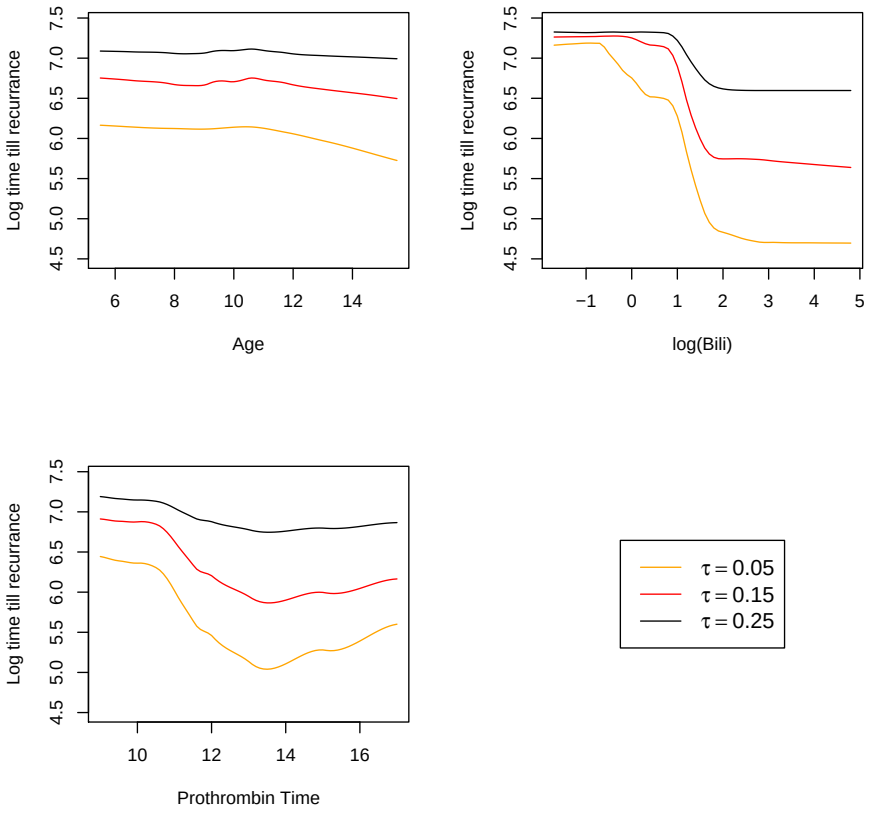


FIGURE 2.1: The marginal quantile relationships for age, log(bilirubin), and prothrombin time. Note the non-linearity in low quantiles for both log(bilirubin) and prothrombin time. The quantile functions were estimated using the bagged survival trees ($bagN = 10$).

ordering of the two estimates are consistent with the anti-conservative bias for Portnoy's estimator in the presence of non-linearity that was observed in the simulation results of Section 2.3. Additionally, the tree-based estimator generally has narrower confidence intervals around $\tau = 0.25$ compared to Portnoy, which is consistent with the simulation results. The tree-based estimator has wider confidence intervals towards the median. However, the censoring rate is above 50% for the PBC data set, hence neither method can accurately estimate the median or higher quantiles.

In the analysis, we focus on the 25th quantile which corresponds to the patients with relatively short survival times. The estimated 25^{th} conditional quantile function using the tree based estimator is:

$$Q_{\log(T)}(0.25|\boldsymbol{x}) \quad = \quad 12.43 - 0.02[\text{Trt}] - 0.11[\frac{\text{age}}{5}] - 0.41[\log_2(\text{bili})] - 0.35[\text{pro. time}],$$

whose coefficients are exponentiated to obtain an interpretation on the original time scale. For example, a two-fold difference in bilirubin is associated with an average -0.41

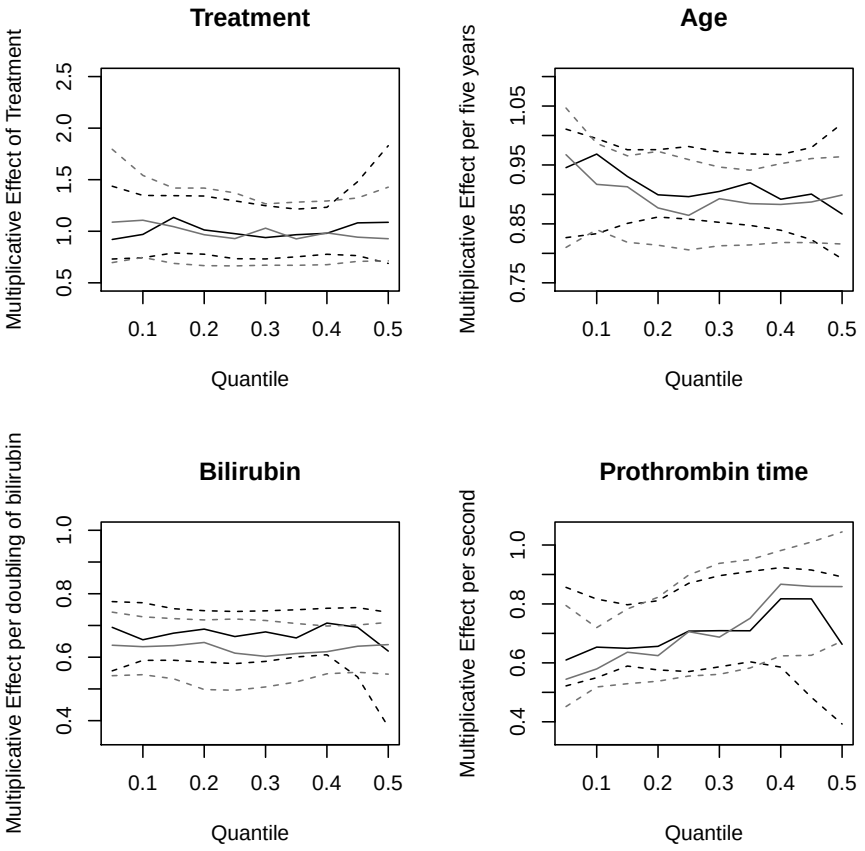


FIGURE 2.2: Estimated multiplicative effects on time to event for 0.05 to 0.5 quantiles (solid lines). 95% confidence intervals (dashed lines) are formed by taking the 2.5^{th} and 97.5^{th} sample quantiles of 1000 bootstrapped samples. The tree-based estimator and Portnoy's estimator are the black and gray lines, respectively.

shorter log time till transplant/death for the 25^{th} quantile. On the original time scale, this corresponds to 33.5% shorter survival time for the 25^{th} quantile on average while adjusting for treatment, baseline age and prothrombin time. On the other hand, a difference of five years of age implies, on average, 10.4% shorter survival time for the 25^{th} quantile while adjusting for treatment, baseline bilirubin and prothrombin time. The other covariates are interpreted in a similar fashion.

2.5 Summary

Motivated in part by the difficulty encountered by the estimator of Wang and Wang [2009] with moderately high dimensional data, we propose a new tree-based weighted censored quantile regression estimator. Under mild conditions, the new estimator is consistent. The simulation study demonstrated that if any variable possesses non-linearity then the Portnoy and PH estimators can suffer from bias and loss of precision in the

estimated coefficients for *all* covariates, not just the covariates with non-linearity. Additionally, the proposed tree-based estimator can improve the bias and MSE in the presence of non-linearity for multivariate scenarios. Interestingly, the largest improvements were for covariates that possessed linearity through all quantiles when adjusting for a covariate with non-linearity. A limitation is, due to strict splitting rules that require the quantile of interest to be defined in each node, the proposed tree-based estimator is more sensitive to a high censoring rate relative to the quantile of interest compared to the Portnoy and PH estimators.

We found that the performance of the estimators depended heavily on the censoring distribution. In particular, in the presence of non-linearity, the Portnoy estimator provides a biased estimate that depends on the location of the censoring distribution. As such, we recommend future investigations of censored quantile regression to explicitly state the censoring distribution used and where the censoring is occurring and to report the percent of observations reweighted, for approaches based on the weighted estimating equation of the form (2.3). The extent of the censoring distribution's impact is less clear for other approaches [e.g., the martingale approach of Peng and Huang [2008]]. Further investigation and benchmarking of relative performance of this issue would be an interesting future research topic.

Compared with the local Kaplan-Meier estimator based weights, i.e., Wang and Wang [2009], the tree-based weights have appealing properties that work better with moderately high dimensional covariates while avoiding the linearity assumption of Portnoy [2003] and Peng and Huang [2008]. An alternative approach to estimating the weights is using flexible spline methods. For example, the polynomial splines developed by Kooperberg et al. [1995] can flexibly estimate the conditional hazard function (the `hare` function in R). This approach could be extended to estimate the conditional survival function used for censored quantile regression. This is an interesting direction to explore in future research.

We briefly described how the sample size within terminal nodes and the overall tree size both need to approach infinity. This does not provide much guidance about how to select a good tuning parameter for the minimum number at risk with a given data set. In practice, cross-validation could be used to select the most appropriate minimum number at risk. However, stacking bagged survival trees across a range of tuning parameters may result in better performance [see Chapter 3].

The bagged survival tree used to estimate the weights can be considered as a non-parametric estimator of the conditional quantile function, i.e., equation (2.1). Essentially, the bagged trees can predict quantile values for particular covariate values similar to Meinshausen [2006]. While this is potentially useful for predicting survival times,

this does not provide information on the relationship of the covariates with the event distribution. Rudser et al. [2012] shows how these predicted values could be used to form linear contrasts, while local regression extensions (e.g., splines) are straightforward.

Chapter 3

Stacked Survival Models

Survival function estimation has long been a major component of survival analysis [Kaplan and Meier, 1958]. Yet estimation of conditional survival functions, i.e., survival functions that depend on covariate values, remains a challenging problem. A common semi-parametric approach combines the Cox proportional hazard model with a baseline hazard estimate, e.g., see Kalbfleisch and Prentice [2002]. However, if the functional form is misspecified or the proportional hazards assumption is violated, then this approach may perform poorly. In terms of the bias-variance tradeoff, the Cox model, and other parametric models, achieve low variance by making distributional and functional form assumptions. If the assumptions are approximately correct, then the bias term is small and the parametric and semi-parametric models perform well. On the other hand, if the assumptions are badly violated, then the bias term can be large and the models perform poorly.

Many non-parametric methods have been proposed to overcome the bias induced by violated assumptions. For example, Kooperberg et al. [1995] proposes a flexible spline approach for the log-hazard that encompasses more than a proportional hazards model. Alternatively, tree-based approaches have been considered by several authors [Bou-Hamad et al., 2011, Ishwaran et al., 2008, Zhu and Kosorok, 2012]. Despite possessing low bias in a wide variety of situations, non-parametric estimators suffer from high variance and can require a large sample size to perform well. This can lead to surprising situations where misspecified parametric models perform better (in terms of mean-squared error) than non-parametric estimators. Specifically, the bias of misspecified parametric models is smaller than the variance of non-parametric estimators, i.e., the bias-variance trade-off.

This chapter pursues a flexible estimator of a conditional survival function, i.e., an estimator that can have low variance when parametric assumptions are approximately correct and robust when parametric assumptions are violated. Traditionally, a single

conditional survival function estimator is chosen from a set of candidate models, e.g., using an information criterion [Kooperberg et al., 1995] or through cross-validation. Rather than select a single survival model, our goal is to estimate an optimally weighted combination of survival models.

A variety of approaches that combine several models, often referred to as ensembles, have been explored in the uncensored setting. One approach, called "stacking," determines the optimally weighted average of models by minimizing predicted error. Wolpert [1992] introduced stacking in the context of neural networks, while Breiman [1996b] extended the idea to uncensored regression models and showed that stacking could improve prediction error. In particular, Breiman [1996b] found that combining fundamentally different regression models, e.g., ridge regression and subset regression, had the largest reduction in prediction error. LeBlanc and Tibshirani [1996] found stacking with a constraint of non-negative weights to be an efficient way to combine models. Van der Laan et al. [2007] independently developed uncensored stacking as the 'Super Learner' algorithm, and presented results regarding the stacked estimator's rate of convergence. More recently, Boonstra et al. [2013] used stacking to improve prediction when incorporating different generation sequencing information in high dimensional genome analysis.

Stacking in a censored data setting presents additional challenges. Polley and Van der Laan [2011] mention stacking within a general censored data framework and provide an example for hazard function estimation. Our approach differs in two significant ways. First, we focus on estimating conditional survival functions rather than a hazard function, which requires a different loss function that is tailored to directly estimating survival functions. We also pursue the potential advantages of stacking parametric, semi-parametric, and non-parametric estimators. In particular, we show that stacked survival models perform well by giving a majority of weight to approximately correct parametric models, while shifting weight to non-parametric estimators when assumptions are violated. This allows stacked survival models to outperform the single model selected via cross-validation and, in some situations, outperform every individual model considered in the stacking procedure. We believe that combining parametric, semi-parametric, and non-parametric estimators is the biggest advantage of stacked survival models.

The remainder of this chapter is organized as follows: uncensored stacking and the extension to censored data are introduced in Section 3.1. Section 3.2 investigates the mean-squared error of stacked survival models. Some asymptotic properties of stacked estimator are discussed in Section 3.3. Section 3.4 investigates the finite sample performance through an extensive simulation study. Stacked survival models are then applied to the German breast cancer study data set in Section 3.5, with a summary presented in Section 3.6.

3.1 Stacking

Throughout the chapter, random variables and observed variables are distinguished by capital and lower case letters, respectively. Our objective is to estimate the survival function of the event time random variable T that depends on p baseline covariates $\boldsymbol{x}$, i.e., $S_o(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x})$. In survival analysis, T may only be partially observed due to a censoring random variable C that may also depend on $\boldsymbol{x}$. Define the conditional survival function of the censoring distribution as $G(t|\boldsymbol{x}) = P(C > t|\boldsymbol{x})$. We assume throughout that the event time and censoring random variables are conditionally independent, i.e., $T \perp C|\boldsymbol{x}$. The observed time is $y_i = \min(t_i, c_i)$, and $\delta_i = I(t_i < c_i)$ indicates whether an event was observed. Hence a sample of right censored survival data of size n consists of triplets $\{y_i, \delta_i, \boldsymbol{x}_i\}$, $i = 1, ..., n$. Using the observed triplets, we can construct, for example, an estimate of the event time survival function from each of m candidate models with the k^{th} estimate denoted as $\hat{S}_k(t|\boldsymbol{x})$.

3.1.1 Uncensored Stacking

Stacking requires predicting outcomes with each model, then finding the combination that minimizes predicted error. In the uncensored case, the event time t_i is observed for all $i = 1, ..., n$, so the predicted values are $\hat{t}_{i,k}$ for the k^{th} model ($k = 1, ..., m$). Since increasing model complexity will improve training set prediction but not necessarily true error, the predicted values are commonly estimated via n-fold cross-validation. That is, $\hat{t}_{i,k}^{(-i)}$ is calculated by fitting the k^{th} model without the i^{th} observation. Then minimizing squared predicted error implies

$$\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}, \alpha_k \geq 0} \sum_{i=1}^{n} (t_i - (\sum_{k=1}^{m} \alpha_k \hat{t}_{i,k}^{(-i)}))^2, \tag{3.1}$$

where the final model predictions are $\hat{t}_i = \sum_{k=1}^{m} \hat{\alpha}_k \hat{t}_{i,k}$. The non-negativity constraint, $\hat{\alpha}_k \geq 0$ for all $k = 1, ..., m$, is not required, but has been shown to perform well [Breiman, 1996b, LeBlanc and Tibshirani, 1996].

3.1.2 Censored Stacking

Stacking survival models is not immediately straightforward. In particular, equation (3.1) is effectively an unknown quantity in the presence of censoring (e.g., t_i is not always observed). We are also interested in the entire survival curve for a given set of covariates and not just a single quantity. Adjustments are therefore required. While optimal measures of predictive error for survival models are not well established, we use

the Brier Score [Graf et al., 1999], which is commonly used and has a connection with squared error. Following Lostritto et al. [2012], the Brier Score at time t can be written as

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t)|\boldsymbol{x}_i)} \times \{Z_i(t) - \hat{S}(t|\boldsymbol{x}_i)\}^2, \tag{3.2}$$

where $Z_i(t) = I(t_i > t)$, $T_i(t) = \min\{t_i, t\}$, $\Delta_i(t) = I(\min\{t_i, t\} \leq c_i)$, and $G(\cdot|\boldsymbol{x}_i)$ is the conditional survival function of the censoring distribution. For a fixed time t, censored observations with $c_i > t$ will contribute to the Brier Score, but when $c_i < t$ the censored observations will only contribute to the Brier Score indirectly through the estimation of $G(\cdot|\boldsymbol{x}_i)$. Using double expectation arguments, it is possible to show that $BS(t)$ estimates the expected squared error of the survival distribution at time t. Thus the true conditional survival function, $S_o(t|\boldsymbol{x})$, is the minimizer of $E\{BS(t)\}$.

Since the goal is to estimate the entire conditional survival function, the Brier Score is minimized over a set of time points, say $t_1, ..., t_s$. As such, the stacking weights are estimated by minimizing the following weighted least squares objective function with the additional constraint that $\sum_{k=1}^{m}\hat{\alpha}_k = 1$

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}, \alpha_k \geq 0}\sum_{r=1}^{s}\sum_{i=1}^{n}\frac{\Delta_i(t_r)}{G(T_i(t_r)|\boldsymbol{x}_i)} \times \{Z_i(t_r) - \sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\}^2, \tag{3.3}$$

where $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ is the survival estimate from the k^{th} model that does not include the i^{th} observation in the fitting process. Finally, the stacked estimate of the conditional survival function with time-independent weights is

$$\hat{S}(t|\boldsymbol{x}) = \sum_{k=1}^{m}\hat{\alpha}_k\hat{S}_k(t|\boldsymbol{x}), \tag{3.4}$$

where $\hat{S}_k(t|\boldsymbol{x})$ is the k^{th} survival model estimated with all the data.

The computational requirements of non-parametric estimators generally prevent estimating $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ with n-fold cross-validation. Instead, v-fold cross-validation can be used ($v < n$) as the important point is that the i^{th} observation is not used to estimate $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$. For example, Breiman [1996b] found that 10-fold cross-validation performed similarly to n-fold cross-validation for stacking uncensored regression models. In addition, the simulation studies in Section 3.4 illustrate that stacked survival models perform well even when $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ is estimated via five-fold cross-validation.

Remark 3.1. The Brier Score measures agreement at only one particular time. As such, the value(s) of t over which it is evaluated, i.e., $t_1, ..., t_s$, have implications for

performance. In particular, care should be taken to avoid picking only very small, or very large t values. Preliminary investigations suggest that using at least nine evenly spaced quantiles of the observed event distribution ensures good performance.

Remark 3.2. Time-dependent stacking, i.e., allowing the weighted combination of models to depend on time, was also considered. Though potentially adding flexibility, a major flaw of time-dependent stacking is that the conditional survival function may, at times, increase, which violates a fundamental property of survival functions. As such, we focus on time-independent stacking while Appendix B discusses time-dependent stacking in greater detail.

3.2 Mean-Squared Error Decomposition

We analyze the decomposition of mean-squared error for stacked survival models. We start by defining the mean-squared error for the stacked estimator as $\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} = E\{E_{\boldsymbol{x}} \int_0^\tau [\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt\}$, where the outer expectation is taken with respect to the estimator. This definition of mean-squared error is motivated, in part, by the Brier Score. In particular, Appendix B shows that $E\{E_{\boldsymbol{x}} \int_0^\tau BS(t)dt\} = \sigma^2 + \text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\}$, where σ^2 is irreducible prediction error. Similar to the analysis of Fumera and Roli [2005], we show in Appendix B that the mean-squared error decomposes into

$$\begin{aligned}\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} \quad = \quad & \sum_{k=1}^{m} \alpha_k^2 \text{MSE}_\tau\{\hat{S}_k(\cdot|\boldsymbol{x})\} + \\ & E_{\boldsymbol{x}} \sum_{k=1}^{m} \sum_{l \neq k} \alpha_k \alpha_l \int_0^\tau \Big[\text{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} \times \text{Bias}\{\hat{S}_l(t|\boldsymbol{x})\} + \\ & \text{Corr}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\} \times \text{Var}\{\hat{S}_k(t|\boldsymbol{x})\}^{\frac{1}{2}} \times \text{Var}\{\hat{S}_l(t|\boldsymbol{x})\}^{\frac{1}{2}}\Big] dt,\end{aligned}$$

where $\text{MSE}\{\hat{S}_k(\cdot|\boldsymbol{x})\}$, $\text{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} = E\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})$, and $\text{Var}\{\hat{S}_k(t|\boldsymbol{x})\} = E\{[\hat{S}_k(t|\boldsymbol{x}) - E\hat{S}_k(t|\boldsymbol{x})]^2\}$ are, respectively, the mean-squared error, bias at time t, and variance at time t for the k^{th} survival model in the stacking procedure, while $\text{Corr}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\}$ is the correlation in predicted survival at time t between the k^{th} and l^{th} survival model.

We note that, given a set of candidate survival models, it is easy to show that there exists a set of stacking weights such that the stacked estimator possesses as good, or better, mean-squared error as the best performing single model in the set of candidate survival models. However, this property is *not* guaranteed after estimating the stacking weights. As such, the careful selection of candidate survival models should help the stacking procedure to achieve good performance.

The decomposition of mean-squared error provides convenient insight into an appropriate set of candidate survival models. In particular, the set of candidate survival models should be relatively independent of each other (i.e., possess a lower correlation between predicted survival functions). Similar to Breiman [1996b], this motivates stacking a diverse set of fundamentally different survival models, e.g., a range of parametric, semi-parametric, and non-parametric survival models. In contrast, non-parametric survival models with different tuning parameter combinations are likely highly correlated. This indicates that selecting a single tuning parameter combination may work better than stacking across a set of tuning parameter combinations. In addition, survival models with the same distributional assumptions but different covariate combinations are likely highly correlated, which indicates that the appropriate covariate combination should be determined prior to stacking the candidate survival models.

3.3 Asymptotic Properties

We show that stacked survival models can ensure consistent model selection and uniform consistency of the conditional survival function estimate. The former refers to the idea that if the set of stacked models contains uniformly consistent models, then all weight is asymptotically given to those models in the stack. Consistent model selection implies uniform consistency as long as the correctly specified model is a uniformly consistent estimator of the conditional survival function. Our main assumption for consistent model selection is that there exists no weighted average of misspecified models that approaches the true survival function for every time point included in equation (3.3). Appendix B contains the specific assumptions and proofs.

Let $\Omega = (0, \tau)$ be the support of interest for estimating the conditional survival function, and consider m estimators for the stacking procedure. Then

Theorem 3.1. *Let $\hat{\boldsymbol{\alpha}}$ be estimated by equation (3.3). Assume that models $1, ..., l$, where $l < m$, are the only uniformly consistent estimators and conditions (B1)-(B3) in Appendix B hold, then $\sum_{k=1}^{l} \hat{\alpha}_k \to 1$, in probability, as $n \to \infty$.*

This ensures that, for time-independent weights, the correct model(s) will asymptotically receive all of the weight for the stacked conditional survival function estimate in equation (3.4). There can be more than one uniformly consistent estimator when considering different tuning parameters for non-parametric estimators. Another example is a correctly specified Weibull model and Cox model. In the special case, when only one model is uniformly consistent, we obtain the corollary:

Corollary 3.2. *If $\hat{S}_1(t|\boldsymbol{x})$ is the only uniformly consistent estimator, then $\hat{\alpha}_1 \to 1$, in probability, as $n \to \infty$.*

Theorem 3.1 and Corollary 3.2 are required for uniform consistency of the stacked estimator with time-independent weights.

Theorem 3.3. *Let the stacked estimate of the conditional survival function be defined as $\hat{S}(t|\boldsymbol{x})$ in equation (3.4). Assume that conditions (B1)-(B3) in Appendix B hold. Then as $n \to \infty$,*

$$\sup_{t \epsilon \Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| \to 0.$$

Both theorems are proved in Appendix B. The rate of convergence of the stacked estimator remains beyond the scope of this Chapter. However, Van der Laan et al. [2007] showed that, in the uncensored case, the stacked estimator's risk converges at either the best rate of a correctly specified model, or slightly slower than the parametric rate.

3.4 Simulations

An extensive simulation study examines the finite sample performance of stacked survival models for several commonly encountered situations. In particular, three settings are investigated: a moderate number of covariates with a modest censoring rate (Section 3.4.1) and a high censoring rate (Section 3.4.2), and then a large number of covariates with a modest censoring rate (Section 3.4.3).

The simulations are comprised of combinations of an event distribution ($d = 1, 2, 3$) and linear form of covariates ($q = 1, 2$). The covariate distributions are multivariate normal: $\boldsymbol{x}_p \sim MVN(\mathbf{0}, \Sigma)$, where Σ is the correlation matrix and for all $i, j = 1, ..., p$, $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.4$ (p is the vector dimension). Sections 3.4.1 and 3.4.2 have a four-dimensional covariate space (i.e., $p = 4$), while Section 3.4.3 has a $p = 40$ dimensional covariate space. For all simulations, only the first and third covariate have a non-zero effect. Specifically, the first three covariate effects are $(1, 0, -1)$. Two different linear combinations are considered: $\boldsymbol{\gamma}^1 = \boldsymbol{x}_p$ and $\boldsymbol{\gamma}^2 = \Phi(4 \times \boldsymbol{x}_p)$ which imply linear and non-linear covariate effects, respectively. The event distributions are defined as

1. $T_1^{(q)} \sim \exp\{\text{Normal}(\boldsymbol{\beta}\boldsymbol{\gamma}^q, \frac{1}{4})\}$
2. $T_2^{(q)} \sim \text{Weibull}(\text{scale} = \exp\{\boldsymbol{\beta}\boldsymbol{\gamma}^q\}, \text{shape} = 1.1)$

3. $T_3^{(q)} \sim \text{Gamma}(\text{scale} = \frac{1}{4}\exp\{\boldsymbol{\beta}\boldsymbol{\gamma}^q\}, \text{shape} = 5)$

Each subsection investigates every combination of the event distribution (d) and linear form (q), i.e., there are six scenarios for each of the three subsections.

Survival models are compared on the basis of integrated squared survival error (ISSE), $E_{\boldsymbol{x}} \int_T (\hat{S}(u|\boldsymbol{x}) - S_o(u|\boldsymbol{x}))^2 du$, which is approximated by

$$\text{ISSE} \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}(t_{(j)} - t_{(j-1)}) \times (\hat{S}(t_{(j)}|\boldsymbol{x}_i) - S_o(t_{(j)}|\boldsymbol{x}_i))^2,$$

where $t_{(j)}$ are the ordered event times, i.e., $t_{(j)} - t_{(j-1)} > 0$ for all j, N is the number of observed event times, and $S_o(\cdot|\boldsymbol{x}_i)$ is the true conditional survival function. Therefore, ISSE measures the squared distance between the estimated and true conditional survival functions.

We use the integrated Brier Score (IBS), which is a similar criterion to the estimation of $\hat{\alpha}_k$, as the measure of the predicted error for selecting an individual model through cross-validation. In particular, the IBS for the k^{th} model is defined as $\hat{\text{IBS}}_k = \int_0^\tau \hat{BS}_k(t)dt$, where τ is the maximum observed time and $\hat{BS}_k(t)$ is the estimated Brier Score at time t for the k^{th} model (with an out-of-bag estimate of the conditional survival function). The cross-validated estimator is then defined as $\hat{S}_l(\cdot|\boldsymbol{x})$, where $l = \arg\min_k \hat{\text{IBS}}_k$.

All simulations were run in R version 3.0.0 [R Development Core Team, 2013]. The constrained minimization problem was solved using the package `alabama` [Varadhan, 2012]. The stacking weights, i.e., equation (3.3), were estimated by minimizing the Brier Score over the $0.1, 0.2, ..., 0.9$ quantiles of the observed event distribution.

3.4.1 Modest Censoring

This setting has relatively few covariates ($p = 4$) with a modest censoring rate (25%) and sample size ($n = 200$). This illustrates stacked survival models in a relatively straightforward and simple scenario.

The stacked survival models include a Weibull model and log-Normal model as parametric models, a Cox proportional hazards model as a semi-parametric model, and random survival forests (RSF) as a non-parametric model. The parametric and semi-parametric models only include first-order main effects and no interactions. All of the parametric and semi-parametric models are estimated using the `survival` package in R [Therneau, 2013], and all of the parametric and semi-parametric models use five-fold cross-validation

to estimate $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$. The RSF is estimated using the `randomSurvivalForest` package in R [Ishwaran and Kogalur, 2013]. The RSF is an ensemble of 250 trees grown using package defaults. For RSF, $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ is estimated using the out-of-bag ensemble from the `rsf` function. The censoring distribution is a uniform distribution for all $T_d^{(q)}$: $C_{d,q} \sim \text{Unif}(0, c(d,q))$, where $c(d,q)$ is a constant that depends on (d,q) and ensures approximately 25% censoring. Kaplan-Meier estimates the survival function of the censoring distribution required for the Brier Score, i.e., $G(\cdot)$ in equation (3.3).

The log-Normal and Weibull scenarios with linear covariate effects illustrate performance when there is a correctly specified parametric or semi-parametric model in the stack. Stacking is not expected to perform better than a correctly specific parametric model, but should remain reasonably efficient in such situations. The Gamma scenario with linear covariate effects illustrates performance when there are approximately correct parametric models in the stack (e.g., a correct mean function). The scenarios with non-linear covariate effects were designed to have badly misspecified parametric and semi-parametric models. Due to the lack of a correctly specified parametric model, stacked survival models should perform relatively well by, in particular, assigning more weight to the non-parametric estimator: random survival forests (RSF).

Table 3.1 presents the results in terms of integrated squared survival error (ISSE). Since the goal is an estimator that performs well in a wide variety of situations, the top two estimators are bolded for each scenario. The stacked survival model, i.e., "Stacking", is a top two estimator for five of the six scenarios. Stacked survival models reduce the ISSE by approximately 20% compared to the best single model for the log-Normal and gamma distributions with non-linear covariate effects. In addition, the stacking procedure outperforms the approach of selecting a single model via cross-validation in every situation.

As an illustration, Table 3.2 presents the stacking weights (averaged over all simulations) for the individual models. For the linear scenarios, the stacking procedure gives a majority of the weight to correctly specified parametric models. The weights are even more interesting for the scenarios with non-linear covariate effects. In particular, the random survival forests (RSF) receive the most weight for the stacking procedure despite having the largest ISSE among single models. This is a good example of stacked survival models combining misspecified parametric models and an inefficient non-parametric model to obtain a new estimator that outperforms every single model considered in the stacking procedure.

Remark 3.3. Random survival forests (RSF) possess tuning parameters that influence performance, e.g., the minimum number of events in a node. While the performance

of RSF could be improved by adaptively selecting tuning parameters (e.g., by cross-validation), stacked survival models are likely to also inherit any improvement in RSF since it is included in the stack.

TABLE 3.1: Simulation results for Section 3.4.1 ($n = 200$, $p = 4$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 100. The two top estimators are bolded for each simulation scenario. 'RSF' stands for random survival forests, 'Stacking' is stacked survival models, and 'CV' is the cross-validated selected estimator.

		Models	log-Normal	Weibull	Gamma
Linear Effects	Single Models	log-Normal	**0.92**	2.48	**1.17**
		Weibull	1.81	**1.26**	**1.29**
		Cox	3.76	2.03	3.49
		RSF	31.5	18.1	36.4
	Flexible Models	Stacking	**1.24**	**1.62**	1.34
		CV	2.97	1.99	2.38
Non-Linear Effects	Single Models	log-Normal	**9.34**	5.73	**11.6**
		Weibull	10.1	**4.77**	12.0
		Cox	10.4	5.02	12.4
		RSF	11.4	10.5	13.0
	Flexible Models	Stacking	**7.59**	**4.80**	**9.02**
		CV	10.3	5.62	12.1

TABLE 3.2: Average weights for the individual models included in the stacked survival model for each of the six scenarios in Section 3.4.1 ($n = 200$, $p = 4$ covariates, and 25% censoring). Each simulation is replicated 2000 times. 'RSF' stands for random survival forests.

	Stacked Models	log-Normal	Weibull	Gamma
Linear Effects	log-Normal	0.64	0.17	0.43
	Weibull	0.21	0.49	0.39
	Cox	0.13	0.29	0.16
	RSF	0.03	0.06	0.02
Non-Linear Effects	log-Normal	0.21	0.12	0.14
	Weibull	0.10	0.31	0.14
	Cox	0.03	0.22	0.04
	RSF	0.67	0.35	0.68

3.4.2 High Censoring

This setting is similar to Section 3.4.1 except that the censoring rate is approximately 75% and the sample size is $n = 1000$, which is designed to mimic large observational

trials that experience substantial administrative censoring at the end of the observed support. To simulate administrative censoring, the censoring is uniformly distributed: $C_{d,q} \sim \text{Unif}(c(d,q), c(d,q) + 0.5)$, where $c(d,q)$ is a constant that depends on (d,q) and ensures approximately 75% censoring.

Table 3.3 presents the results in terms of integrated squared survival error (ISSE). Again, the top two estimators are bolded to highlight flexibility in a wide range of scenarios. Stacked survival models is a top two estimator for five of the six scenarios, while none of the alternatives is a top two estimator for more than two scenarios. Additionally, stacking possesses approximately 30% higher ISSE than correctly specified parametric models (i.e., log-Normal and Weibull distributions with linear effects), and as good or better ISSE when the parametric models are slightly misspecified (i.e., Gamma distribution with linear effects). The stacking procedure also outperforms the model selected via cross-validation in every situation.

An interesting point is that the parametric and semi-parametric models perform very poorly for the log-Normal and Gamma scenarios with non-linear covariate effects, where RSF is the best performing single survival model. Despite the poor performance of the parametric models, stacking is able to improve on the performance of RSF to achieve more than a 10% reduction in integrated squared survival error. In contrast, cross-validation, which always selects RSF as the best performing model, is naturally unable to improve upon the performance of RSF. This again illustrates the ability of stacked survival models to perform well in a wide range of situations *and*, at times, improve the estimation of the conditional survival function above any model in the stack.

3.4.3 High Dimensional Covariate Space

This setting has $p = 40$ covariates with a relatively small sample size ($n = 200$). The censoring distributions are the same as Section 3.4.1. This represents the situation where the number of covariates is large relative to the sample size. In general, the parametric and semi-parametric models used (and stacked) in Sections 3.4.1 and 3.4.2 will not perform well in high dimensional settings without regularization. As such, these models are not included for these high dimensional scenarios. We instead stack a Cox model with an l_1 penalty (i.e., lasso), a boosted version of the Cox model, and random survival forests (RSF). The l_1 penalized version of the Cox model is fit using the R package `penalized` with the penalty parameter chosen via cross-validation [Goeman, 2012]. The boosted Cox model is fit using the package `CoxBoost` in R with default tuning parameters [Binder, 2013]. RSF is fit in the same manner as Sections 3.4.1 and 3.4.2.

TABLE 3.3: Simulation results for Section 3.4.2 ($n = 1000$, $p = 4$ covariates, and 75% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 1000. The two top estimators are bolded for each simulation scenario. 'RSF' stands for random survival forests, 'Stacking' is stacked survival models, and 'CV' is the cross-validated selected estimator.

		Models	log-Normal	Weibull	Gamma
Linear Effects	Single Models	log-Normal	**0.26**	1.02	**0.38**
		Weibull	0.80	**0.33**	0.45
		Cox	0.99	**0.41**	0.69
		RSF	7.64	4.84	8.16
	Flexible Models	Stacking	**0.34**	0.42	**0.37**
		CV	0.93	0.71	0.55
Non-Linear Effects	Single Models	log-Normal	13.2	2.96	14.9
		Weibull	14.1	**2.84**	15.6
		Cox	14.0	2.90	15.6
		RSF	**5.20**	5.17	**5.63**
	Flexible Models	Stacking	**4.56**	**2.29**	**5.01**
		CV	**5.20**	3.47	**5.63**

The stacked survival model is the best performing model for every non-proportional hazards scenario (see Table 3.4). Stacked survival models are one of the top two estimators for five of the six scenarios and, relative to Sections 3.4.1 and 3.4.2, stacked survival models offer smaller improvements (approximately $5\% - 15\%$ reductions in ISSE). However, the improvements in ISSE are more consistent across the scenarios. In addition, the stacking procedure outperforms the model selected via cross-validation in every situation.

3.5 German Breast Cancer Study

Stacked survival models are illustrated on a well-known survival benchmark data set: the German breast cancer study (GBCS) described by Hosmer et al. [2008], and accessible at the University of Massachusetts website for statistical software information. There are eight covariates included in the analysis: age at diagnosis, tumor size, tumor grade, number of nodes, menopausal status, the number of progesterone receptors, the number of estrogen receptors, and hormone therapy status. The outcome of interest is the time till death, and there is complete data on 686 patients with approximately 75% censoring, which is similar to the simulation scenarios in Section 3.4.2. The stacking procedure uses the same models as Sections 3.4.1 and 3.4.2. That is, the Weibull and log-Normal model are the parametric models, the Cox proportional hazards model is the semi-parametric

TABLE 3.4: Simulation results for Section 3.4.3 ($n = 200$, $p = 40$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 100. The two top estimators are bolded for each simulation scenario. 'RSF' stands for random survival forests, 'Stacking' is stacked survival models, and 'CV' is the cross-validated selected estimator.

		Models	log-Normal	Weibull	Gamma
Linear Effects	Single Models	Cox - Lasso	5.57	3.66	6.03
		Cox - Boosting	5.60	**3.14**	6.82
		RSF	49.6	27.4	57.4
	Flexible Models	Stacking	**4.65**	**3.30**	**4.96**
		CV	**4.82**	3.33	**5.13**
Non-Linear Effects	Single Models	Cox - Lasso	11.1	6.32	13.2
		Cox - Boosting	**10.6**	**6.21**	**12.5**
		RSF	16.0	12.9	18.8
	Flexible Models	Stacking	**9.81**	**6.26**	**11.6**
		CV	10.8	6.40	12.9

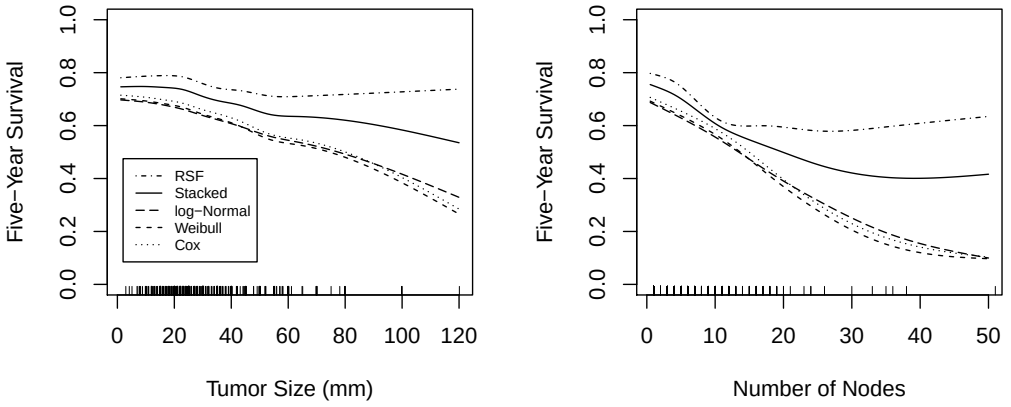
model, and a random survival forest is the non-parametric model. The minimum number of deaths (for RSF) is set at 12, which was selected by minimizing predicted error among five potential values: $3, 6, 12, 24, 48$.

We are particularly interested in the association of five-year survival with tumor size and the number of nodes. To evaluate the association, the stacked survival model and each model included in the stacking procedure predicts the five-year survival rate for each patient in the study. After predicting five-year survival, a generalized additive model with penalized B-splines for the continuous covariates [i.e., the `gam` function from the `mgcv` package [Wood, 2006]] estimates the association of five-year survival with tumor size and the number of nodes while adjusting for the other covariates.

Figure 3.1 presents the estimated five-year survival as a function of tumor size and the number of nodes at the median of the other covariates. The parametric/semi-parametric models suggest worse five-year survival with increasing tumor size and number of nodes. In contrast, RSF suggests that five-year survival dips slightly around $40mm$ for tumor size, while five-year survival for the number of nodes has a sharp early decrease but plateaus after about 10 nodes. The stacked survival model - which gives weight to the Weibull model (0.36), the Cox model (0.04), and RSF (0.58) - is a compromise between the parametric/semi-parametric models and RSF.

The GBCS data set has a marginal five-year survival rate of 70% due, in part, to a censoring rate of 75%. As such, predicted five-year survival rates less than 20% are surprising (i.e., the parametric/semi-parametric models for the number of nodes). Due

FIGURE 3.1: The association of tumor size (mm) and the number of nodes with five-year survival for the GBCS data set with the other covariates to their median value. The tick marks at the bottom of the plots indicate the skewness of both covariates.



to the sparsity of patients with more than 20 nodes, the low model based predicted probabilities are likely due to parametric/semi-parametric models being heavily influenced by a strong negative association with survival for patients with less than 20 nodes (98% of patients have less than 20 nodes) through the first order linear effect (note that the patient with over 50 nodes was censored after two years). In contrast, RSF does not require any linearity assumptions and is more influenced by local observations in predicting five-year survival [Ishwaran et al., 2008]. From this perspective, the stacked survival model is balancing model based predictions that require assumptions of linearity with locally based predictions.

3.6 Summary

We propose stacking survival models to flexibly estimate conditional survival functions. Stacked survival models can combine several models, spanning the full range of parametric, semi-parametric, and non-parametric estimators, in the hope of potentially gaining efficiency from the parametric estimators while maintaining the robustness of non-parametric estimators. As illustrated in the simulation study, stacked survival models give more weight to parametric and semi-parametric models when assumptions are approximately correct, but shift weight to non-parametric estimators when assumptions are badly violated. In this manner, stacked survival models perform well across a wide range of scenarios. In particular, for a given scenario, stacked survival models are consistently one of the two best estimators and, at times, perform better than any single model considered in the stacking procedure.

In practice, the true underlying data generation process is never known, i.e., one does not choose the event distribution or functional form of the covariates. This motivates an adaptive approach that can perform well in a wide variety of situations. Cross-validation is currently the most common adaptive approach. Yet the simulation study illustrates that stacked survival models perform as good, or better, than the model selected through cross-validation, which picks a single model to receive all the weight (i.e., $\alpha_k = 1$ for some k). As such, stacked survival models warrant consideration whenever cross-validated models are used. Other predictive models could also have been considered, though stacked survival models could inherit any particular advantages of such models through inclusion in the stack.

Appendix B includes an investigation of time-dependent stacking, but the resulting survival function is not guaranteed to be non-increasing. This is a major flaw that compromises the conceptual cohesion of time-dependent stacking. As an alternative, time-dependent stacking on the conditional hazard function would ensure a non-increasing survival function. However, this requires a different loss function than the Brier Score and remains beyond the scope of this thesis.

Similar to time-dependent stacking, covariate dependent stacking (or, allowing the α_k to depend upon $\boldsymbol{x}$) is a potential avenue for improving stacked survival models. LeBlanc and Tibshirani [1996] mention this approach for uncensored stacking, and a collaborative group using covariate dependent stacking won the Netflix Prize competition to improve movie recommendations [Sill et al., 2009]. However, extending the stacking procedure to include covariate dependent weights with the constraints introduced here is not straightforward. For example, Sill et al. [2009] do not constrain their covariate dependent weights despite prior experiences suggesting regularization improves performance [Breiman, 1996b, LeBlanc and Tibshirani, 1996]. Investigation of covariate dependent stacking and different approaches to constraining the covariate dependent weights deserves further investigation.

The set of m survival models included in the stacking procedure will influence the performance of the final stacked survival model. As discussed in Section 3.2, the mean-squared error of the stacked survival model depends, in part, on the correlation between predicted survival functions. Similar to Breiman [1996b], this motivates stacking survival models that are fundamentally different. As such, models based on different classes are recommended, e.g., parametric, semi-parametric, and non-parametric estimators. In addition, the set of important covariates should be determined prior to forming the set of candidate survival models. Further characterization, e.g., through empirical studies, of this topic is warranted.

The Brier Score, used to estimate the weighted combination of survival models, is essentially an inverse probability-of-censoring weighted (IPCW) estimate of prediction error. The IPCW estimate requires estimating the (possibly conditional) censoring distribution. The simulation scenarios introduced in Section 3.4 use a Kaplan-Meier estimator for the censoring distribution that is correctly specified. In our experience, the stacking procedure maintains good operating characteristics when the censoring model is misspecified. However, if there is strong evidence of differential censoring among the covariates, then a conditional estimator may be warranted (e.g., a Cox proportional hazards model).

The importance of efficient, yet robust, estimators of conditional survival functions (or, equivalently, conditional distribution functions) continues to grow. Methods in a wide range of areas require estimating a conditional survival function as a nuisance parameter, for example, censored quantile regression (see Chapter 2), time-dependent ROC curves [Zheng and Heagerty, 2004], inverse probability-of-censoring weighted estimators, e.g., Fine and Gray [1999], model-free contrast approaches [Rudser et al., 2012], and dynamic treatment regime methods [Zhao et al., 2011]. The simulations presented here suggest that stacking parametric, semi-parametric, and non-parametric models for the nuisance parameter will likely result in better estimation of regression parameters of interest, though these topics warrant further investigation.

Chapter 4

Restricted Mean Treatment Effects

Patients with end-stage lung disease (e.g., advanced cystic fibrosis) may be eligible for lung transplantation after other treatment options fail. Unfortunately, post-transplant survival is poor after lung transplantation, especially in comparison to other solid organ transplants, with one and three-year graft survival of 79% and 64%, respectively. Given the poor prognosis, understanding the factors related to post-transplant survival remains an important but controversial task. For example, transplant center volume (i.e., the number of lung transplants at a center over two years) is a factor that is related to post-transplant survival; in particular, higher center volume is associated with lower mortality [Thabut et al., 2010, Weiss et al., 2009]. In the absence of censoring, the difference in survival between high volume centers and low volume centers would be traditionally summarized by the difference in mean survival time. However, the mean for a non-negative random variable (e.g., survival time) is defined as $E\{T\} = \int_0^\infty S(t)dt$, where $S(t) = P(T > t)$ is the survival function of the random variable T. The estimate of the mean is therefore not defined when $\hat{S}(t) > 0$ for all observed t, a situation regularly experienced with even light censoring. Since substantial censoring is experienced in lung transplantation, a different summary measure is required.

The restricted mean is an alternative summary measure that is always estimable under censoring [Royston and Parmar, 2013]. The τ-restricted mean truncates observations at some time point τ [i.e., $E\{\min(T, \tau)\} = \int_0^\tau S(t)dt$]. By choosing a value for τ within the observed support, the restricted mean is a well-defined summary measure with a direct interpretation that is closely related to the mean. For example, the average difference in post-transplant survival between high volume and low volume centers over one year is the difference in one-year restricted means. However, estimating the difference of

the restricted mean survival time in observational studies, such as lung transplantation, is difficult due to potential confounding from observed covariates. In particular, the difference in the area under the Kaplan-Meier survival curve up to time τ between the two treatment groups is not a consistent estimator of the restricted mean difference.

To account for imbalances between treatments in confounding variables, several researchers have proposed estimating the covariate-adjusted restricted mean difference by estimating the covariate-adjusted survival distribution and then marginalizing over the covariate distribution to obtain the estimated restricted mean difference (referred to as the "regression" approach). For example, Karrison [1987] proposed modeling the survival time distribution as a proportional hazards model with a piecewise constant baseline hazard function. As a natural alternative, Zucker [1998] proposed a proportional hazards model with an unspecified baseline hazard, i.e., a Cox proportional hazards model [Cox, 1972]. Both Karrison and Zucker assume that the covariate effects are the same for each treatment. Chen and Tsiatis [2001] relax this assumption by estimating separate baseline hazard functions and covariate effects for each treatment.

All current approaches to estimating the difference in restricted means rely on a proportional hazards model to estimate the covariate-adjusted survival distribution. Unfortunately, the proportional hazards assumption may not hold in many applications. For example, centers with greater lung transplant volume are more likely to perform bilateral lung transplants (as opposed to single lung transplants), and the type of lung transplant is well-known to violate the proportional hazards assumption [Thabut et al., 2010]. As such, current approaches, which rely on the proportional hazards assumption, may produce biased estimators of the difference in restricted mean of post-transplant survival between high volume and low volume centers. As an alternative, we could estimate the survival time distribution with an accelerated failure time model (e.g., a log-Normal model), but the estimator is biased if the accelerated failure time assumption is violated. Rather than rely on either a proportional hazards model or an accelerated failure time model, we pursue a flexible estimator of the restricted mean that performs well across many situations.

This is a natural setting for the stacked survival models introduced in Chapter 3. As noted in that chapter, stacking finds the optimally weighted average of several conditional survival function estimators by minimizing predicted error. Since the minimization is based on predicted error, stacking can include parametric models, semi-parametric models (e.g., the Cox model) and non-parametric models. This allows the weight to shift to the model that most accurately estimates the underlying survival function for a given sample size. In this way, stacked survival models can accurately estimate a conditional

survival function, and the corresponding restricted mean, in situations that extend past proportional hazards scenarios.

Section 4.1 introduces restricted mean treatment effect estimation and stacked survival models. A simulation study evaluates the finite sample performance of the proposed estimator in Section 4.2. The proposed estimator is applied to a observational registry of post-lung transplantation survival from the United Network for Organ Sharing in Section 4.3. Concluding remarks are presented in Section 4.4.

4.1 Proposed Estimator

Some notation is required before introducing the estimator. Throughout the chapter, random variables and observed variables are distinguished by capital and lower case letters, respectively. The treatment, or condition, is denoted by a_i, where i denotes the patient, and follows the Bernoulli random variable A (i.e., $A = \{0, 1\}$). Additional covariates, denoted by vector $\boldsymbol{x}_i$, are measured at the beginning of the study and follow the distribution of the random variable $\boldsymbol{X}$. For this chapter, we define the non-negative survival time random variable as $T = T^0 \cdot I(A = 0) + T^1 \cdot I(A = 1)$, where T^0 and T^1 are the (possibly unobserved) survival time random variables had a patient received treatment 0 and 1, respectively. We assume that there are no unmeasured confounders; that is, the set of potential outcomes, (T^0, T^1), is conditionally independent of A given $\boldsymbol{X}$ (i.e., $(T^0, T^1) \perp A \mid \boldsymbol{X}$, where $\perp$ denotes statistical independence). The censoring time is c_i, which follows the distribution of the continuous non-negative random variable C and is assumed to be conditionally independent of (T^0, T^1) (i.e., $(T^0, T^1) \perp C \mid \{\boldsymbol{X}, A\}$). Hence a sample of right censored survival data for n patients is $\{y_i, \delta_i, a_i, \boldsymbol{x}_i\}$, $i = 1, ..., n$, where $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i < c_i)$.

Now let $S^{(A=a)}(t|\boldsymbol{X} = \boldsymbol{x}) = P(T > t|\boldsymbol{X} = \boldsymbol{x}, A = a)$ and $G^{(A=a)}(t|\boldsymbol{X} = \boldsymbol{x}) = P(C > t|\boldsymbol{X} = \boldsymbol{x}, A = a)$ be the treatment-specific conditional survival functions for the survival time and censoring time random variables, respectively, of treatment a. For the rest of the chapter, we suppress the covariate and treatment random variable notation in the conditional survival function; that is, $S^{(A=a)}(t|\boldsymbol{X} = \boldsymbol{x}) = S^{(a)}(t|\boldsymbol{x})$ and $G^{(A=a)}(t|\boldsymbol{X} = \boldsymbol{x}) = G^{(a)}(t|\boldsymbol{x})$.

4.1.1 Restricted Mean Treatment Effects

Following the outline of Chen and Tsiatis [2001], we estimate restricted means with the "regression" approach which involves modeling the conditional survival time distribution. In particular, the restricted mean for treatment a is defined as

$$\begin{aligned} \mu(\tau, a) \equiv E\{\min(T^a, \tau)\} &= E_{\boldsymbol{x}}[E\{\min(T^a, \tau)|\boldsymbol{X} = \boldsymbol{x}\}] \\ &= E_{\boldsymbol{x}}[E\{\min(T, \tau)|\boldsymbol{X} = \boldsymbol{x}, A = a\}] \qquad (4.1) \\ &= E_{\boldsymbol{x}}\left\{\int_0^\tau S^{(a)}(t|\boldsymbol{x})dt\right\}, \end{aligned}$$

where (4.1) holds due to the assumption that $(T^0, T^1) \perp A|\boldsymbol{X}$ (i.e., the assumption of no unmeasured confounders). It is important to note that the expectation is taken with respect to the marginal, rather than conditional, covariate distribution.

If we can estimate the conditional survival distribution $S^{(a)}(t|\boldsymbol{x})$, then the empirical covariate distribution estimates the expectation over the covariate space. As such, the estimator for the τ-restricted mean of T^a, which adjusts for potential confounding effects, is

$$\hat{\mu}(\tau, a) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \hat{S}^{(a)}(t|\boldsymbol{x}_i)dt, \qquad (4.2)$$

where $\hat{S}^{(a)}(t|\boldsymbol{x}_i)$ is the estimate of $S^{(a)}(t|\boldsymbol{x}_i)$. In practice, a closed form solution of equation (4.2) may not exist, and we therefore approximate equation (4.2) by

$$\hat{\mu}(\tau, a) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_\tau}(t_{(j)} - t_{(j-1)}) \times \hat{S}^{(a)}(t_{(j-1)}|\boldsymbol{x}_i), \qquad (4.3)$$

where $t_{(j)}$ are the ordered event times, e.g., $t_{(j)} - t_{(j-1)} > 0$ for all $j = 1..., n$, and N_τ is one more than the number of event times less than τ. If no event time equals τ (i.e., $t_{(j)} \neq \tau$ for all $j = 1, ..., n$), then $t_{(N_\tau)} = \tau$ and $t_{(N_\tau - 1)}$ is the largest event time less than τ. For two treatments, the estimated difference in restricted mean survival time is

$$\hat{\gamma}(\tau) = \hat{\mu}(\tau, a = 1) - \hat{\mu}(\tau, a = 0), \qquad (4.4)$$

which also corresponds to the difference in the area under the survival curves for the two treatments up to time τ.

At this point, it is obvious that the estimator of $S^{(a)}(t|\boldsymbol{x})$ influences the estimation of the difference in restricted means. In particular, a better estimator of $S^{(a)}(t|\boldsymbol{x})$ should result in a better estimator of the restricted mean difference with the approach outlined

here. This is, in fact, the main motivation for estimating restricted mean treatment effects with stacked survival models.

4.1.2 Stacked Survival Models

This chapter reintroduces stacked survival models in order to make appropriate modifications for estimating restricted means. However, we refer to Chapter 3 for an in-depth discussion on stacked survival models for estimating conditional survival functions.

The Brier Score [Graf et al., 1999] measures the predicted squared error of a conditional survival function at a particular time point. Following Lostritto et al. [2012], the estimated Brier Score for treatment a at a single time point t can be written as

$$\hat{BS}^{(a)}(t) = \frac{1}{n}\sum_{i\epsilon\Gamma_a}\frac{\Delta_i(t)}{\hat{G}^{(a)}(\min\{t_i,t\}|\boldsymbol{x}_i)} \times \{Z_i(t) - \hat{S}^{(a)}(t|\boldsymbol{x}_i)\}^2, \tag{4.5}$$

where $Z_i(t) = I(t_i > t)$, $\Delta_i(t) = I(\min\{t_i, t\} \leq c_i)$, $\hat{G}^{(a)}(\cdot|\boldsymbol{x}_i)$ is the estimated conditional survival function of the censoring distribution for the a^{th} treatment, and Γ_a is the set of patients that received treatment a. For a fixed time t, censored observations with $c_i > t$ will contribute to the Brier Score, but when $c_i < t$ the censored observations will only contribute to the Brier Score indirectly through the estimation of $G^{(a)}(T_i(t)|\boldsymbol{x}_i)$. For this chapter, $\hat{G}^{(a)}(\cdot|\boldsymbol{x}_i)$ is an (unconditional) treatment-specific Kaplan-Meier denoted here after as $\hat{G}^{(a)}(\cdot)$.

For each treatment group, the stacking procedure considers the same set of m candidate models, and each model has a corresponding conditional survival function estimate, say $\hat{S}_k^{(a)}(t|\boldsymbol{x})$ for $k = 1, ..., m$ survival models. Since the goal is estimating the entire conditional survival function to time τ, stacked survival models minimize $\hat{BS}^{(a)}(t)$ over a set of time points, say $t_1, ..., t_s$. Since there are m models to stack, the stacking weights are estimated by a weighted least squares problem with the additional constraints that $\alpha_k^{(a)} \geq 0$ for $k = 1, ..., m$ and $\sum_{k=1}^m \hat{\alpha}_k^{(a)} = 1$:

$$\hat{\boldsymbol{\alpha}}^{(a)} = \arg\min_{\boldsymbol{\alpha}^{(a)}, \alpha_k^{(a)} \geq 0} \sum_{r=1}^{s}\sum_{i\epsilon\Gamma_a}\frac{\Delta_i(t_r)}{\hat{G}^{(a)}(\min\{t_i,t_r\})} \times \{Z_i(t_r) - \sum_{k=1}^{m}\alpha_k\hat{S}_k^{(a,-i)}(t_r|\boldsymbol{x}_i)\}^2, \tag{4.6}$$

where $\hat{S}_k^{(a,-i)}(t|\boldsymbol{x}_i)$ is the survival estimate from the k^{th} model while leaving the i^{th} observation out during the fitting process. This ensures that stacking does not reward models that over-fit the data. This is traditionally done by leaving only the i^{th} observation out in the fitting process. However, the computational requirements induced

by bootstrapping prevent fitting the set of candidate survival models n separate times. The data is instead randomly split into five roughly equally sized sets and $\hat{S}_k^{(a,-i)}(t|\boldsymbol{x}_i)$ is obtained for observations in a given set by fitting the survival models to the observations in the other four sets. As such, five survival models, rather than n survival models, are fit for each of the m candidate survival models.

After minimizing equation (4.6), the stacked estimate of the conditional survival function for treatment a is

$$\hat{S}^{(a)}(t|\boldsymbol{x}) \quad = \quad \sum_{k=1}^{m} \hat{\alpha}_k^{(a)} \hat{S}_k^{(a)}(t|\boldsymbol{x}), \tag{4.7}$$

where $\hat{S}_k^{(a)}(t|\boldsymbol{x})$ is the k^{th} survival model estimated with all observations on treatment a. The treatment-specific restricted means and the restricted mean treatment effects are then estimated with equations (4.3) and (4.4), respectively.

The set of survival models has an influence on the performance of the stacked survival models. For example, Breiman [1996b] and Chapter 3 found that stacking a diverse set of models performs well. Due to their accessibility in the `survival` package, we use the Weibull and log-Normal models as the parametric survival models, and two versions of the Cox proportional hazards model as the semi-parametric survival models. In particular, the first Cox model has linear main effects, while the second Cox model has penalized splines for the continuous covariates. We also consider including bagged survival trees as the non-parametric survival model, which are easily estimated by the `randomSurvivalForest` package.

We estimate confidence intervals with the non-parametric bootstrap. In particular, we randomly sample with replacement n of the observed $\{y_i, \delta_i, a_i, \boldsymbol{x}_i\}$; this is called the b^{th} bootstrap data set. The b^{th} bootstrapped estimate of the k^{th} conditional survival function is defined as $\hat{S}_{k,b}^{(a)}(t|\boldsymbol{x})$. Since the stacking weights are re-estimated for each bootstrap, the final conditional survival function estimate of the b^{th} bootstrap for treatment a is $\hat{S}_b^{(a)}(t|\boldsymbol{x}) = \sum_{k=1}^{m} \hat{\alpha}_{k,b}^{(a)} \hat{S}_{k,b}^{(a)}(t|\boldsymbol{x})$. The bootstrap estimates are then defined similarly for the treatment specific restricted means and restricted mean treatment effect [see equations (4.3) and (4.4)]. Finally, the 95% confidence intervals are estimated using the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distributions.

Remark 4.1. Chapter 3 shows that, given the stack contains a uniformly consistent estimator, stacked survival models are uniformly consistent for the underlying conditional survival function. As such, when at least one model within the stacking procedure is correctly specified, $\hat{\mu}(\tau, a = l)$ is consistent for the true restricted mean treatment effect by the dominated convergence theorem [Ferguson, 1996]. The proposed estimator is

therefore consistent in a wider range of scenarios than previous methods that only use a proportional hazards model to estimate the conditional survival distribution.

Remark 4.2. The Brier Score measures agreement at only one particular time. As such, the set values (i.e., $t_1, ..., t_s$) over which the Brier Score is minimized [see equation (3.3)] have implications for performance. In particular, care should be taken to avoid picking only very small or very large values. At least nine evenly spaced quantiles of the observed event distribution is usually sufficient to ensure good estimation of the conditional survival function.

4.2 Simulations

The simulation study evaluates the finite sample performance of estimating restricted mean treatment effects with stacked survival models. Proportional hazard scenarios are investigated with an exponential distribution, while a gamma distribution represents non-proportional hazard scenarios. Each data-generating distribution includes a scenario with linear and non-linear covariate effects for a total of four scenarios.

For each simulation scenario, the covariate distribution is a four-dimensional multivariate Normal with a zero mean vector, unit variances, and a positive $AR(1)$ correlation structure ($\rho = 0.4$). To mimic observational studies with confounding, the treatment assignment depends on the covariate distribution. In particular, the probability of receiving treatment, i.e., $P(a_i = 1|\boldsymbol{x}) = p_i$, is defined by $\text{logit}(p_i) = 0.1 \times (x_1 + x_2 + x_3 + x_4)$. The event distributions for treatment a and scenario b are defined: when $b = 1, 2$ then $T_b^{(a)} \sim \text{Exp}[\exp\{\lambda_b^{(a)}\}]$, where $E\{T_b^{(a)}\} = 1/\exp\{\lambda_b^{(a)}\}$; when $b = 3, 4$ then $T_b^{(a)} \sim \text{Gamma}[\text{scale} = \exp\{\lambda_b^{(a)}\}, \text{shape} = 2.5]$, where $E\{T_b^{(a)}\} = 2.5 \cdot \exp\{\lambda_b^{(a)}\}$. The specific covariate effects for the control group (i.e., $\lambda_b^{(0)}$) are

$$
\begin{aligned}
\lambda_1^{(0)} &= -4.50 - 0.125 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\lambda_2^{(0)} &= -0.70 - 1.000 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\} \\
\lambda_3^{(0)} &= 3.50 - 0.125 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\lambda_4^{(0)} &= 4.50 - 0.500 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\},
\end{aligned}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Normal distribution (i.e., the non-linear effect is a 'smooth step function'). The covariate effects for the

treatment group (i.e., $\lambda_b^{(1)}$) are

$$\begin{aligned}
\lambda_1^{(1)} &= -3.50 - 0.125 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\lambda_2^{(1)} &= -1.70 - 1.000 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\} \\
\lambda_3^{(1)} &= \ 3.00 - 0.125 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\lambda_4^{(1)} &= \ 4.00 - 0.500 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\}.
\end{aligned}$$

The censoring distributions are defined similarly with $\lambda_b^{(a)}$ replaced by $\gamma_b^{(a)}$, and are designed to achieve a marginal censoring rate of approximately 30%. The censoring distributions for the control group (i.e., $\gamma^{(0)}$) are

$$\begin{aligned}
\gamma_1^{(0)} &= -4.895 - 0.0625 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\gamma_2^{(0)} &= -3.680 - 0.5000 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\} \\
\gamma_3^{(0)} &= \ 3.780 - 0.0625 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\gamma_4^{(0)} &= \ 4.765 - 0.5000 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\},
\end{aligned}$$

while the censoring distributions for the treatment group (i.e., $\gamma^{(1)}$) are

$$\begin{aligned}
\gamma_1^{(1)} &= -5.395 - 0.0625 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\gamma_2^{(1)} &= -4.680 - 0.5000 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\} \\
\gamma_3^{(1)} &= \ 3.280 - 0.0625 \cdot \{x_1 + x_2 + x_3 + x_4\} \\
\gamma_4^{(1)} &= \ 4.265 - 0.5000 \cdot \{\Phi(4 \times x_1) + \Phi(4 \times x_2) + \Phi(4 \times x_3) + \Phi(4 \times x_4)\}.
\end{aligned}$$

Each simulation scenario is evaluated 2500 times with a sample size of 300. All simulations were run in R version 3.0.0 [R Development Core Team, 2013]. The stacking weights are estimated by minimizing the Brier Score over nine equally spaced quantiles of the observed events, while the constrained minimization problem was solved using the `alabama` package [Varadhan, 2012].

The parametric models in the set of candidate models are the Weibull model and log-Normal model with only linear main effects. Both models are special cases of an accelerated failure time model, while the Weibull is also a special case of a proportional hazards model. The semi-parametric models are two versions of the Cox model. The first Cox model has only linear main effects, while the second Cox model uses penalized splines for main effects with the roughness penalty set to 0.5. The `survival` package estimates both the parametric and semi-parametric models [Therneau, 2013]. The non-parametric estimator in the set of candidate survival models is bagged survival trees, which are estimated with the `randomSurvivalForest` package [Ishwaran et al., 2008] with 1000 trees

grown. Bagged survival trees require setting a tuning parameter: minimum number of unique failure times (i.e., n_{min}).

We consider three different versions of stacked survival models. Two versions of the stacked estimator are distinguished by the value of n_{min} for bagged survival trees, while one version excludes bagged survival trees altogether from the set of candidate survival models. In particular,

1. The 'Stacked' estimator only includes the parametric and semi-parametric survival models in the set of candidate survival models.
2. The 'Stacked ($n_{min} = 3$)' estimator includes the parametric, semi-parametric, and non-parametric survival models with the minimum number of unique failure times set to three.
3. The 'Stacked ($n_{min} = 10$)' estimator includes the parametric, semi-parametric, and non-parametric survival models with the minimum number of unique failure times set to ten.

There are two motivations for investigating three different versions of the stacked estimator. First, bagged survival trees are computationally expensive, which is significantly compounded by the non-parametric bootstrap used to estimate confidence intervals. As such, we can save substantial computational time if the stacked estimator without bagged survival trees performs as well, or better, than the stacked estimators with bagged survival trees. Secondly, by including two versions of the stacked estimator with bagged survival trees, we can investigate the sensitivity of the restricted mean to the selection of the tuning parameter for bagged survival trees. In general, the ability of bagged survival trees to estimate the conditional survival function depends on the appropriate selection n_{min}. As such, the n_{min} value for bagged survival trees may affect the estimate of the restricted mean.

Each restricted mean estimator in this simulation study uses the "regression" approach described in Section 4.1.1. The different methods of estimating the restricted mean are, in essence, different approaches to estimating the conditional survival function in equation (4.3). The three versions of the stacked estimator are compared to a Cox proportional hazards model with first-order main effects (referred to as the 'Cox estimator'), and a Cox proportional hazards model with penalized splines (referred to as the 'Splines estimator'). The Cox estimator was proposed by Chen and Tsiatis [2001] and is the most common approach, while the Splines estimator is a straightforward extension of the Cox estimator that should be more robust in a variety of situations. Note that each

stacked estimator includes both the Cox and Splines estimators in the set of candidate survival models.

The methods are compared on the basis of point estimation and confidence interval performance. The quality of point estimation is measured by bias, i.e., $E\hat{\delta}(\tau) - \delta(\tau)$, and mean squared error (MSE), i.e., $E\{\hat{\delta}(\tau) - \delta(\tau)\}^2$. Confidence interval performance is assessed with three measures: average confidence interval length (ACL), coverage probability, and power. For each method, the estimated confidence intervals use the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution with 300 bootstrap replicates.

In the exponential scenario with linear covariate effects, the correctly specified Cox estimator should perform well, while the Splines estimator - despite being correctly specified - should be slightly less efficient due to increased flexibility. In contrast, for the exponential scenario with non-linear covariate effects, the Cox estimator should perform poorly due to model misspecification, while the Splines estimator remains correctly specified and should perform relatively well. Despite including both the Cox estimator and the Splines estimator, all three stacked estimators will likely perform relatively worse in both exponential scenarios due to the increased flexibility of additional models in the set of candidate survival models.

Each parametric and semi-parametric model is misspecified in both gamma scenarios. However, the parametric models in the stacked estimators closely approximate the truth in the linear scenario (e.g., same mean function). Thus, in the linear gamma scenario, the stacked estimators should perform better than both the Cox and Splines estimator due to approximately correct parametric models. The non-linear gamma scenario assesses the robustness of each estimator since the parametric and semi-parametric models may not closely approximate the underlying truth. The stacked estimators should perform better than the Cox and Splines estimators due to the increased flexibility of additional models in the set of candidate models. This is also a good opportunity for the stacked estimators with bagged survival trees [i.e., the Stacked ($n_{min} = 3$) and Stacked ($n_{min} = 10$) estimators] to perform well relative to the Stacked estimator without bagged survival trees.

Across the different scenarios (Tables 4.1 and 4.2), the Stacked estimator (without bagged survival trees) possesses as good, or better, bias and MSE than both the Stacked ($n_{min} = 3$) and Stacked ($n_{min} = 10$) estimators with bagged survival trees. The Stacked ($n_{min} = 3$) estimator also consistently fails to achieve nominal coverage, and possesses coverage levels as low 86%. In addition, the Stacked ($n_{min} = 3$) and Stacked ($n_{min} = 10$) estimators actually perform slightly worse than the Stacked estimator in the non-linear gamma scenario; the most advantageous scenario for bagged survival trees. Due to the similar, or better, performance and substantially less computational requirements, the

Stacked estimator (without bagged survival trees) is preferred to the Stacked estimators with bagged survival trees. As such, for the rest of this section, we only compare the Cox and Splines estimators to the Stacked estimator (without bagged survival trees).

For the exponential scenarios (Table 4.1), the Stacked estimator possesses slightly more bias than the Cox estimator when the covariate effects are linear, and the Stacked estimator has similar, or more, bias than the Splines estimators for both linear and non-linear covariate effect scenarios. These results are expected as the Cox and Splines estimators are correctly specified in, respectively, the linear and non-linear exponential scenarios. In the exponential scenario with non-linear covariate effects, the Stacked estimator has approximately $\sim 40\% - 85\%$ less absolute bias than the misspecified Cox estimator. In the linear scenario, the Stacked estimator has MSE similar to as both the Cox and Splines estimators, while in the non-linear scenario the Stacked estimator has $10\% - 15\%$ lower MSE than both the Cox and Splines estimators. In the non-linear scenario, the Stacked estimator also has approximately 5% shorter confidence intervals than both the Cox estimator and, surprisingly, the Splines estimator. The Stacked estimator is therefore more efficient than both the Cox and Splines estimators in the non-linear exponential scenario. All estimators maintain a coverage probability within 1% of the nominal coverage level.

For both gamma scenarios (Table 4.2), the Stacked estimator possesses less bias and $5\% - 10\%$ lower MSE than both the Cox and Splines estimators. For the linear gamma scenario, the Stacked and Cox estimators possess similar confidence interval length, while the Stacked estimator has $\sim 5\%$ shorter confidence intervals than the Splines estimator. For the non-linear gamma scenario, the Stacked estimator has approximately 5% shorter confidence intervals than both the Cox and Splines estimators, while maintaining nominal coverage. Thus, the Stacked estimator performs better than both the Cox and Splines estimators under model misspecification.

Chapter 3 motivated stacking with the goal of balancing, for a given sample size, the efficiency of (potentially misspecified) parametric models with robust (but potentially inefficient) semi-parametric and non-parametric models. This effect is illustrated in the simulation study here even when the bagged survival trees (i.e., the non-parametric estimator) are excluded from the set of candidate survival models. For example, in the linear exponential scenario, the Stacked estimator possesses MSE as good, or better, than the correctly specified Cox and Splines estimators. This is likely due to the inclusion of a correctly specified parametric Weibull model. Yet, even the Weibull model is misspecified in the non-linear gamma scenario, the Stacked estimator still performs better than both the Cox and Splines estimators despite the lack of a non-parametric estimator. This ability to adaptively find a good balance between the low variance

TABLE 4.1: Simulation results for the exponential distributed scenarios: $N = 300$, $N_{SIM} = 2500$, and a marginal censoring of 30%. The confidence intervals are estimated using the 2.5^{th} and 97.5^{th} quantiles of the non-parametric bootstrap distribution with 300 bootstrap replicates. 'ACL' is the average confidence interval length. The $\gamma(20)$ and $\gamma(50)$ are the true restricted mean treatment effects for $\tau = 20$ and $\tau = 50$, respectively.

	Estimator	Bias	MSE	ACL	Cov.	Power
Linear $\gamma(20) = -2.965$	Cox	-0.02	0.49	2.69	0.94	0.99
	Splines	-0.05	0.51	2.83	0.95	0.99
	Stacked	-0.09	0.50	2.74	0.94	0.99
	Stacked ($n_{min} = 3$)	-0.09	0.50	2.64	0.93	0.99
	Stacked ($n_{min} = 10$)	-0.09	0.50	2.72	0.95	0.99
Non-Linear $\gamma(20) = 2.690$	Cox	0.08	0.57	2.87	0.94	1.00
	Splines	0.01	0.53	2.87	0.95	1.00
	Stacked	0.01	0.48	2.73	0.95	1.00
	Stacked ($n_{min} = 3$)	-0.04	0.48	2.56	0.91	1.00
	Stacked ($n_{min} = 10$)	0.07	0.51	2.81	0.94	1.00
Linear $\gamma(50) = -12.318$	Cox	0.00	4.34	7.90	0.94	1.00
	Splines	-0.08	4.56	8.29	0.94	1.00
	Stacked	0.04	4.20	7.90	0.94	1.00
	Stacked ($n_{min} = 3$)	0.11	4.17	7.27	0.90	1.00
	Stacked ($n_{min} = 10$)	-0.01	4.23	7.91	0.93	1.00
Non-Linear $\gamma(50) = 7.929$	Cox	0.26	3.40	6.94	0.94	1.00
	Splines	0.06	3.19	6.92	0.95	1.00
	Stacked	0.15	2.89	6.57	0.94	1.00
	Stacked ($n_{min} = 3$)	-0.08	3.00	5.97	0.86	1.00
	Stacked ($n_{min} = 10$)	0.16	2.91	6.63	0.94	1.00

of (potentially misspecified) parametric models with the robustness of semi-parametric models (e.g., a Cox model with penalized splines) or non-parametric models is the most appealing aspect of stacked survival models.

A central concept of this chapter (and, in general, this thesis) is that better conditional survival function estimation is associated with better estimation of restricted mean treatment effects. However, the validity of this concept depends on the measure of restricted mean performance. In particular, the MSE of the restricted mean treatment effect is strongly correlated with the mean-squared error of the conditional survival function estimate, while the absolute bias has a surprisingly weak correlation (see Figure 4.1). It turns out that the MSE is bounded by the mean-squared error of the conditional survival function estimator (see Appendix C for details). This implies that the absolute bias is also bounded, but the bound is less tight due to a positive variance term. It is therefore not surprising that the absolute bias would be less strongly correlated with the performance of the conditional survival function.

TABLE 4.2: Simulation results for the gamma distributed scenarios: $N = 300$, $N_{SIM} = 2500$, and a marginal censoring of 30%. The confidence intervals are estimated using the 2.5^{th} and 97.5^{th} quantiles of the non-parametric bootstrap distribution with 300 bootstrap replicates. 'ACL' is the average confidence interval length. The $\gamma(20)$ and $\gamma(50)$ are the true restricted mean treatment effects for $\tau = 20$ and $\tau = 50$, respectively.

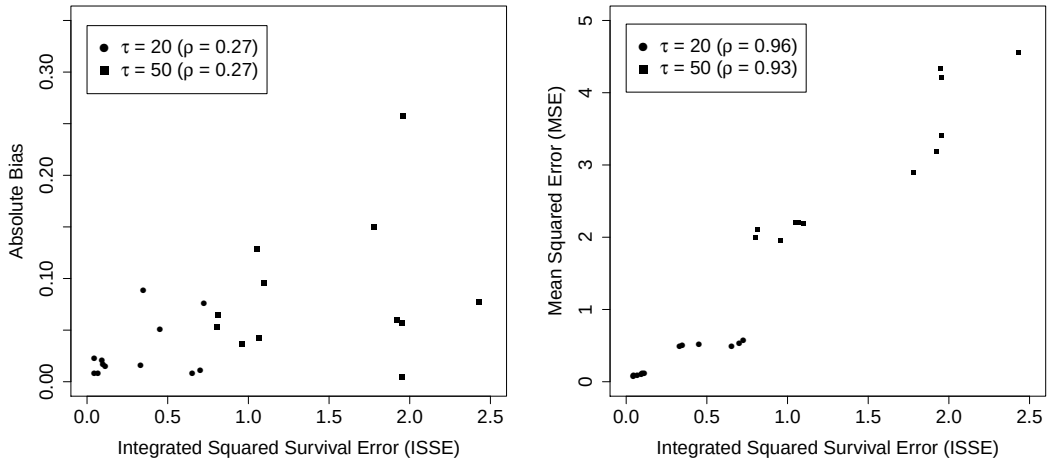
	Estimator	Bias	MSE	ACL	Cov.	Power
Linear $\gamma(20) = -0.753$	Cox	-0.02	0.09	1.18	0.94	0.74
	Splines	-0.01	0.10	1.24	0.95	0.69
	Stacked	0.01	0.08	1.13	0.95	0.75
	Stacked ($n_{min} = 3$)	0.00	0.08	1.19	0.93	0.76
	Stacked ($n_{min} = 10$)	0.00	0.08	1.17	0.94	0.76
Non-Linear $\gamma(20) = -0.931$	Cox	-0.02	0.12	1.34	0.94	0.78
	Splines	-0.02	0.12	1.40	0.95	0.75
	Stacked	0.02	0.10	1.29	0.95	0.79
	Stacked ($n_{min} = 3$)	0.00	0.11	1.36	0.92	0.81
	Stacked ($n_{min} = 10$)	-0.01	0.11	1.34	0.94	0.81
Linear $\gamma(50) = -6.599$	Cox	-0.06	2.10	5.60	0.94	0.99
	Splines	-0.04	2.20	5.91	0.95	0.99
	Stacked	0.04	2.02	5.61	0.95	1.00
	Stacked ($n_{min} = 3$)	0.04	2.01	5.46	0.95	1.00
	Stacked ($n_{min} = 10$)	0.00	2.02	5.70	0.95	1.00
Non-Linear $\gamma(50) = -6.407$	Cox	-0.13	2.20	5.72	0.94	0.99
	Splines	-0.10	2.19	5.83	0.96	0.99
	Stacked	0.04	1.95	5.56	0.95	0.99
	Stacked ($n_{min} = 3$)	0.05	1.98	5.36	0.92	1.00
	Stacked ($n_{min} = 10$)	-0.13	2.08	5.70	0.94	1.00

4.3 Effect of Center Volume in Lung Transplantation

We applied the proposed estimator to data from an observational registry of post-lung transplant survival to estimate the effect of large center volume. In particular, we want to estimate the difference in the restricted mean of post-transplant survival between high volume centers [defined as > 100 lung transplants over the past two years [Tsuang et al., 2013]] and low volume centers (defined as ≤ 100 lung transplants over the past two years). However, previous research has demonstrated that transplant type, which is an important confounder, likely possesses non-proportional hazards after surgery [Weiss et al., 2009]. Therefore, this example represents an ideal setting for estimating restricted mean treatment effects with stacked survival models.

The United Network for Organ Sharing (UNOS) collects patient information, donor information and survival status of every solid organ transplant performed in the United States. This analysis only includes lung transplants performed between January 1, 2008

FIGURE 4.1: An investigation into the relationship between restricted mean treatment effect performance and the quality of the conditional survival function estimation [excluding the Stacked ($n_{min} = 3$) and Stacked ($n_{min} = 10$) estimators], which is measured by the mean of the treatment-specific integrated squared survival errors, or $\text{ISSE}(a) = E_{\boldsymbol{x}}\{\int_0^\tau [\hat{S}^{(a)}(t|\boldsymbol{x}) - S_o^{(a)}(t|\boldsymbol{x})]^2 dt\}$. The legend contains the restricted mean of interest and Spearman's ρ in parentheses.



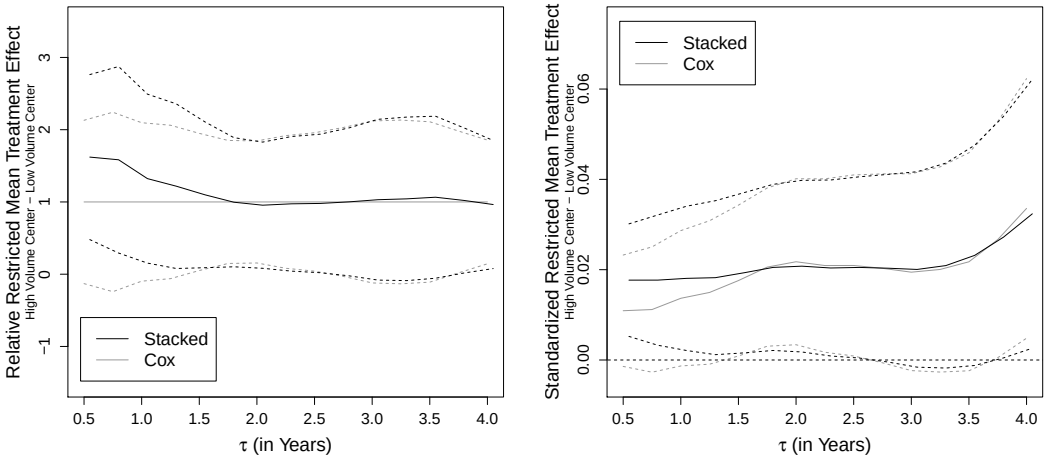
and December 31, 2011 in adult recipients receiving their first lung-only transplantation. We adjust for potential confounding from several patient related covariates including gender, age, lung allocation score, native disease grouping (obstructive, vascular, cystic and restrictive), distance walked in six minutes, ventilator use, level of oxygen use, and type of transplant (single versus bilateral lung transplant). We also adjust for several donor related covariates: age over 55 years, African American race, smoking history greater than 20 pack years, and height difference between donor and recipient. The event of interest is time to death or retransplantation; known as graft survival. A total of $5,499$ transplanted patients were included in this analysis. Approximately 76% of the survival times were censored.

Similar to Section 4.2, the stacked survival model includes two versions of the Cox model, a Weibull model, and a log-Normal model. The Weibull, log-Normal, and the first Cox model fit only linear main effects to all continuous covariates except the height difference between donor and recipient, which also has a quadratic term. The second Cox model fits penalized splines to each continuous covariate. The 95% confidence intervals are estimated from the $2.5th$ and $97.5th$ percentiles of the bootstrap distribution with 1000 bootstrap replications.

Figure 4.2 gives the estimated restricted mean treatment effect for high volume versus low volume centers from $\tau = 0.5$ years to $\tau = 4$ years. Both the Stacked and Cox estimators consistently estimate a restricted mean difference between large and small volume centers greater than zero, which indicates better post-transplant survival for

high volume centers. The Stacked estimate is approximately $30 - 60\%$ larger than the Cox estimate from 0.5 years to 1 years, while at 4 years the Stacked estimate is approximately 5% smaller than the Cox estimate. In addition, the confidence intervals for the Stacked estimate do not include zero until about 3.5 years, while the Cox estimate is not significant until about 1.5 years and becomes non-significant from 2.5 to 3.5 years. This is due, in part, to the Stacked confidence interval being approximately 5% shorter than the Cox confidence interval for the first two years. As illustrated in the simulation scenario, the difference in significance between the two estimators may be a function of the shorter confidence intervals and consistently lower MSE of the Stacked estimate in non-proportional hazards scenarios.

FIGURE 4.2: The estimated difference in restricted means between high volume centers and low volume centers across a range of τ vaues. The left graph plots the restricted mean treatment effect for the Stacked estimator relative to the Cox Estimator, while the right graph standardizes the restricted mean treatment effect by τ (i.e., $\hat{\gamma}(\tau)/\tau$). The dashed lines are the 95% confidence interval limits. The confidence intervals are based on the 2.5^{th} and 97.5^{th} quantiles of a non-parametric bootstrap with 1000 bootstrap replications. The Stacked estimator does *not* include random survival forests.



Weiss et al. [2009] were specifically interested in the mortality risk for center volume at one year as it is a potential indicator of peri-operative and early post-operative mortality (which is relatively high). The Stacked estimator shows a 30% larger difference than the Cox estimator in average survival over one year. In addition, the confidence interval for the Cox estimate includes zero, while the Stacked estimator does not include zero. The lack of significance for the Cox estimator after one-year is surprising due to sustained evidence in the transplantation literature that center volume is significantly associated with mortality [Thabut et al., 2010, Weiss et al., 2009].

4.4 Summary

We propose flexibly estimating restricted mean treatment effects from observational studies with stacked survival models. Restricted mean survival is traditionally estimated by first estimating the conditional survival distribution with a Cox proportional hazards model, yet the simulation study illustrates that stacked survival models can improve the mean squared error (MSE) even under proportional hazards. When the proportional hazards assumption is violated, stacked survival models can reduce the bias and variance of the restricted mean treatment effect. We also demonstrated that the Stacked estimator identifies a statistically significant difference in lung transplantation between high volume and low volume centers for the one year restricted mean, while the proportional hazards estimator fails to identify the difference.

We considered three different approaches to stacked survival models: one estimator (the 'Stacked' estimator) excluded the non-parametric model from the set of candidate models, and two estimators [the 'Stacked ($n_{min} = 3$)' and 'Stacked ($n_{min} = 10$)' estimators] included bagged survival trees (i.e., the non-parametric model) in the set of candidate survival models. We found that the Stacked estimator (without bagged survival trees) performed as good, or better, than the stacked estimators with bagged survival trees. This illustrates that additional candidate survival models do not necessarily improve the performance of the stacked survival model for estimating restricted mean treatment effects. Thus, the impact of the set of candidate survival models deserves further characterization in a wider range of scenarios.

We note that in estimating the conditional survival distribution, the method presented here estimates a separate model for each treatment. This flexible approach allows all covariate effects to vary between treatments, i.e., an implicit assumption of treatment by covariate interactions. If there are no treatment by covariate interactions, then the method is less efficient than a model that assumes no interactions [see Chen and Tsiatis [2001] for a discussion].

There are two main approaches to estimating the casual restricted mean treatment effect: the "regression approach" pursued in this chapter and an approach based on inverse-probability weighting (IPW) for treatment assignment and censoring [Schaubel and Wei, 2011, Zhang and Schaubel, 2012]. The IPW approach requires forming models for the censoring and treatment distributions. The "regression" approach is a more efficient estimator of the restricted mean difference when the conditional survival model has been correctly specified. However, the IPW approach is sometimes preferred because standard methods to estimating the conditional survival distribution may be overly restrictive (e.g., the Cox proportional hazards model). The flexibility of stacked survival models

may mitigate some of the concerns of the "regression" approach. We note, though, the difference between misspecified censoring/treatment models and a misspecified survival time model (i.e., the regression approach) is unclear, and these topics deserve further attention.

The Cox proportional hazards model is mathematically tractable, which allows intuitive extensions to situations more difficult than right censoring. These extensions are easily adapted to estimate restricted mean treatment effects. For example, Zhang and Schaubel [2011] estimate restricted mean treatment effects under dependent censoring using a Cox proportional hazards model with inverse-probability-of-censoring weights. Pan and Gastwirth [2013] also use the proportional hazards framework to estimate restricted mean treatment effects under semi-competing risks. These extensions are generally more complicated outside of a proportional hazards framework. A future research interest is extending stacked survival models to situations such as dependent censoring or competing risks. This could improve restricted mean estimation *and* survival prediction in such situations.

This chapter focuses on estimating restricted mean treatment effects. Yet statisticians regularly want to form linear contrasts for summary measures of interest. A common approach to restricted mean regression uses pseudo-observations from the leave-one-out jackknife of the Kaplan-Meier restricted mean estimator as the outcome variable in a generalized linear model [Andersen et al., 2004, 2003]. The potential benefit for estimating pseudo-observations with stacked survival models is not clear. However, an alternative to pseudo-values is the model-free contrast approach proposed by Rudser et al. [2012], which would likely benefit from stacked survival models. The investigation into these areas is a future research interest.

There has been recent research on combining information from relevant covariates in an outcome model (i.e., the model for the survival time random variable) with the relevant covariates in an exposure model (i.e., the model for treatment assignment) [Cefalu et al., 2013, Wang et al., 2012]. Specifically, the efficiency of the outcome model is improved by selecting covariates based, in part, on the covariates included in the treatment model. The idea exploits the fact that confounders have to be associated with both the treatment and the exposure. A similar approach to covariate selection may improve the performance of stacked survival models. Although, a major advantage of stacking is the ability to gain efficiency by considering different distributional assumptions and functional forms. As such, an interesting avenue for future research would consider the selection of covariates based on both the outcome and treatment models, while also considering different distributional assumptions and functional forms for the outcome model.

A central concept of this chapter is that a better estimator of the conditional survival function should be associated with a better estimator of the restricted mean treatment effect. Yet a conditional survival function is required by many methods besides restricted mean treatment effects: for example, censored quantile regression [Chapter 2], time-dependent ROC curves [Zheng and Heagerty, 2004], inverse probability-of-censoring weighted estimators [Fine and Gray, 1999], model-free contrast approaches [Rudser et al., 2012], and dynamic treatment regime methods [Zhao et al., 2011]. Thus, a better estimator of a conditional survival function may be able to improve the estimation of many survival analysis methods. The characterization of this connection, both formally and empirically, is a future research interest.

Chapter 5

Conclusion

Due to the difficult interpretation of the hazard ratio, this thesis explored potential alternatives to the Cox proportional hazards model. We specifically considered censored quantile regression and restricted mean treatment effects. However, both censored quantile regression and restricted mean treatment effects depend on the estimation of a conditional survival function, which acts as a nuisance parameter. Thus, we aimed to improve the estimation of censored quantile regression and restricted mean treatment effects by improving the estimation of the conditional survival function.

Chapter 2 applied bagged survival trees (i.e., a non-parametric estimator) to the estimation of censored quantile regression. When the underlying assumptions of current methods were badly violated, bagged survival trees improved the estimation of censored quantile regression. However, when the underlying assumption were correct, bagged survival trees performed relatively worse due to the inefficiency of non-parametric estimators. This illustrates the need for an estimator of the conditional survival function that performs well in a wide variety of situations.

Chapter 3 proposed stacking survival models to effectively estimate conditional survival functions in a wide range of situations. In particular, stacked survival models can combine parametric, semi-parametric and non-parametric estimators of the conditional survival function. In this manner, stacking can exploit the low variance of approximately correct parametric models, while maintaining the robustness of non-parametric models. This allows stacked survival models to perform better than the single model chosen through cross-validation and, at times, better than every model in the set of candidate survival models. This is a promising result for survival methods that require an estimate of the conditional survival function.

For example, Appendix C showed that the mean-squared error of the restricted mean treatment effect is bounded by the mean squared error of the conditional survival function. It is therefore not surprising that the estimator of the restricted mean treatment effect based on a stacked survival model consistently achieved mean-squared error as good, or better, than estimators based on proportional hazards; interestingly, this relationship was observed even in proportional hazards scenarios. This is a promising development as numerous survival analysis methods that require an estimator of the conditional survival function rely on a Cox proportional hazards model.

Stacked survival models were found to have, at times, interesting problems within inference settings. First, in Chapter 4, confidence intervals for the stacked survival models with bagged survival trees occasionally failed to maintain nominal coverage. Preliminary investigations suggest that a biased bootstrap distribution may play a role. Bias-corrected and accelerated bootstrap confidence intervals is one approach that could improve confidence interval performance [Efron, 1987, Efron and Tibshirani, 1993]. The second issue faced by stacked survival models is fitting failures for parametric models during bootstrap resampling. In particular, the parametric models fail to converge for bootstrapped data sets, which is likely due to tied event times. A potential solution for this problem is randomly perturbing the observed times (e.g., multiplying by a random exponential). This would avoid tied event times and may provide similar performance as the bootstrap [Jin et al., 2001].

There are some potential extensions to stacked survival models. The most interesting extension is time-dependent stacking on the hazard function. Unlike time-dependent stacking on the survival function, this maintains the fundamental properties of a conditional survival function. Covariate-dependent stacking is another interesting avenue. However, this approach would require careful regularization of the weights (e.g., adaptively turning off certain covariate dependence). Otherwise, the advantages of increased flexibility might be lost due to an increased variance (i.e., the bias-variance trade-off). As illustrated in Chapters 2 and 4, these potential improvements in stacked survival models could improve survival prediction *and* survival analysis methods.

Chapters 2, 3, and 4 each mentioned a variation of tree-based survival forests. In particular, Chapters 2 and 4 used bagged survival trees, while Chapter 3 used random survival forests. In addition, each chapter mentioned that tree-based survival forests required the appropriate selection of tuning parameters to perform well. Stacking across tuning parameters is an interesting approach to this problem. However, tree-based survival forests with different tuning parameter combinations may not capture fundamentally different aspects of the conditional survival function and, as shown in Chapter 3, stacking is less effective when the candidate survival functions are highly correlated. This

suggests that cross-validation may work better than stacking across tuning parameters. Regardless, a comprehensive investigation into the effect of tuning parameters would increase accessibility of tree-based survival forests.

Bibliography

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* **10,** 335–350.

Andersen, P. K., Klein, J. P., and Rosthøj, S. (2003). Generalised linear models for correlated pseudoobservations, with applications to multistate models. *Biometrika* **90,** 15–27.

Binder, H. (2013). *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*. R package version 1.4.

Boonstra, P. S., Taylor, J. M. G., and Mukherjee, B. (2013). Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics* **14,** 259–272.

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

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

Breiman, L. (1996b). Stacked regressions. *Machine Learning* **24,** 49–64.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

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

Butler, J. H., Gilpin, E. A., Gordon, L., and Olshen, R. A. (1989). Tree-structured survival analysis, ii. Technical Report 133, Division of Biostatistics, Stanford University, Stanford University.

Cefalu, M., Dominici, F., and Parmigiani, G. (2013). Model averaged double robust estimation. Technical Report 149, Division of Biostatistics, Harvard University, Harvard University.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* **57,** 1030–1038.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71,** 1591–1608.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health.*

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82,** 171–185.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* Chapman and Hall.

Fan, J. and Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society. Series B* **62,** 303–322.

Ferguson, T. S. (1996). *A course in large sample theory.* Chapman and Hall/CRC.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94,** 496–509.

Fitzmaurice, G. M., Lipsitz, S. R., Arriage, A., Sinha, D., Greenberg, C., and Gawande, A. A. (2014). Almost efficient estimation of relative risk regression. *Biostatistics* .

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley.

Fumera, G. and Roli, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27,** 942–956.

Goeman, J. J. (2012). *Penalized R Package.* R package version 0.9-42.

Gordon, L. and Olshen, R. A. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis* **15,** 147–163.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18,** 2529–2545.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley.

Ishwaran, H. and Kogalur, U. (2013). *Random Survival Forests*. R package version 3.6.4.

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

Jin, Z., Ying, Z., and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88,** 381–390.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association* **82,** 1169–1176.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2011). *quantreg: Quantile Regression*. R package version 4.69.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46,** 33–50.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90,** 78–94.

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

LeBlanc, M. and Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association* **91,** 1641–1650.

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* **52,** 721–725.

Lostritto, K., Strawderman, R. L., and Molinaro, A. M. (2012). A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Biometrics* **68,** 1146–1156.

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

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *The Handbook of Econometrics*, volume 4. Amsterdam: North-Holland.

Pan, Q. and Gastwirth, J. L. (2013). Estimating restricted mean job tenures in semi-competing risk data compensating victims of discrmination. *Annals of Applied Statistics* **7,** 1474–1496.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association* **103,** 637–649.

Polley, E. C. and Van der Laan, M. (2011). Super learning for right-censored data. In *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association* **98,** 1001–1012.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics* **32,** 143–155.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Royston, P. and Parmar, M. K. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology* **13,** 1–15.

Rudser, K. D., LeBlanc, M. L., and Emerson, S. S. (2012). Distribution-free inference on contrasts of arbitrary summary measures of survival. *Statistics in Medicine* **31,** 1722–1737.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Schaubel, D. E. and Wei, G. (2011). Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics* **67,** 29–38.

Sill, J., Takacs, G., Mackey, L., and Lin, D. (2009). Feature-weighted linear stacking. *Arxiv* .

Thabut, G., Christie, J. D., Kremers, W. K., Fournier, M., and Halpern, S. D. (2010). Survival differences following lung transplantation among us transplant centers. *Journal of the American Medical Association* **304,** 53–60.

Therneau, T. (2013). *Survival analysis, including penalized likelihood*. R package version 2.37-4.

Tsuang, W. M., Vock, D. M., Copeland, C. A. F., and Lederer, D. J. (2013). An acute change in lung allocation score and survival after lung transplantation: a cohort study. *Annals of Internal Medicine* **158,** 650–657.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6,**.

Varadhan, R. (2012). *alabama: Constrained nonlinear optimization.* R package version 2011.9-1.

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68,** 661–686.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association* **103,** 1117–1128.

Weiss, E. S., Allen, J. G., Meguid, R. A., Patel, N. D., Merlo, C. A., Orens, J. B., Baumgartner, W. A., Conte, J. V., and Shah, A. S. (2009). The impact of center volume on survival in lung transplantation: An analysis of more than 10,000 cases. *The Annals of Thoracic Surgery* **88,** 1062–1070.

Wolpert, D. H. (1992). Stacked generalization. *Neural Network* **5,** 241–259.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall.

Yang, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association* **94,** 137–145.

Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association* **90,** 178–184.

Zhang, M. and Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics* **67,** 740–749.

Zhang, M. and Schaubel, D. E. (2012). Double-robust semiparametric estimator for differences in restricted mean lifetimes in observations studies. *Biometrics* **68,** 999–1009.

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,** 1422–1433.

Zheng, Y. and Heagerty, P. (2004). Semiparametric estimation of time-dependent roc curves for longitudinal marker data. *Biostatistics* **5,** 615–632.

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association* **107,** 331–340.

Zucker, D. M. (1998). Restricted mean life with covariates: Modification and extension of a useful survival analysis method. *Journal of the American Statistical Association* **83,** 702–709.

Appendix A

Censored Quantile Regression

This appendix shows that recursive partitioning based weights lead to consistent estimation of regression quantiles. We first introduce some notation and regularity conditions, then show that survival trees are uniformly consistent for conditional survival functions on a certain support.

In order to clearly state the regularity conditions, some concepts from the tree literature are introduced. Consider the partition $Q^{(n)}$ of the covariate space, i.e., as produced by a tree, then $B_k^{(n)}$ is the k^{th} *box*, or terminal node, of $Q^{(n)}$ such that $\bigcup_k B_k^{(n)} = Q^{(n)}$. Now define the *mesh*, or diameter, of the box k as

$$D_n(k) = \sup\{||y - z||, \text{ such that y, z } \epsilon\ B_k^{(n)}\}, \tag{A.1}$$

where it is assumed that $B_k^{(n)}$ is contained within the support of $\boldsymbol{x}$ for all k. Define $\hat{F}(t|B_k^{(n)})$ as the within terminal node cumulative distribution estimator for all $\boldsymbol{x}\ \epsilon\ B_k^{(n)}$. We adopt the following conditions:

A1. For $\beta(\tau)$ in the neighborhood of $\boldsymbol{\beta}_o(\tau)$, $E[\boldsymbol{x}\boldsymbol{x}^T f_T(\boldsymbol{x}^T\boldsymbol{\beta}(\tau)|\boldsymbol{x})\{1 - F_C(\boldsymbol{x}^T\boldsymbol{\beta}(\tau)|\boldsymbol{x})\}]$ is positive definite.

A2. There exists a constant $K_{\boldsymbol{x}}$ such that $E[||\boldsymbol{x}||^3] \leq K_{\boldsymbol{x}}$. In addition, $\max_{1\leq i\leq n} ||\boldsymbol{x}_i|| = O_p(n^{1/2}(\log n)^{-1})$, and $E[\boldsymbol{x}\boldsymbol{x}^T]$ is a positive definite $p \times p$ matrix.

A3. The conditional distribution functions $F_T(t|\boldsymbol{x})$ and $F_C(t|\boldsymbol{x})$ have first derivatives with respect to t, $f_T(t|\boldsymbol{x})$ and $f_C(t|\boldsymbol{x})$ respectively, which are uniformly bounded away from infinity. Also, $F_T(t|\boldsymbol{x})$ and $F_C(t|\boldsymbol{x})$ have bounded (uniformly in t) second-order partial derivatives with respect to $\boldsymbol{x}$.

A4. $E[|T|^r] < \infty$ for some $r > 1$ and $Q^{(n+1)}$ is a refinement of $Q^{(n)}$. Let $k(n)$ be a nondecreasing sequence of integers which approaches infinity, $n^{1/r}\log n/k(n)$,

$D_n(\vec{x})$, and $I_{\hat{H}_n(B_k^{(n)})<k(n)/n}$ approach 0 as $n \to \infty$, where $\hat{H}_n(B_k^{(n)})$ is the empirical probability of box k on the sample.

A5. Let $\xi > 0$ and for any fixed $t < \xi$, $\hat{F}(t|B_k^{(n)}) \to F(t|B_k)$, in probability.

Assumptions A1 and A2 are common to the censored quantile regression literature, see for example Wang and Wang [2009]. Assumptions A3, A4 and A5 imply that the conditional distribution function converges uniformly in t. Assumption A4 ensures the conditions for theorems needed from Gordon and Olshen [1984] are satisfied. Assumption A5 requires that the within terminal node cumulative distribution estimator is pointwise consistent on a subset of the support, where ξ is typically chosen to ensure a well defined survival distribution.

In Chen et al. [2003], "*Theorem 1*" states five sufficient conditions for consistency. Condition (1.3) is satisfied by assumptions while conditions (1.1), (1.2) and (1.5') are satisfied by identical arguments used by Wang and Wang [2009]. Condition (1.4) requires us to show that survival tree estimators, denoted $\hat{F}_T(t|\boldsymbol{x})$, that use a general within node cumulative distribution estimator, say $\hat{F}(t|B^{(n)})$, is uniformly consistent for $F_T(t|\boldsymbol{x})$. Define the following quantities as

$$\begin{aligned} F_Y(t|\boldsymbol{x}) &= P_Y(Y \leq t|\boldsymbol{x}) = 1 - (1 - F_T(t|\boldsymbol{x}))(1 - F_C(t|\boldsymbol{x})), \\ F_{T1}(t|\boldsymbol{x}) &= P_{T1}(T \leq t, \delta = 1|\boldsymbol{x}) = \int_0^t (1 - F_C(u|\boldsymbol{x}))f_T(u|\boldsymbol{x})du, \end{aligned}$$

which can be thought as the observed distribution function and observed event distribution function, respectively. Consider the survival tree partitioning the covariate space into a set of 'boxes' denoted by B_N. Now define the distribution estimators for $F_Y(t|\boldsymbol{x})$ and $F_{T1}(t|\boldsymbol{x})$ as $\hat{F}_Y(t|\boldsymbol{x})$ and $\hat{F}_{T1}(t|\boldsymbol{x})$, respectively.

For fixed $t > 0$, define two types of convergence as described in Gordon and Olshen [1984],

(i) $E\{|\hat{F}_n(t|\boldsymbol{x}, \zeta_n) - F(t|\boldsymbol{x})||\zeta_n\} \to 0$, and

(ii) $\hat{F}_n(t|\boldsymbol{x}, \zeta_n) - F(t|\boldsymbol{x}) \to 0$, almost surely,

where ζ_n is the observed data, and the set $\Omega = \{\xi : F_Y(\xi|\boldsymbol{x}) < 1 - \delta\}$ for $\delta > 0$. Now we can concisely state the following lemma,

Lemma A.1. *Let the conditional distribution functions $F_{T1}(t|\boldsymbol{x})$ and $F_C(t|\boldsymbol{x})$ be continuous. For all $s \leq \xi$, where $\xi > 0$ is a fixed time, $\hat{F}_Y(s|\boldsymbol{x})$ and $\hat{F}_{T1}(s|\boldsymbol{x})$ are recursive*

partitioning based estimators that are type (i) and/or (ii) consistent tree estimators for $F_Y(s|\boldsymbol{x})$ *and* $F_{T1}(s|\boldsymbol{x})$*, respectively, for each single* s*. Then*

$$\int_0^{\xi} \frac{d\hat{F}_{T1}(s|\boldsymbol{x})}{1-\hat{F}_Y(s|\boldsymbol{x})} \quad \rightarrow \quad -\log\{1-F_T(\xi|\boldsymbol{x})\},$$

almost surely on Ω*.*

This lemma is *Theorem 1* from Butler et al. [1989]. Assumptions (A4) and (A5) combined with either *Theorem 3.6* or *Theorem 4.1* from Gordon and Olshen [1984] satisfy the conditions for *Lemma 1.1*. Now define "*Lemma 1*" from Breslow and Crowley [1974],

Lemma A.2. *Let* $N(t) = \Sigma_{i=1}^{N} I[Y_i \geq t]$ *be the number of individuals still "at risk" at time* t*. Then with probability 1, for all* $0 < t < max_{1\leq i\leq N} Y_i$*,*

$$0 < -\log\{1-\hat{F}_T(t|\boldsymbol{x})\} \quad - \quad \int_0^{\tau} \frac{d\hat{F}_{T1}(s|\boldsymbol{x})}{1-\hat{F}_Y(s|\boldsymbol{x})} < \frac{N-N(t)}{N\cdot N(t)}.$$

Lemma 1.1 and *Lemma 1.2* imply that for all $t \,\epsilon\, \Omega$,

$$-\log\{1-\hat{F}_T(t|\boldsymbol{x})\} \quad \rightarrow \quad -\log\{1-F_T(t|\boldsymbol{x})\},$$

as $N(t) \rightarrow \infty$. This means that $\hat{F}_T(t|\boldsymbol{x}) \rightarrow F_T(t|\boldsymbol{x})$ at all the continuity points of F_T on Ω. Thus, by definition,

$$\hat{F}_T(t|\boldsymbol{x}) \quad \rightarrow_D \quad F_T(t|\boldsymbol{x}),$$

in distribution. By problem 1.6 of Ferguson [1996], convergence in law with previously stated conditions implies uniform convergence which satisfies condition (1.4) of Chen et al. [2003]. This completes the proof for consistency.

Appendix B

Stacking Survival Models

Section B.1 presents the derivation of the mean-squared error decomposition presented in Section 3.2. In addition, a simple example illustrates the effect of stacking parametric and non-parametric survival models. Section B.2 proves the asymptotic properties of stacked survival models presented in Section 3.3. Section B.3 discusses time-dependent stacking and compares the performance to time-independent stacking [equation (3.4) in Section 3.1.2].

B.1 Mean-Squared Error Decomposition

We want to define certain quantities and make a connection to the Brier Score before showing the derivation of the mean-squared error decomposition presented in Section 3 of the main paper. We define the mean-squared error for a conditional survival function estimator as the integral of the squared error at time t over $\Omega = (0, \tau)$: $\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} = E\{E_{\boldsymbol{x}} \int_0^\tau [\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt\}$, where the outer expectation is taken with respect to the estimator. As mentioned in Section 3.2, the mean-squared error has a direct connection to the Brier Score. In particular,

$$
\begin{aligned}
E\left\{E_{\boldsymbol{x}} \int_0^\tau BS(t)dt\right\} &= E_{\boldsymbol{x}} \int_0^\tau E\left[\frac{\Delta(t)}{G(T(t)|\boldsymbol{x})} \times \{Z(t) - \hat{S}(t|\boldsymbol{x})\}^2\right] \\
&= E_{\boldsymbol{x}} \int_0^\tau E\left[\{Z(t) - \hat{S}(t|\boldsymbol{x})\}^2 dt\right], \text{ by iterative expectation} \\
&= E\left\{E_{\boldsymbol{x}} \int_0^\tau \{Z(t) - S_o(t|\boldsymbol{x}) + S_o(t|\boldsymbol{x}) - \hat{S}(t|\boldsymbol{x})\}^2 dt\right\} \\
&= EE_{\boldsymbol{x}} \int_0^\tau \left\{[Z(t) - S_o(t|\boldsymbol{x})]^2 + [S_o(t|\boldsymbol{x}) - \hat{S}(t|\boldsymbol{x})]^2\right\} dt \\
&= \sigma^2 + \text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\},
\end{aligned}
$$

where σ^2 is irreducible prediction error.

We define the bias and variance of a conditional survival function estimator at time t as, respectively, $\text{Bias}\{\hat{S}(t|\boldsymbol{x})\} = E\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})$ and $\text{Var}\{\hat{S}(t|\boldsymbol{x})\} = E[\hat{S}(t|\boldsymbol{x}) - E\hat{S}(t|\boldsymbol{x})]^2$, where these expectations are not taken with respect to the covariate space. The mean squared error of the stacked estimator is then decomposed as

$$
\begin{aligned}
\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} &= E\{E_{\boldsymbol{x}} \int_0^\tau [\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt\} \\
&= E\{E_{\boldsymbol{x}} \int_0^\tau [\hat{S}(t|\boldsymbol{x}) - E\hat{S}(t|\boldsymbol{x}) + E\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt\} \\
&= E_{\boldsymbol{x}} \int_0^\tau E\{[\hat{S}(t|\boldsymbol{x}) - E\hat{S}(t|\boldsymbol{x})]^2 + [E\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2\} dt \\
&= E_{\boldsymbol{x}} \int_0^\tau E\{[\sum_{k=1}^m \alpha_k \{\hat{S}_k(t|\boldsymbol{x}) - E\hat{S}_k(t|\boldsymbol{x})\}]^2 + \\
&\qquad [\sum_{k=1}^m \alpha_k \{E\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})\}]^2\} dt, \text{ because } \sum_{k=1}^m \alpha_k = 1 \\
&= E_{\boldsymbol{x}} \int_0^\tau \sum_{k=1}^m \sum_{l=1}^m \alpha_k \alpha_l \Big\{ E\{[\hat{S}_k(t|\boldsymbol{x}) - E\hat{S}_k(t|\boldsymbol{x})][\hat{S}_l(t|\boldsymbol{x}) - E\hat{S}_l(t|\boldsymbol{x})]\} + \\
&\qquad [E\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})][E\hat{S}_l(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})] \Big\} \\
&= \sum_{k=1}^m \alpha_k^2 E_{\boldsymbol{x}} \int_0^\tau [\text{Bias}^2\{\hat{S}_k(t|\boldsymbol{x})\} + \text{Var}\{\hat{S}_k(t|\boldsymbol{x})\}] dt + \\
&\qquad E_{\boldsymbol{x}} \sum_{k=1}^m \sum_{l \neq k} \alpha_k \alpha_l \int_0^\tau [\text{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} \times \text{Bias}\{\hat{S}_l(t|\boldsymbol{x})\} + \\
&\qquad \text{Cov}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\}] dt \\
&= \sum_{k=1}^m \alpha_k^2 \text{MSE}_\tau\{\hat{S}_k(\cdot|\boldsymbol{x})\} + \\
&\qquad E_{\boldsymbol{x}} \sum_{k=1}^m \sum_{l \neq k} \alpha_k \alpha_l \int_0^\tau [\text{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} \times \text{Bias}\{\hat{S}_l(t|\boldsymbol{x})\} + \\
&\qquad \text{Corr}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\} \times \text{Var}\{\hat{S}_k(t|\boldsymbol{x})\}^{\frac{1}{2}} \times \text{Var}\{\hat{S}_l(t|\boldsymbol{x})\}^{\frac{1}{2}}] dt.
\end{aligned}
$$

As specifically outlined in Section 3.2, this decomposition motivates stacking a diverse set of survival models in order to lower the correlation between predicted survival curves.

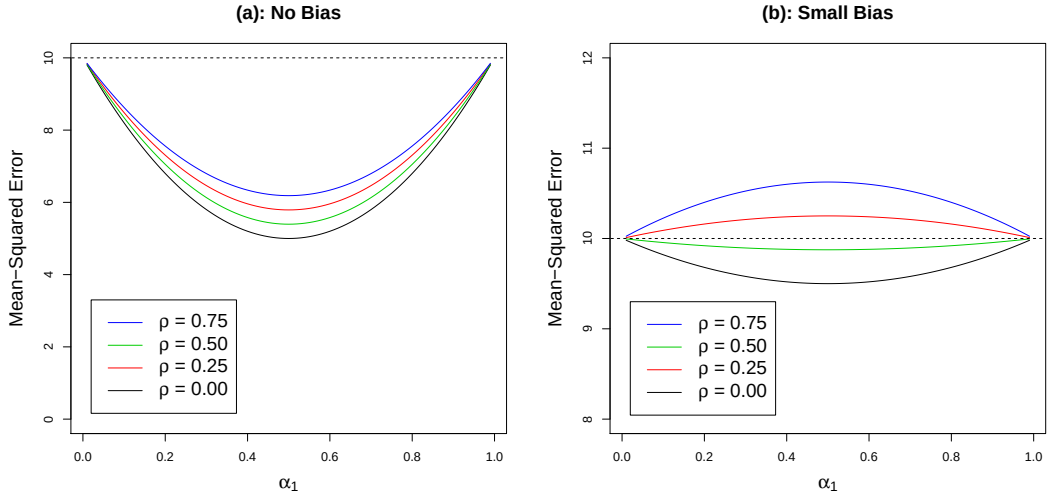
We now consider a simple example to illustrate the potential of stacking a survival model with low bias but high variance and a badly misspecified parametric model (i.e., high bias; low variance). For example, consider a set of two independent candidate survival models where both survival models possess the same mean-squared error, say $\text{MSE}_\tau\{\hat{S}_k(\cdot|\boldsymbol{x})\} = 10$ for $k = 1, 2$. For the sake of simplicity, the bias and variance of both survival models are constant across time and the covariate space.

We set the bias of the first model to $\int_0^\tau \text{Bias}^2\{\hat{S}_1(t|\boldsymbol{x})\}dt = 9$

and consider two different values of bias for the second model: $\text{Bias}\{\hat{S}_2(t|\boldsymbol{x})\} = 0$ and $\int_0^\tau \text{Bias}^2\{\hat{S}_2(t|\boldsymbol{x})\}dt = 1$. These fully define the operating characteristics of both survival models.

The first model corresponds to a badly misspecified parametric model, while the second model corresponds to a non-parametric estimator (first with no bias, then with some small sample bias). For this simple scenario, Figure B.1 illustrates that the stacked estimator always achieves lower MSE than the individual models when the non-parametric estimator possesses no bias (left plot). However, this advantage can decrease substantially when the non-parametric estimator is slightly biased depending on the degree of correlation (right plot). In addition, the effectiveness of the stacked estimator in both situations decreases when the correlation increases between the survival models.

FIGURE B.1: The example of stacking a misspecified, but efficient, parametric model and a low biased, but highly variable, non-parametric model (see Section B.1). Figure (a) shows a non-parametric estimator with no bias, while Figure (b) shows a non-parametric estimator with a small amount of bias. Note that the effectiveness of stacking decreases as the correlation (ρ) increases. The mean-squared error of both candidate survival models is 10, which is represented by the dotted line.



B.2 Asymptotic Properties

For both proofs, we consider the general case where l of the m estimators for the stacking procedure are uniformly consistent. Without loss of generality, let $\hat{S}_k(t|\boldsymbol{x})$ be uniformly consistent for $k = 1, ..., l$, i.e., the first l estimators are correctly specified. Recall that the conditional survival function is estimated on $\Omega = (0, \tau)$. Now assume the following conditions throughout Appendix B.

B1. There exists l estimators that are uniformly consistent within the set of estimators used for the stacking procedure. Without loss of generality, let these estimators be $\hat{S}_k(t|\boldsymbol{x})$ for $k = 1, ..., l$ and, by uniform consistency, $\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| \to 0$ for $k = 1, ..., l$. Additionally, for $k = l+1, ..., m$, $\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| \to c_k$ where $c_k > 0$.

B2. The estimator for the censoring distribution is uniformly consistent: $\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{G}(t|\boldsymbol{x}) - G(t|\boldsymbol{x})| \to 0$, and there exists a $\delta > 0$ such that $G(\tau|\boldsymbol{x}) > \delta$ for all $\boldsymbol{x}$.

B3. Let $\Gamma = \{t_1, t_2, ..., t_s\}$ where $t_r \epsilon \Omega$ for $r = 1, ..., s$, i.e., the set of time points used to minimize the Brier Score for the stacking procedure. Define the sum of misspecified model weights as $\widetilde{\alpha} = \sum_{k=l+1}^{m} \alpha_k$. Then for all $\widetilde{\alpha} > 0$ there exists at least one $t_r \epsilon \Gamma$ such that,

$$\sup_{\boldsymbol{x}} | \sum_{k=l+1}^{m} \frac{\alpha_k}{\widetilde{\alpha}} \hat{S}_k(t_r|\boldsymbol{x}) - S_o(t_r|\boldsymbol{x})| \to c,$$

where $c > 0$.

B.2.1 Proof of Theorem 1

We first must show that the expected value of our objective function, i.e., the Brier Score, is bounded by some function that has finite expectation. By assumption B2,

$$\begin{aligned} E[\frac{\Delta_i(t_r)}{G(T_i(t_r)|\boldsymbol{x}_i)}\{Z_i(t_r) - \sum_{k=1}^{m} \alpha_k \hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\}^2] &< E[\frac{1}{\delta}\{Z_i(t_r) - \sum_{k=1}^{m} \alpha_k \hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\}^2] \\ &< \infty. \end{aligned}$$

Then Lemma 2.4 of Newey and McFadden [1994] justifies the law of large numbers, i.e., the Brier Score asymptotically approaches its expectation. We therefore can determine the asymptotic minimizer by considering the expectation of the Brier Score. Since assumption B2 implies $E[\frac{\Delta(t_r)}{G(T(t_r)|\boldsymbol{x})}|T, \boldsymbol{x}] = \frac{1}{G(T(t_r)|\boldsymbol{x})} \times E[\Delta(t_r)|T, \boldsymbol{x}] = 1$, we obtain by double expectation that

$$\begin{aligned} E[BS(t_r)|\boldsymbol{x}] &= E[\{Z(t_r) - \sum_{k=1}^{m} \alpha_k S_k(t_r|\boldsymbol{x})\}^2|\boldsymbol{x}] \\ &= E[\{Z(t_r) - S_o(t_r|\boldsymbol{x})\}^2|\boldsymbol{x}] + \{S_o(t_r|\boldsymbol{x}) - \sum_{k=1}^{m} \alpha_k S_k(t_r|\boldsymbol{x})\}^2 \\ &= S_o(t_r|\boldsymbol{x})\{1 - S_o(t_r|\boldsymbol{x})\} + \{S_o(t_r|\boldsymbol{x}) - \sum_{k=1}^{m} \alpha_k S_k(t_r|\boldsymbol{x})\}^2 \end{aligned}$$

The asymptotic minimization problem becomes

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}} \sum_{r=1}^{s}[S_o(t_r|\boldsymbol{x})(1-S_o(t_r|\boldsymbol{x})) + \\ &\qquad \{S_o(t_r|\boldsymbol{x}) - [\sum_{k=1}^{l}\alpha_k S_o(t_r|\boldsymbol{x}) + \sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})]\}^2] \\ &= \arg\min_{\boldsymbol{\alpha}} \sum_{r=1}^{s}[S_o(t_r|\boldsymbol{x}) - \{S_o(t_r|\boldsymbol{x})\sum_{k=1}^{l}\alpha_k + \sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})\}]^2.\end{aligned}$$

At this point, we know there exists $\boldsymbol{\alpha}$ such that $\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x})$, e.g., $\alpha_1 = 1$. However, we need to show that the sum of correctly specified model weights equals one, i.e., $\sum_{k=1}^{l}\alpha_k = 1$. Suppose the sum of misspecified model weights is greater than zero, i.e., $\widetilde{\alpha} = \sum_{k=l+1}^{m}\alpha_k > 0$, then

$$\begin{aligned}\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) &= S_o(t_r|\boldsymbol{x}) \\ \Rightarrow S_o(t_r|\boldsymbol{x})\sum_{k=1}^{l}\alpha_k &+ \sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x}).\end{aligned}$$

Subtracting the correctly specified models from each side, we obtain by the sum-to-one constraint that $\sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x})\sum_{k=l+1}^{m}\alpha_k$. This implies that

$$\sum_{k=l+1}^{m}\frac{\alpha_k}{\widetilde{\alpha}}S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x}),$$

which contradicts assumption A3. Therefore, by the non-negativity constraint, $\widetilde{\alpha} = 0$ and hence $\sum_{k=1}^{l}\hat{\alpha}_k \to 1$ as $n \to \infty$.

B.2.2 Proof of Theorem 2

Define the random variables: $A_n^k = \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})|$. By assumption B1, as $n \to \infty$, $A_n^k \to 0$ for $k = 1, ...l$, while $A_n^k \to c_k$ for some $c_k > 0$ $(k = l+1, ..., m)$. Now

$$\begin{aligned}
\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| &= \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}) - \sum_{k=1}^{m} \hat{\alpha}_k S_o(t|\boldsymbol{x})| \quad \text{(B.1)} \\
&= \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \{\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})\}| \\
&\leq \sum_{k=1}^{m} \hat{\alpha}_k \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| \quad \text{(B.2)} \\
&= \sum_{k=1}^{l} \hat{\alpha}_k A_n^k + \sum_{k=l+1}^{m} \hat{\alpha}_k A_n^k \\
&\to 1 \times 0 + \sum_{k=2}^{m} \{0 \times c_k\} = 0, \quad \text{(B.3)}
\end{aligned}$$

where line (B.1) holds due to $\sum_{k=1}^{m} \hat{\alpha}_k = 1$, line (B.2) by the triangle inequality, and line (B.3) by Slutsky's lemma and Theorem 1. This implies that the stacked estimator is uniformly consistent as long as the correctly specified models are uniformly consistent.

B.3 Time-Dependent Stacking

Potential added flexibility for stacked survival models allows the weights to depend on time, i.e., $\hat{\alpha}_k(t)$. Similar to the approach proposed by Fan and Zhang [2000] for functional data analysis, a two-step estimation procedure is investigated that first obtains "raw estimates" at event times, then smoothes the raw estimates to obtain the final refined time-dependent weights.

The first step estimates the stacking weights for each N event times (i.e., $t_{(1)}, ..., t_{(N)}$). That is, for a given $t_{(r)}$,

$$\hat{\boldsymbol{\alpha}}(t_{(r)}) = \arg \min_{\boldsymbol{\alpha}(t_{(r)}), \alpha_k(t_{(r)}) \geq 0} \sum_{i=1}^{n} \frac{\Delta_i(t_{(r)})}{G(T_i(t_{(r)})|\boldsymbol{x}_i)} \times \{Z_i(t_{(r)}) - \sum_{k=1}^{m} \alpha_k(t_{(r)}) \hat{S}_k^{(-i)}(t_{(r)}|\boldsymbol{x}_i)\}^2,$$

with the additional constraint that $\sum_{k=1}^{m} \hat{\alpha}_k(t_{(r)}) = 1$. The $\hat{\alpha}_k(t_{(r)})$ are called the "raw estimates". Since the raw estimates can vary substantially across time, the second step smoothes the raw estimates to decrease variability.

While there are several potential avenues for smoothing the raw estimates, local constant regression, e.g., see Ruppert et al. [2003], stabilizes the estimates while maintaining both

the sum-to-one and non-negativity constraints. In particular,

$$\hat{\alpha}_k^{TD}(t) \;=\; \sum_{r=1}^{N} K\left(\frac{t_{(r)}-t}{h}\right)\hat{\alpha}_k(t_{(r)}) \Big/ \sum_{r=1}^{N} K\left(\frac{t_{(r)}-t}{h}\right),$$

where $t_{(r)}$ the r^{th} ordered event time, and $K(\cdot)$ is a symmetric probability density. The final estimate for the time-dependent stacking procedure is

$$\hat{S}^{TD}(t|\boldsymbol{x}) = \sum_{k=1}^{m} \hat{\alpha}_k^{TD}(t)\hat{S}_k(t|\boldsymbol{x}), \tag{B.4}$$

where $\hat{S}_k(t|\boldsymbol{x})$ is the k^{th} conditional survival estimate using all the data. This two-step approach to estimating time-dependent weights is appealing for its simplicity and straightforward computational implementation.

Conceptually, time-dependent weights may perform better by shifting weight between survival models as the appropriateness of the distributional assumptions vary across time. However, time-dependent weights increase the flexibility of stacked survival models and, therefore, generally possess a larger variance than time-independent weights. As such, when the stack includes a correctly specified model, time-independent weights will likely perform better than time-dependent weights. In addition, the conditional survival function with time-dependent weights, i.e., $\hat{S}^{TD}(t|\boldsymbol{x})$, is not guaranteed to be a non-increasing function with respect to survival time (which is an essential characteristic of survival functions). In fact, an increasing survival function occurred for a handful of points in the GBCS analysis (see Section 3.5). As such, time-dependent weights may improve predictive performance, but the conceptual cohesion of the conditional survival function is potentially compromised.

We note that adding a non-decreasing constraint for all of the time-dependent weights would ensure a non-increasing survival function. However, it is easy to show that a non-decreasing constraint on all of the time-dependent weights would result in constant (i.e., time-independent) weights due to the sum-to-one constraint. In addition, we note that a non-decreasing constraint on one survival model and a non-increasing constraint on a separate survival model will *not* ensure a non-increasing survival function. As such, it is difficult to fix the conceptual cohesion of the time-dependent stacking.

Remark B.1. Estimating time-dependent weights requires the selection of a bandwidth h. A reasonable approach sets the bandwidth to the standard error of the observed event times. This ensures h approaches 0 at a correct speed for the asymptotic results. However, there may exist a more optimal approach.

Remark B.2. Tables B.1, B.2, and B.3 compare time-dependent stacking to time-independent stacking (i.e., the approach in Chapter 3). Time-independent stacking performs better for scenarios with linear effects across all simulation setups. In contrast, time-dependent stacking performs as well as, or slightly better, than time-independent stacking for scenarios with non-linear effects.

TABLE B.1: Simulation results for Section 4.1 ($n = 200$, $p = 4$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 100. 'Stacking (TI)' is stacking with time-independent weights, and 'Stacking (TD)' is stacking with time-dependent weights.

	Models	log-Normal	Weibull	Gamma
Linear	Stacking (TI)	1.24	1.62	1.34
Effects	Stacking (TD)	1.86	1.92	1.88
Non-Linear	Stacking (TI)	7.59	4.80	9.02
Effects	Stacking (TD)	6.50	4.75	7.76

TABLE B.2: Simulation results for Section 4.2 ($n = 1000$, $p = 4$ covariates, and 75% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 1000. 'Stacking (TI)' is stacking with time-independent weights, and 'Stacking (TD)' is stacking with time-dependent weights.

	Models	log-Normal	Weibull	Gamma
Linear	Stacking (TI)	0.34	0.42	0.37
Effects	Stacking (TD)	0.44	0.49	0.44
Non-Linear	Stacking (TI)	4.56	2.29	5.01
Effects	Stacking (TD)	4.61	2.28	4.98

TABLE B.3: Simulation results for Section 4.3 ($n = 200$, $p = 40$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 100. 'Stacking (TI)' is stacking with time-independent weights, and 'Stacking (TD)' is stacking with time-dependent weights.

	Models	log-Normal	Weibull	Gamma
Linear	Stacking (TI)	4.65	3.30	4.96
Effects	Stacking (TD)	4.78	3.39	5.16
Non-Linear	Stacking (TI)	9.81	6.26	11.6
Effects	Stacking (TD)	9.56	6.15	11.4

Appendix C

Restricted Mean Treatment Effects

We investigate the association between the performance of the conditional survival function and the restricted mean treatment effect. While the conditional survival function plays an obvious role in restricted mean estimation, we want to know whether better estimation of the conditional survival function guarantees better estimation of restricted mean treatment effects. Chapter 4 illustrates that better conditional survival function estimation is weakly associated with better absolute bias, while better conditional survival function estimation is strongly associated with better MSE. This Appendix presents a discussion of bounding the MSE of the restricted mean treatment effect by the mean-squared error of the conditional survival function.

Similar to Chapter 3, we define the mean-squared error for a conditional survival function estimator of the a^{th} treatment as the integral of the squared error at time t over $\Omega = (0, \tau)$:

$$\text{MSE}_\tau\{\hat{S}^{(a)}(\cdot|\boldsymbol{x})\} = E\{E_{\boldsymbol{x}} \int_0^\tau [\hat{S}^{(a)}(t|\boldsymbol{x}) - S_o^{(a)}(t|\boldsymbol{x})]^2 dt\}.$$

A significant difference between this investigation and Chapter 3 is the addition of treatment a. Note that the inner expectation is with respect to the marginal, rather than conditional, covariate distribution. Since restricted mean treatment effect estimation requires two conditional survival functions (one for each treatment), we take the simple average of the mean-squared error for both treatments. That is, Chapter 4 uses

$$\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} \quad = \quad \frac{1}{2} \times \{\text{MSE}_\tau\{\hat{S}^{(0)}(\cdot|\boldsymbol{x})\} + \text{MSE}_\tau\{\hat{S}^{(1)}(\cdot|\boldsymbol{x})\}\}, \qquad \text{(C.1)}$$

as the mean-squared error for an estimator of the conditional survival function. We can then show that the mean squared error of the restricted mean treatment effect is bounded by the mean-squared error of the conditional survival function:

Theorem C.1. *Let the mean squared error of an estimator of the restricted mean treatment effect be* $MSE[\hat{\gamma}(\tau)] = E\{\hat{\gamma}(\tau) - \gamma(\tau)\}^2$*, then*

$$MSE[\hat{\gamma}(\tau)] \quad \leq \quad 2\tau \times MSE_\tau\{\hat{S}(\cdot|\boldsymbol{x})\}.$$

The result — which is a consequence of a sequential application of the Jensen, triangle, and Schwarz inequalities — helps explain the strong association of the mean squared error of the restricted mean treatment effect with the performance of the conditional survival function estimator. The bias is also bounded, but the limit is less tight due to a positive variance term. This results in a less strong, but still positive, association of bias with the mean-squared error of the conditional survival function.

Proof:

$$\begin{aligned}
E\{\hat{\gamma}(\tau) - \gamma(\tau)\}^2 &= E\left\{E_{\boldsymbol{x}}\int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]dt - \right. \\
&\qquad \left. E_{\boldsymbol{x}}\int_0^\tau [S^{(1)}(t|\boldsymbol{x}) - S^{(0)}(t|\boldsymbol{x})]dt\right\}^2 \\
&\leq EE_{\boldsymbol{x}}\left\{\int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - S^{(1)}(t|\boldsymbol{x})]dt + \right. \qquad \text{(C.2)} \\
&\qquad \left. \int_0^\tau [S^{(0)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]dt\right\}^2 \\
&\leq EE_{\boldsymbol{x}}\left(\{\int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - S^{(1)}(t|\boldsymbol{x})]dt\}^2 + \right. \qquad \text{(C.3)} \\
&\qquad \left. \{\int_0^\tau [S^{(0)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]dt\}^2\right) \\
&\leq \tau \times EE_{\boldsymbol{x}}\left(\int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - S^{(1)}(t|\boldsymbol{x})]^2 dt + \right. \qquad \text{(C.4)} \\
&\qquad \left. \int_0^\tau [S^{(0)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]^2 dt\right) \\
&= \tau \times \{\mathrm{MSE}_\tau\{\hat{S}^{(0)}(\cdot|\boldsymbol{x})\} + \mathrm{MSE}_\tau\{\hat{S}^{(1)}(\cdot|\boldsymbol{x})\}\} \\
&= 2\tau \times \mathrm{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\},
\end{aligned}$$

where line (C.2) holds by Jensen's inequality, line (C.3) holds by the triangle inequality, and line (C.4) holds by Schwarz's inequality.