2017

Survival analysis with Bayesian additive regression trees and its application

Satabdi Saha

ABSTRACT

SURVIVAL ANALYSIS WITH BAYESIAN ADDITIVE REGRESSION TREES AND ITS APPLICATION

Satabdi Saha, MS
Division of Statistics
Northern Illinois University, 2017
Dr Duchwan Ryu, Director

The objective of this study is to compare the performance of Bayesian Additive Regression Trees (BART) with Cox Proportional Hazards (CPH) and Random Survival Forests (RSF) models using simulation studies and a real data application on breast cancer survival data as provided by the U.S. SEER database for the year 2005. In the simulation study, we compared the three models across varying sample sizes and censoring rates on the basis of bias and prediction accuracy. Results obtained indicate that the performance of the CPH model depreciates when the PH assumption is violated, however BART continues to perform with almost equal effectiveness. In the real data application, a retrospective analysis was performed in 1500 patients having invasive ductal carcinoma. According to several performance assessment measures, BART and RSF based on log-rank splitting rule fare equivalently and BART marginally outperforms CPH. BART is shown to have similar functioning capacities as RSF, however being in the Bayesian paradigm, BART additionally allows for natural quantification of uncertainty and construction of credible and prediction intervals. The prognostic competence of BART along with the interpretative results obtained using the partial dependence survival functions and variable importance measures can thus be effectually used to solve potential future survival problems.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

AUGUST 2017

SURVIVAL ANALYSIS WITH BAYESIAN ADDITIVE REGRESSION TREES AND ITS APPLICATION

BY

SATABDI SAHA

A THESIS SUBMITTED TO THE GRADUATE SCHOOL IN

PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

DIVISION OF STATISTICS

Dissertation Director:
Dr Duchwan Ryu

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Dr. Duchwan Ryu for his patience, encouragement and aid while I completed this thesis. I would also have to thank the members of my masters committee, Dr Nader Ebrahimi and Dr Alan Polansky for agreeing to be in my committee and being patient readers of my work. Special thanks to Dr Haiming Zhou for introducing me to Survival Analysis and answering my endless doubts.

I extend my gratitude to my fellow graduate students for the stimulating discussions we had while working together for courses, and all the fun we have had in the last two years. Thanks to Joel, Michael, Claudine and Dr. McCullough for helping me with my TA responsibilities.

Above all else, I am forever grateful to my husband and mom, who were always willing to listen to my complaints, encourage me during my periods of hopelessness, and incessantly pushing me towards my dreams. I could not have made it anywhere without their unconditional love and support.

DEDICATION

To my loving husband and Ma for their endless support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background

Invasive ductal carcinoma (IDC) (Invasive Ductal Breast Cancer-Figure 1.1), sometimes called infiltrating ductal carcinoma, is the most common type of breast cancer. Nearly 80% of all breast cancers are invasive ductal carcinomas. Invasive refers to the condition where the cancer cells have attacked or spread to the contiguous breast tissues and ductal signifies that the cancer originated in the milk ducts; which are the channels that carry milk from the milk-producing lobules to the nipple. Carcinoma refers to any cancer that starts in the skin or other tissues that shelter internal organs such as breast tissue. Summing up it can be said that, invasive ductal carcinoma refers to cancer that has broken through the wall of the milk duct and started to invade the tissues of the breast. Over time, invasive ductal carcinoma can spread to the lymph nodes and possibly to other areas of the body. According to the American Cancer Society, more than 180,000 women in the United States are diagnosed with invasive breast cancer each year and most of these are cases of invasive ductal carcinoma. Although invasive ductal carcinoma can affect women at any age, it is more common among women of older age. According to the American Cancer Society, about two-thirds of women are 55 or older when they are diagnosed with an invasive breast cancer. There have been several studies reporting the effects of tumor size, tumor stage and tumor grade on the survival of breast cancer patients (D'Eredita et al. (2001), Rosenberg et al. (2005), Delen et al. (2005), Omurlu et al. (2009) Faradmal et al. (2014)). Our study

uses a dataset taken from the U.S. National Cancer Institutes Surveillance, Epidemiology, and End Results (SEER) database (Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), April 2016 based on November 2015 submission) which contains huge magnitude of data on several factors affecting breast cancer from 1973-2013.

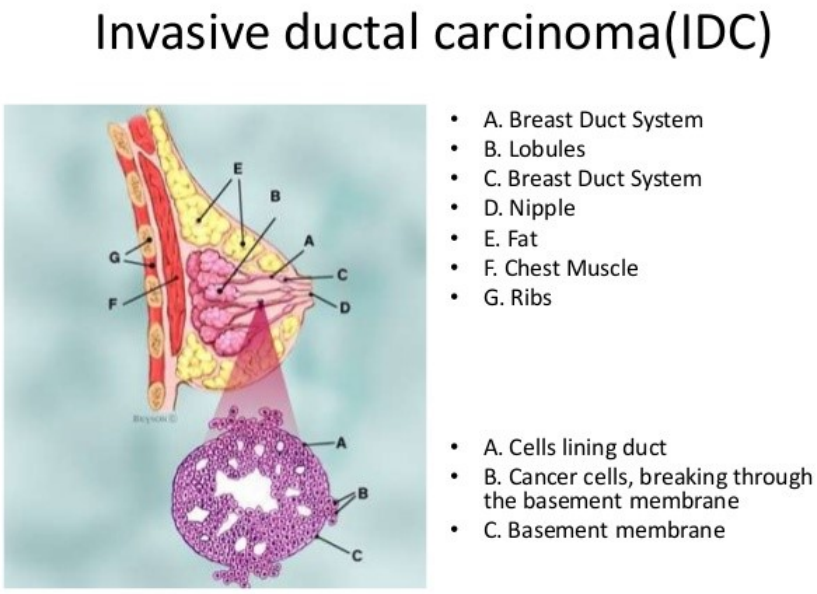

Figure 1.1: Invasive Ductal Carcinoma

This study is primarily concerned about studying the effect of several important covariates on the survival times of female breast cancer patients for the year 2005. The application of the Cox proportional hazards model seems to be the first natural step since it is the most commonly used method in modelling survival times using several covariates. It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest. The model assumes a non-parametric form for the baseline hazard, which implies that the hazard function of the comparing groups will be always proportional to the baseline hazard, irrespective of its form. This grants us the flexibility to estimate

hazard ratios without concerning ourselves about the functional form of the baseline hazard. The CPH model however, results in biased parameter estimates if the covariates fail to follow the proportional hazards assumption; and the assumption is often violated due to the presence of complex relationships in the data structure. Failure of the proportional hazards assumption sometimes leads to the use stratified Cox models that stratify with respect to covariates or time dependent Cox models that incorporate a time dependent interaction term to deal with the non-proportionality of hazards. However several other problems including non-linearity, interactions between covariates and high dimensional parameter spaces cannot be effectively addressed using these methods. Various non-parametric modeling methods such as penalized regression (Tibshirani et al. (1997), Zhang and Lu (2007)), survival trees (LeBlanc and Crowley, 1993), boosting with Cox-gradient descent (Ma and Huang, 2007) and random survival forests (Ishwaran et al., 2008) have been devised to address the various problems faced by the CPH model.

The use of Classification and regression trees (CART) and other recursive partitioning methods have allowed for a more detailed study of the effects of variables on the survival distribution. It has proved to be a very efficient tool in determining covariate importance and revealing covariate interactions in the data. Analyzing and interpreting complex non-linear and high dimensional datasets effectively and capably reducing its dimensionality are CARTs main advantages over other methods. But the CART method often suffers from high variance. This means that if we split the training data into two parts at random, and fit a decision tree to both halves, the results that we get could be quite different. In contrast, a procedure with low variance should yield similar results if applied repeatedly to distinct data sets. Various methods which combine a set of tree models, so called ensemble methods, have attracted much attention it helps in decreasing the variance and increasing the prediction accuracy of CART. These include boosting (Freund and Schapire (1995), Friedman (2001)), bagging (Breiman, 1996) and random forests (Breiman, 2001), each of which use different

techniques to fit a linear combination of trees. Boosting fits a sequence of single trees, using each tree to fit data variation not explained by earlier trees in the sequence. Bagging and random forests use randomization to create a large number of independent trees, and then reduce prediction variance by averaging predictions across the trees. These flexible non parametric modeling methods have been successfully applied in the context of survival analysis in the form of Bagging survival trees (Hothorn et al., 2004), Survival Ensembles (Hothorn et al., 2006) and Random Survival Forests (Ishwaran et al., 2008), to address the various problems faced by the CPH models.

Yet another approach that results in a linear combination of trees is Bayesian model averaging applied to the posterior arising from a Bayesian single-tree model as in Chipman et al. (1998), Mallick et al. (1999) and Blanchard (2004). Such model averaging uses posterior probabilities as weights for averaging the predictions from individual trees. This idea has been further developed by constructing a Bayesian sum-of-trees model where each tree is constrained by a regularization prior to be a weak learner, and fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm that generates samples from a posterior. Motivated by ensemble methods, this "sum of trees" model known as Bayesian additive regression trees (BART) provide a framework for flexible nonparametric modeling of relationships of covariates to outcomes. BART models have been shown to provide excellent predictive performance, for both continuous and binary outcomes, and exceeding that of its competitors (Random Forests, Boosting, etc.) (Chipman et al., 2010). Very recently the idea of BART has been extended to analyse survival data (Sparapani et al., 2016). The BART model for survival analysis has been formed by expressing the nonparametric likelihood for the KaplanMeier (KM) estimator in a form suitable for BART. In application to survival data, the BART model was shown to perform very well in terms of prediction error and had medium variance and medium bias. It successfully identified

complex non-linear relationships and interactions in the dataset and performed efficient variable selection (Sparapani et al., 2016).

Survival prediction is often formulated in terms of categorical outcomes(e.g. poor versus good prognosis), which may be useful for guiding decisions about cancer management and treatment (Ross, 2009). However, due to a large degree of heterogeneity observed within prognostic classes, efficient prediction of time to a clinical event may not be successful. Improved accuracy of survival prediction can be attained by relating time-to-event measures directly with covariates, which requires specific survival analysis methods that account for the presence of right censored outcomes and effectively deals with the problem of covariate interaction, non-linearity and high dimensionality.

In this research, we try to extract greater survival prediction accuracy while attempting to understand the importance of the individual covariates affecting overall survival. We use the CPH, RSF and BART models to predict survival probabilities, perform variable selection and determine the marginal effects of the covariates. The models are then compared and contrasted using simulation studies and a benchmark breast cancer dataset.

1.2 Overview of the Project

Chapter 2 begins with a basic overview of the widely used the parametric and semparametric models in survival analysis, with a special attention to the CPH model. Chapter 3 describes two non parametric ensemble tree structured models, RSF based on the frequentist random forest approach of bootstrapping and randomization and BART, a non parametric Bayesian regression approach which uses dimensionally adaptive random basis elements. Several performance assessment measures are reviewed to compare and contrast the predictive abilities of the models described. In chapter 4 we conduct simulation studies

to compare the effectiveness of BART with the non-parametric KM estimator, in one sample studies and the CPH and RSF models in regression scenarios. In Chapter 5 we use a real life dataset to demonstrate the potential of BART. The survival experience of 1500 patients having invasive ductal carcinoma is studied using CPH, RSF and the BART models. For evaluating the performance of the three models, the dataset is split randomly in the ratio 2:1 into a training and a validation set. We train each of the classifiers on the training set and test them on the validation set. Finally a discussion of our work in Section 6, and some potential prospects of future research, concludes the thesis.

CHAPTER 2

PARAMETRIC AND NON PARAMETRIC SURVIVAL ANALYSIS

2.1 Survival Analysis

Survival analysis generally comprises of a set of statistical methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest like death of a patient or relapse of a disease (Klein and Moeschberger, 2005). Several subjects are followed for a specified period of time until occurrence of the event of interest. For example, if the event of interest is relapse of breast cancer, then survival time is defined as the time in years until a person with a cured case of breast cancer is again diagnosed with the same malignancy.

Observation times are called censored when exact information about the subject's event time is not available. Censoring could occur when a patient withdraws from the study, is lost to follow up, or does not experience the event of interest before the study ends. These are all examples of right censoring. Left censoring occurs when the event of interest has already occurred before a patient enrolls into the study and interval censoring is encountered when the times of the event of interest is not exactly known. Non-informative or random censoring occurs when the subject's censoring time is independent of its failure time and is usually required in order to avoid bias in survival analysis.

Generally, we use logistic regression models to understand how the risk factors are associated with the absence or presence of a disease. However, in survival analysis, we have

censored cases where we are unaware of the disease status of subjects who have dropped out. In such scenarios, survival regression models are used in order to estimate model parameters since they are able to incorporate information from both the censored and uncensored subjects. The response variable, generally the time to event, along with the censoring indicator, is modeled using relevant predictors to estimate the survival and the hazard functions, which forms the key component of analysis in survival studies. The survival function for any given time provides the probability of surviving upto that time for a given subject. The hazard function gives the potential that the event will occur per unit time, given that the subject has survived upto the specified time. Therefore for any given time t the survival function is given as

$$S(t) = P(T > t) = 1 - F(t) \tag{2.1}$$

where $F(t)$ denotes the cumulative distribution function of time t and the hazard function is given as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.2}$$

where $f(t)$ denotes the probability density function of time t. Generally survival regression models aim at evaluating the relationship between the time to event and a factor of interest in the presence of several other covariates. Several parametric, semi-parametric and non-parametric modelling approaches are used to estimate the survival and the hazard functions.

2.2 Kaplan-Meier Estimator

The survival function $S(t)$, which calculates the probability of a subject surviving longer than time t, is theoretically a smooth curve. The Kaplan-Meier (KM) product limit estimator by (Kaplan and Meier, 1958) is used to estimate the survival probabilities as a function of time using a non-parametric approach. Let $t_i, i = 1, 2, ..., n$, denote the event times for i individuals in the study, d_i denote the number of events occuring at this time and r_i be the corresponding number of subjects remaining in the study. For any event time $t \in [t_i, t_{i+1})$, the KM estimator of the survival function $\hat{S(t)}$ can be written as:

$$\hat{S(t)} = \prod_{i=1}^{n} (1 - \frac{d_i}{r_i}) \tag{2.3}$$

The KM estimator works very well in case of right censored data (Klein and Moeschberger, 2005).

2.3 Parametric Survival Models

Parametric survival models assume a known probability distribution for the underlying survival times. This leads to a concise equation and a smooth function for estimating $S(t)$ and $h(t)$. If the assumed parametric form is correct, its survival estimates are more precise than the KM estimates. Some popular distributions for survival times are the Weibull, and

exponential distributions (Klein and Moeschberger, 2005). For an exponential model, when $T \sim Exp(\frac{1}{\lambda})$

$$f(t) = \frac{1}{\lambda} \exp(-\frac{t}{\lambda}) \tag{2.4}$$

$$h(t) = \frac{1}{\lambda} \tag{2.5}$$

$$S(t) = \exp(-\frac{t}{\lambda}) \tag{2.6}$$

For an Weibull model, when $T \sim W(\lambda, \alpha)$

$$f(t) = \frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left(-\frac{t}{\lambda}\right)^{\alpha} \tag{2.7}$$

$$h(t) = \left(\frac{\alpha}{\lambda^{\alpha}}\right) t^{\alpha-1} \tag{2.8}$$

$$S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \tag{2.9}$$

2.4 Cox Proportional Hazards Regression Model

Survival models in general consists of two parts: baseline hazard function and model parameters. The baseline hazard function indicates the change of risk of an individual per unit time at base levels of the covariates. The model parameters indicate the association of the hazard with the covariates. The proportional hazards assumes that the covariates are multiplicatively associated to the hazard (Breslow, 1975) . When the proportional hazards assumption holds, the model parameters can be estimated without any consideration to the hazard function (Cox, 1992). The Cox regression model assumes proportional hazards.
Now let us consider data in the form (t_i, δ_i, x_i) where t_i is the time to event, δ_i is the censoring indicator and x_i forms the vector of covariates for indexes $i = 1, 2, ..., n$. The censoring indicator takes values 1 when the event has occurred and value 0 when the time

is right censored. We order the distinct event and censoring times t_i as $0 < t_{(1)} < t_{(2)} < , ..., < t_{(k)} < \infty$, where $t_{(j)}$ denotes the j_{th} ordered statistic among distinct observation and censoring times, assigning $t_{(0)} = 0$.

The Cox proportional hazards model for survival data is given as

$$h_{\text{Cox}}(t|x_i) = h_0(t)exp(\beta^T x_i) \tag{2.10}$$

where $h_{Cox}(t|x_i)$ is the proportional hazard at time t given the vector of covariates x_i, β is the vector of unknown regression coefficients and $h_0(t)$ is the arbitrary baseline hazard function.The cumulative hazard function $H_{Cox}(t|x_i)$ and survival function $S_{Cox}(t|x_i)$ are given as

$$H_{\text{Cox}}(t|x_i) = \int_0^t h_{\text{Cox}}(v|x_i)dv \tag{2.11}$$

$$S_{\text{Cox}}(t|x_i) = exp[-H(t|x_i)] \tag{2.12}$$

2.5 Tree based survival models

CART proposed by Breiman et al. (1984) were initially developed to model categorical and continuous outcomes using a set of covariates from available data. The basic idea of a tree model is to partition the covariate space recursively to form groups (tree nodes) of subjects which have similar response values. This is often achieved by minimizing node impurity. Gini and entropy measures are very popular for a categorical response while the sum of squared deviations from the mean are used in case of regression trees. For a single continuous covariate X, a binary split of the form $X \leq c$ is conducted where c is a threshold value. For a categorical covariate X, a split has the form $X \in (c_1, c_2, ..., c_k)$, where $(c_1, c_2, ..., c_k)$ are the possible values of X. The splitting algorithm starts with the root

node containing all the observations, searches exhaustively for all potential binary splits with the covariates; and selects the best split according to a splitting criteria such as an impurity measure. In the CART algorithm, the splitting process is repeated recursively on the daughter nodes until a stopping criterion is reached (often until a minimum node size is attained). This procedure generally leads to overfitting. To overcome overfitting problems, a pruning and selection method is then applied to find the best subtree. Ensemble methods like Bagging and Random Forests are also used to deal with the problem of overfitting. Node summaries obtained in the terminal nodes are used as the predicted values. Generally, the node mean is used in case of continuos covariates and the most frequent value appearing in the node is selected in case of categorical covariates.

As opposed to the linear CPH model which assumes that a single line or surface is sufficient to estimate the response over varying predictors, tree based approaches provide alternative partitions of the covariate space which sometimes lead to more accurate response estimates. In this regard the CART methodology has been used extensively in modeling survival data. The basic ideas of splitting and pruning still remain the same, however different approaches (detailed in the next chapter), relevant in survival settings, are used. When modeling survival data, it is frequently of interest to determine which covariates affect the survival distribution and to judge whether the effect is valid across all individuals or within subsets. Statistically, these questions are often posed as variable selection and detection of interactions. CART forms a very significant tool for revealing structure in the data and answering these questions. The main advantage of CART over other survival methods lies in its ease of interpretation of the results and its ability to analyze complex nonlinear datasets with many covariates, by effectively reducing the dimensionality of the data.

CHAPTER 3

TREE BASED REGRESSION MODELS IN SURVIVAL ANALYSIS

3.1 Trees in Survival Analysis

Several extensions of CART to censored survival data, have been proposed in the literature. These extensions generally provide modifications to the four basic building ideas expressed in CART: the prediction rule, the splitting rule, the pruning algorithm and the tree selection. The prediction rule for survival analysis is typically based on the estimate of the distribution function, which implies that the three other ideas work non-parametrically on the space of distributions. The extensions of CART to survival data fall can be broadly divided into two groups. The first approach uses a test statistic that determines within-node homogeneity, ie. similarity in the survival experiences of observations in a node. The alternative approach is based on separation measures, ie. calculating a test statistic that distinguishes between survival experiences.

Gordon and Olshen (1985) presented the first extension of CART to censored survival data, which involves a distance measure, Wasserstein metric (Olkin and Pukelsheim, 1982) between Kaplan- Meier curves and certain point masses. Their approach assumes a piecewise exponential model with one data-determined knot. Ciampi et al. (1995) proposed a different recursive partitioning technique based on between node splitting instead of splitting based on reducing within node variability. Davis and Anderson (1989) suggested a method based on the observed likelihood, while assuming an exponential baseline hazard function. LeBlanc

and Crowley (1993) used the logrank statistic as a splitting criterion, but they introduced a new method for pruning and selecting a final tree built around a measure of split complexity. These extensions of CART are based on a definition of a within-node homogeneity measure. The use of within-node homogeneity based on likelihood statistics allows the inheritance of all subsequent CART methodology, since the measures defined are all subadditive, allowing comparison between subtrees. Segal (1988) argued that tests for between-node separation can tell more about the important prognostic factors associated with the survival phenomenon under study than within-node homogeneity. He developed an algorithm that grows and prunes a tree based on the logrank test, however he didnot provide an algorithm for selection of an optimal tree size. In the recent years, complex situations relating to the usage of survival trees in analyzing multivariate and correlated survival data have been researched (Gao et al. (2004), Gao et al. (2006), Fan et al. (2006)). Extensive studies have also been carried out regarding the use of ensemble tree methods to study survival data. (Hothorn et al. (2004), Ishwaran et al. (2008), Krketowska (2004), Krketowska (2006)). The random survival forest approach has been described in subsection 3.1.1. Studies have also dealt with time varying covariates and discrete time survival data (Xu et al. (2001), Xu and Adak (2002), Yin et al. (2002), Bou-Hamad et al. (2011), Huang et al. (1998), Fahrmeir (1998)). A review of tree-based methods for survival can be found in (Zhang and Singer, 2013). The Bayesian versions of CART (Chipman et al. (1998), Mallick et al. (1999)) and other Bayesian binary trees (Loh (2011), Pittman et al. (2004)) have also been used in predicting survival data. Research has shown that the prediction accuracy of such models can be improved through Bayesian Model Averaging (Madigan et al., 1996), Bagging (Breiman, 1996), Boosting (Schapire et al., 1998) and related methods. Bayesian Additive Regression Tree model (Chipman et al., 2010), described in subsection 3.1.2, forms the basis of one such approach.

3.1.1 Random Survival Forests

Random Forests are a machine learning ensemble method that combines the idea of bootstrap aggregation and random selection of features. Introduced by Brieman in 2001, it is considered to be an extension of the bagged CART model (Breiman, 1996). Survival trees are binary trees grown by recursive partition of the root nodes. In random survival forests by (Ishwaran et al., 2008), a predictor is developed by combining the results obtained from growing survival trees on many bootstrap samples of the same data. The algorithm can be summarized as follows

1. Draw B bootstrap samples from the data.

2. A survival tree is grown for each of the bootstrap samples by recursive splitting of the root nodes into interior and terminal nodes.

   (a) At each tree node a random selection of a subset of predictor variables is made.

   (b) Among all the binary splits made by the predictor variables selected in (a), the best split is determined using the predictor variable that maximizes the survival difference between daughter nodes.This splitting rule can be based on log-rank splitting (Segal (1988), LeBlanc and Crowley (1993)) or log-rank score splitting (Hothorn and Lausen, 2003)

   (c) Repeat (a) and (b) recursively unless a terminal node has a minimum of $d_0 > 0$ deaths.

3. Calculate the cumulative hazard function (or survival function) for each tree and then average over the B bootstrap samples to obtain a ensemble cumulative hazard function (or ensemble survival function).

3.1.1.1 Ensemble Cumulative Hazard Function

Let $(t_{i,l}, \delta_{i,l})$ be the survival time and the censoring indicator for individual cases in the l^{th} terminal node, where $(\delta_{i,l} = 0)$ indicates that the event is right censored and $(\delta_{i,l} = 1)$ indicates that the individual has experienced an event. Let x_i be the vector of covariates for the individual cases. We now order the distinct event and censoring times in the l^{th} terminal node $t_{i,l}$ as $0 < t_{(1,l)} < t_{(2,l)} <, ..., < t_{(k,l)} < \infty$, where $t_{(j,l)}$ denotes the jth ordered statistic among distinct observation and censoring times, assigning $t_{(0,l)} = 0$. Let $d_{j,l}$ and $Y_{j,l}$ be the number of deaths and the individuals at risk at time $t_{(j,l)}$. Then the estimate for the cumulative hazard function (CHF) and the survival function is given as

$$H(t|x_i) = \sum_{t_{j,l} \leq t} \frac{d_{j,l}}{Y_{j,l}} \qquad \text{for } x_i \in l \tag{3.1}$$

$$S(t|x_i) = exp[-H(t|x_i)] \qquad \text{for } x_i \in l \tag{3.2}$$

Now let $H_b(t|x_i)$ be the CHF obtained for a tree grown on the b^{th} bootstrap sample, then the bootstrap ensemble CHF obtained as an average over the B survival trees can be represented as,

$$H_{\mathrm{RF}}(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} H_b(t|x_i) \tag{3.3}$$

The survival function can then be obtained using 3.2

3.1.1.2 Out of Bag ensemble CHF estimation

In a random forest, each of the trees is grown using an independent bootstrap sample from the set of training observations. It can be shown that each of these trees makes use of only

two thirds of the training sample observations. The remaining one-third of the observations that are not used to construct a tree from the particular bootstrap sample is referred to as out of bag (OOB) observations. We can predict the ensemble CHF (or ensemble survival function) for the i^{th} observation using the trees for which it was OOB. This will give us B/3 predictions for B bootstrap samples. For calculating a single predictive estimate, we obtain an average over these predictions. The resulting OOB predicted ensemble CHF (or ensemble survival function) is a valid test set prediction obtained from the random forest model since the predicted response obtained for each of the observation is calculated using the trees that were not grown using that observation. Further we can calculate the c-index, as described later in section 3.3.3, based on the OOB data, which serves as the test set c-index. The OOB prediction error can then be obtained by subtracting the OOB c-index from one.

3.1.1.3 Ensemble Mortality

The basic idea underlying the approach of random survival forests is a conservation-of-events principle for survival trees (Naftel et al., 1985) and that has been used by Ishwaran et al. (2008) to define ensemble mortality, a new predicted outcome for survival data . Let (T_i, δ_i) denote the survival times and the censoring indicator for the nonbootstrapped data. then mortality can be defined as the expected value of the CHF summed over time T_j, conditioned on a specific covariate x_i. It measures the expected number of deaths under a null hypothesis of similar survival experience. Mathematically the expression for mortality can be written as

$$M_i = E_i \left(\sum_{j=1}^{n} H(T_j | x_i) \right) \tag{3.4}$$

where E_i is the expectation of similar surival behaviour under the null hypothesis. In a survival tree, each of the terminal nodes share a common estimated hazard function thus experiencing a similar survival behaviour. Therefore we can define the ensemble mortality estimator as

$$M_{\mathrm{RF},i} = \sum_{j=1}^{n} H_{\mathrm{RF}}(T_j|x_i) \tag{3.5}$$

3.1.2 Bayesian Additive Regression Trees

BART method of Chipman et al. (2010) is a non-parametric Bayesian method that uses a Bayesian sum of trees model, which enables full posterior inference including point and interval estimates of the unknown regression function. Motivated by ensemble methods, each tree in the model is constrained by a regularization prior to be a weak learner, fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm that generates samples from the posterior. The successive iterations of the backfitting MCMC algorithm forms an MCMC sample from the posterior distribution over the sum of trees model space. Posterior mean estimate of any input value x is simply obtained by averaging over the consecutive sum of trees model draws, calculated at x. Partial dependence functions which reveal the marginal effects of the x components can also be determined similarly. Variable selection is made possible by calculating the relative frequencies of the x components appearing in the sum of trees model iterations. The corresponding relative frequencies help in explaining the relative importance of the x components in determining the variation in Y. For classification purposes using BART, a probit extension of the described method is used.

3.1.2.1 The BART Model

To explain the form of a sum of trees model we start with a single tree model. Let y_i denote an outcome and x_i denote the vector of covariates modelled by the structure $y_i = g(x_i; T, M) + \epsilon_i$. Notationally, T denotes a binary tree consisting of interior and terminal nodes. A branch decision rule at each interior node, typically splits the predictor space into two regions $\{x \in S\}$ vs $\{x \notin S\}$,where S is a subset of the range of x, based on a single component of the covariate vector. $M = \{\mu_1, \mu_2, ..., \mu_b\}$ denotes a set of functional values associated with the b terminal nodes. For a given T and M we use $g(x_i; T, M)$ to denote a function that assigns a $\mu_i \in M$ to x_i. Thus,

$$y_i = g(x_i, T, M) + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \tag{3.6}$$

is a single tree model. Therefore referring to this notation, the BART model can be expressed as

$$y_i = f(x_i) + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \tag{3.7}$$

$$f(x_i) = \sum_{i=1}^{m} g(x_i; T_j, M_j) \tag{3.8}$$

For the Bayesian specification of the sum of trees model we need to impose priors over the parameters of the BART model, namely $(T_1, M_1), ...(T_m, M_m)$ and σ. Priors are carefully chosen to regularize the fit and to curtail strong individual tree effects. Thus the regularization prior is specified as

$$p((T_1, M_1), ..., (T_m, M_m) = [\prod_j p(T_j, M_j)]\ p(\sigma)$$
$$= [\prod_j p(M_j|T_j)p(T_j)]\ p(\sigma)$$

and

$$p(M_j|T_j) = \prod_j p(\mu_{jk}|T_j) \tag{3.9}$$

where $\mu_{jk} \in M_j$. Following Chipman et al. (2010), we specify $p(T_j)$ by three aspects, the probability that a node at depth d is non-terminal is calculated as $\alpha(1+d)^{-\zeta}$ where $\alpha \in (0,1)$ and $\zeta \in [0, \infty)$, the choice of the splitting covariate at each interior node is uniformly distributed over the set of available covariates and the choice of a branching rule given a covariate at the interior node also follows an uniform distribution over the discrete set of available splitting values. The default values as specified by (Chipman et al., 2010) are $\alpha = 0.95$ and $\zeta = 2$. For $p(M_j|T_j) = \prod_j p(\mu_{jk}|T_j)$ we use the conjugate normal prior $\mu_{jk} \sim N(0, 2.25/m)$ on the values of the terminal nodes. This prior has the effect of curbing strong individual effects of the trees so that every tree forms only a small part of the entire sum of trees model. Now we need $p(\sigma)$ to complete the Bayesian model in 3.9, but it is not required in the reformulation of the BART model for survival analysis, described in 3.1.2.2, as it uses a probit regression with latent variables having unit variance.

3.1.2.2 BART Model in Survival Analysis

Let us consider data in the form (t_i, δ_i, x_i) where t_i is the time to event, δ_i is the censoring indicator where $(\delta_i = 1)$ indicates that an event has occurred whereas $(\delta_i = 0)$ indicates that the observed time was right censored and x_i forms the vector of covariates for indices

$i = 1, 2, ..., n$. We now order the distinct event and censoring times t_i as $0 < t_{(1)} < t_{(2)} < , ..., < t_{(k)} < \infty$, where $t_{(j)}$ denotes the jth ordered statistic among distinct observation and censoring times, assigning $t_{(0)} = 0$. Now we formulate event indicators y_{ij} for each subject i and distinct time $t_{(j)} \leq t_i$ where $t_i = t_{(n_i)}$ where n_i constitutes the total number of events occurred before the j^{th} ordered time reaches time t_i. Let us now define p_{ij} as the unconditional probability of an event occurring at time t_j. We then use the non-parametric probit regression structure to regress y_{ij} on time $t_{(j)}$ and the covariates x_i, and then use truncated normal latent variables z_{ij} to convert y_{ij} to a continuous variate as required by the BART model for continuous outcomes as formulated in equation (3.7) and (3.8). Therefore,

$$j = 1, ..., n_i; \qquad y_{ij} = \begin{cases} \delta_i, & t_i = t_{(j)} \\ \\ 0, & t_i \neq t_{(j)} \end{cases} \tag{3.10}$$

$$y_{ij}|p_{ij} \sim Bernoulli(p_{ij}) \tag{3.11}$$

$$f \sim BART \tag{3.12}$$

$$p_{ij}|f = \Phi(\mu_{ij}), \mu_{ij} = \mu_0 + f(t_{(j)}, x_i) \tag{3.13}$$

$$z_{ij}|y_{ij}, f \sim \begin{cases} N(\mu_{ij}, 1)I(-\infty, 0); & y_{ij} = 0 \\ \\ N(\mu_{ij}, 1)I(0, \infty); & y_{ij} = 1 \end{cases} \tag{3.14}$$

We considered $\mu_0 = \Phi^{-1}(\hat{p})$ for centering the latent variables.This described model is estimated using the existing software for binary BART. The (t_i, δ_i) data pairs available for survival analysis are then used to create the binary response vector y_{ij} for using the survival using BART model as described by Sparapani et al. (2016). The event times, constructed as described above are treated as the first column of the covariate matrix and the rest of the

columns are constructed by duplicating the the individual level covariate values to match the repition pattern of the first subscript of y. Thus using the BART model for binary data, we are able to get samples from the the posterior distribution of f. Then for any given covariate containing time t and rest of the covariates x, we can calculate the posterior distribution as

$$p(t, x) = \Phi(\mu_0 + f(t, x)) \tag{3.15}$$

Finally our intended purpose lies in estimating the survival and hazard functions as possible targets of inference. The survival and hazard function at event or censoring time $t_{(j)}, j = 1, 2, ..., k$: can be estimated as,

$$S_{\text{BART}}(t_{(j)}|x) = P(T > t_{(j)}|x) \tag{3.16}$$

$$h_{\text{BART}}(t_{(j)}|x) = \frac{p(t_{(j)}, x)}{(t_{(j)} - t_{(j-1)})} \tag{3.17}$$

These functions can be only calculated at distinct survival times, however using the constant hazard assumption,interpolation between these times can be accomplished.

Partial dependence survival function. In survival analysis we are generally interested in exploring the effects of individual covariates and their specific interactions on overall survival. For this purpose we need to calculate marginal survival functions involving single covariates or a subset of covariates. These marginal survival functions derived in Chipman et al. (2010), using Friedman's partial dependence function (Friedman, 2001) computes the individual effect of a covariate or a subset of covariates on overall survival while averaging over the others. Mathematically we partition the covariate space x as the set containing

the covariates of interest x_a and the remaining covariates x_b. Thus the partial dependence function can be defined as

$$f(x_a) = \frac{1}{n}\sum_{i=1}^{n} f(x_a|x_{ib}) \tag{3.18}$$

and the survival function (Sparapani et al., 2016) can be written as

$$S(t|x_a) = \frac{1}{n}\sum_{i=1}^{n} S(t|x_a, x_{ib}) \tag{3.19}$$

The means or medians over the samples of the posterior distribution can then be treated as the required estimates.

3.2 Variable selection in survival models

3.2.1 Stepwise variable selection using Cox Regression

We select the variables using a backward stepwise variable selection approach (implemented using the selectCox function in the R package pec, (Mogensen et al., 2012)). The variables in each step are selected using the Akaike information criteria (AIC) and then a Cox regression model is fit using the chosen predictors. The AIC criteria is closely related to the logarithmic scoring rule, which is strictly proper, (Gneiting and Raftery, 2007) and thus can be used for identifying a prediction model. The estimates of the regression coefficients and the baseline hazard so obtained are then used to estimate the survival function in 2.12.

3.2.2 Variable selection using Random survival forest

We start by calculating the OOB prediction error for each of the predictor variables x. For each x, this is achieved by dropping OOB cases down their in-bag survival tree and then assigning a daughter node randomly as soon as a split for the predictor variable is encountered. A variable importance measure is then defined as the difference between the original OOB prediction error and the new OOB prediction error. The variable importance measure, so calculated, evaluates the increase or drop in the misclassification rates on the OOB data, given that the variable x was not available. Predictor variables having a large variable importance measure are considered to have greater prognostic capacities, whereas the variables with zero or negative variable importance measure can be dropped from the original model, as they add nothing to its predictive ability.

3.2.3 Variable selection using BART

Variable selection in BART is carried out by selecting those variables which occurs most in the fitted sum of trees model. This method works better when the number of trees grown is small, as growing a large number of trees can give rise to an inappropriate mix of relevant and irrelevant predictors, leading to redundancy (Chipman et al., 2010). Variable selection is thus accomplished by observing the individual predictor usage frequency in a sequence of MCMC samples as the number of trees grown becomes smaller and smaller. Thus predictors with a higher usage frequency in the MCMC samples are considered to have higher prognostic competence as compared to the other predictors.

3.3 Performance assessment of the prediction models

An extremely important process in model building is assessing the model's prognostic competence. An important feature of this prognostic competence is discrimination, which refers to the ability of a predictive model to correctly classify subjects for their actual outcomes. In order to compare the discriminating potential of the aforementioned risk prediction models, we utilize some pre-existing methodologies such as concordance index (c-index), time-dependent Recieving Operating Characteristic (ROC) curves, Area under the ROC curve (AUC), Integrated area under the ROC curve (IAUC) and Integrated Brier Score (IBS) statistics.

In a standard binary classification setting, the c-index, is equivalent to the area under the ROC curve and thus serves as an overall summary for the curve. However in survival settings, this relationship stands negated, since the ROC curves become time dependent and therefore the c-index and the AUC statistic also become time specific. Thus all the methodologies referred to above have to be formulated differently to deal with censored data in survival settings.

3.3.1 Time dependendent ROC curves

ROC curves are very popular in displaying the sensitivity and specificity of a diagnostic marker say x and a disease variable say D. In most of the cases the disease variables are time dependent and therefore the ROC curve must be made to vary with time. The disease status variable in survival analysis is often a time dependent variable where $D(u) = 1$ if an event has occurred prior to time u, and zero otherwise. A method proposed by Heagerty et al. (2000) summarizes the discriminating potential of a diagnostic marker x, by calculating

ROC curves for event incidence by time u. Alternatively it is possible to calculate a risk score (cumulative hazard, survival and mortality) R, from a survival scenario and calculate estimates of the specificities and sensitivities. Let us consider data in the form (t, δ, R) where t is the time to event, δ is the censoring indicator where $(\delta = 1)$ indicates that an event has occurred whereas $(\delta = 0)$ indicates that the observed time was right censored and R forms the vector of risk set predictions. We define $D(u) = 1$ if $t \leq u$ and $D(u) = 0$ if $t > u$, with $D(u) = 1$ indicating that the event has occurred prior to time u. At given time u, and a cut-off value c, we can define

$$sensitivity(c, u) = Pr(R > c|D(u) = 1) = S_n(c, u) \tag{3.20}$$

$$specificity(c, u) = Pr(R \leq c|D(u) = 1) = S_p(c, u) \tag{3.21}$$

The ROC curve at time u is defined as

$$ROC(c, u) = S_n(c)[1 - S_p^{-1}(c, u)] \tag{3.22}$$

This definition is often referred to as "cumulative/dynamic" ROC curve in literature. "Cumulative" means that all the events that occur before time u are considered as "cases". Other definitions of ROC curves can also be found in Heagerty and Zheng (2005).
The AUC statistic at time u is defined as the area under the ROC curve at time u:

$$AUC(c, u) = \int ROC(c, u)du \tag{3.23}$$

There are several available methods like Inverse Probability of Censoring Weights (IPCW) (Uno et al., 2007), Conditional Kaplan-Meier (Heagerty et al., 2000), Nearest Neighbour Estimator (NNE) (Heagerty et al., 2000) and Recursive method (Chambless and Diao, 2006)

for estimating the time dependent ROC curves. For our analysis we have used the IPCW method by Uno et al. (2007). Unos estimators, based on IPCW method, do not assume a specific working model for deriving the risk predictor. In addition to the non-informative censoring that the other methods assume, the IPCW method also assumes that censoring occurs independently of all the covariates.

Inverse Probability of Censoring Weights Approach. Let T be a continuous failure time, C be the corresponding censoring variable and $\tilde{x}$ be the vector of predictors. We assume that T and C are independent and that the survival function G(.) of C is free of $\tilde{x}$. For the i-th individual ($1 \leq i \leq n$) in a sample let $(t_i, \delta_i, \tilde{x_i})$ be the observed survival time ($t_i = \min(T_i, C_i)$), censoring indicator ($\delta_i = I(t_i = T_i)$) where I is the indicator function) and the covariate vector respectively. Based on this data we are interested in formulating a rule that can correctly predict whether or not the survival time T^0 of a future subject with $\tilde{x} = \tilde{x}^0$ is shorter than u-year, where u is a specified time point and $P(t > u) > 0$. Now let x, a function of $\tilde{x}$, be a p dimensional vector with the first component being 1, and consider the following model

$$P(T \leq u|x) = g(\beta' x) \tag{3.24}$$

where g(.) is a known strictly increasing differentiable function and β is a p-dimensional vector of unknown parameters. In particular if $g(.)$ is $1 - exp(-exp(.))$, then it is the CPH model. Now we suppose that β is estimated by $\hat{\beta}$, based on (t_i, δ_i, x_i) and $\hat{G(.)}$ is the KM estimate of the censoring distribution(assuming no covariates). To evaluate a prediction rule for a binary outcome we consider its specificity and sensitivity under a specific threshold value c. For the prediction rule $I(g(\hat{\beta}' x) > c)$ the sensitivity is $S_n(c, u) = P(g(\beta' x) > c|T \leq u)$ and the specificity is $S_p(c, u) = P(g(\beta' x) \leq c|T > u)$. These conditional probabilities are estimated consistently by

$$S_n(c,u) = \frac{\sum_{i=1}^{n} \delta_i I(g(\hat{\beta}' x_i) > c, t_i \leq u)/\hat{G(t_i)}}{\sum_{i=1}^{n} \delta_i I(t_i \leq u)/\hat{G(t_i)}} \tag{3.25}$$

$$S_p(c,u) = \frac{\sum_{i=1}^{n} I(g(\hat{\beta}' x_i) \leq c, t_i \geq u)}{\sum_{i=1}^{n} I(t_i > u)} \tag{3.26}$$

ROC(c,u) can be estimated by substituting these estimated sensitivities and specificities. The estimated AUC(c,u) is calculated by using the trapezoidal rule to integrate the estimated ROC(c,u) curve. The IAUC summary measure is given by the integral of AUC on $(0, max(t_i))$ (weighted by the estimated probability density of the time-to-event outcome).

Therefore the discriminating potential of the several modeling approaches discussed earlier can be measured by plotting the time dependent ROC curve. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity and 100% specificity. The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line (the line of no-discrimination) from the left bottom to the top right corners (regardless of the positive and negative base rates). The area under the ROC curve is equal to the probability that a prognostic survival model will predict a high survival probability (or a lower hazard rate and mortality) for a subject having a high observed survival time and a lower survival probability (or a higher hazard rate and mortality) for a person with a shorter observed survival time. Thus a model with a higher prognostic competence will have a higher AUC value at time t and a higher IAUC value .

3.3.2 Integrated Brier Score

The Brier score (Brier, 1950) is a quadratic score function, calculating the squared differences between actual binary outcomes Y and its predictions p. In survival settings, let $Y_i(u) = 1 - D_i(u)$ be the observed status of the subject i and $\hat{S}(u|x_i)$ be the predicted survival probability at time u for subject i with predictor variables x_i. Then the time dependent expected Brier score is given as

$$BS(u, \hat{S}) = E(Y_i(u) - \hat{S}(u|x_i))^2 \tag{3.27}$$

where the expectation is taken with respect to the data of a subject i belonging to the test set. In censored survival outcomes, the squared residuals are weighted using inverse probability of censoring weights (Gerds and Schumacher, 2006) given by

$$\hat{W_i}(u) = \frac{(1 - Y_i(u))\delta_i}{\hat{G}(t_i - |x_i)} + \frac{Y_i(u)}{\hat{G}(u|x_i)} \tag{3.28}$$

where $\hat{G(}.)$ is the KM estimate of the censoring distribution (assuming no covariates). If an independent test dataset K_M is considered then the expected Brier score (Mogensen et al., 2012) is estimated by

$$BS(u, \hat{S}) = \frac{1}{M} \sum_{i \in K_M} \hat{W}_i(u)(Y_i(u) - \hat{S}(u|x_i))^2 \tag{3.29}$$

where M is the number of subjects in K_M and $\hat{S}$ is based on the training data.

Finally the Integrated Brier Score (Mogensen et al., 2012) can be calculated as

$$IBS(BS, \tau) = \frac{1}{\tau} \int_0^{\tau} BS(b, \hat{S}) db \tag{3.30}$$

where $\tau > 0$ can be set to any value smaller than the maximum time of the test sample. The IBS ranges from 0 to 1; the smaller the score, the better the fit. Useful benchmark values for the IBS are 33%, which corresponds to predicting the risk by random number drawn from an Uniform (0,1) distribution, and 25% which corresponds to predicting 50% risk for everyone. The most important benchmark is the IBS of a prediction model without any covariates. In survival analysis the Kaplan-Meier estimate of survival calculated with all training samples yields such a null model.

3.3.3 Concordance Index

The concordance index (c-index) by (Harrell et al. (1982), Harrell et al. (1996)) can be interpreted as a global discrimination index for understanding the predictive competence of a survival model. It is similar to a rank correlation measure between the predicted and observed survival times. In calculating the c-index, we consider all possible pairs of patients (i, j) with survival times (t_i, t_j) and censoring indicator (δ_i, δ_j). We delete pairs (i, j) for which $(t_i < t_j)$ and $(\delta_i = 0)$ or $(t_j < t_i)$ and $(\delta_j = 0)$ or $(t_i = t_j)$ and both ($\delta_i = 0$ and $\delta_j = 0$). The remaining pairs can be referred to as the utilizable pairs. For each utilizable pair $(t_i \neq t_j)$ we count 1, if the predicted outcome is worse for the patient with the shorter observed survival time and count 0.5, if $(t_i = t_j)$. For each utilizable tied time pair $(t_i = t_j)$ if both ($\delta_i = 1$ and $\delta_j = 1$), we count 1, if the predicted outcomes are tied and count 0.5; otherwise. For each utilizable tied time pair $(t_i = t_j)$ when not both ($\delta_i = 1$ and $\delta_j = 1$)

we count 1, if the patient who died has a worse predicted outcome and count 0.5; otherwise. We sum over the count values calculated using all the utilizable pairs to reach at the total number of concordant pairs. The c-index can then be calculated as a proportion of all the utilizable patient pairs that are concordant. Thus the c-index estimates the agreement probability between the observed and predicted survival outcomes. The close the value of the c-index is to 1, the better is the fit. A c-index value of 0.5 indicates no prognostic discerning capacity between patients with different outcomes, whereas a value of 1, indicates perfect prognostic discernment. The confidence interval for the c-index can be calculated using the concordance.index function found in the R package survcomp.

CHAPTER 4

SIMULATION STUDY

4.1 Performance in one-sample studies

Event times were generated from a Weibull survival curve , as described in 2.9, with $\alpha = 0.8$ and $\lambda = 2.5$. Censoring times were generated independently from an expoential distribution with parameters selected to induce 20% and 50% censoring respectively.Samples of sizes 50 and 200 were considered and for each sample size, 100 datasets were generated. We calculated the 95% confidence intervals for the posterior survival distribution obtained using BART and compared them with the 95% confidence intervals obtained using the KM estimator(using log transformation of the confidence intervals). For each sample size and censoring percentage performance of both the methods are compared based on coverage, accuracy and bias. The root mean square error, coverage and width of the confidence intervals and bias are calculated for both the BART and KM estimates.The expected value of the simulated estimates is then used as a measure of performance. The performance measures used for comparison and their formulas have been shown in Table 4.1. The posterior intervals obtained from BART have very good coverage probabilities as compared to the KM estimator. The BART model's root mean square error is slightly lower than the KM estimate when a larger sample size is considered. The bias of the BART model estimate, very close to zero, is slightly higher than that of the KM estimate. The detailed results by sample size and censoring percentages have been summarized in Table 4.1 and boxplot summaries of the coverage probability, bias and root mean square error have been shown in Figure. From the

results it is evident that BART performs as well as the KM estimator in fitting a survival function.

Table 4.1: Measures to evaluate the performance of different methods

Bias	$\delta = \bar{\hat{\theta}} - \theta$
Root Mean Square Error (RMSE)	$\sqrt{(\bar{\hat{\theta}} - \theta)^2 + SE(\hat{\theta})^2}$
Coverage	Proportion of times the $100(1-\alpha)\%$ confidence interval $\hat{\theta}_i \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\theta}_i$ include θ for $i = i, ..., B$

Key: θ is the true value of the estimate of interest,$\bar{\hat{\theta}} - \theta = \sum_{i=i}^{B} \frac{\hat{\theta}_i}{B}$ where B is the number of simulations performed $\hat{\theta}_i$ is the estimate of interest calculated from each of the $i = 1, ..., B$ simulations. $SE(\hat{\theta})$ is the empirical standard error of the estimate of interest over all simulations,$SE(\hat{\theta}_i)$ is the standard error of the estimate of interest calculated for each simulation and $Z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution.

Table 4.2: Comparison of results between KM and BART in a one sample study

Size of the dataset	Censoring	Method	Mean coverage probability of 95 % CI at 50th percentile	Mean RMSE of estimates at 50th percentile	Mean Bias of estimates at 50th percentile
50	20%	KM	0.9588	0.0674	−0.0017
50	20%	BART	0.9155	0.0671	−0.0036
50	50%	KM	0.9583	0.0636	−0.0157
50	50%	BART	0.9234	0.0624	−0.0184
200	20%	KM	0.9658	0.0575	−0.0043
200	20%	BART	0.9250	0.0558	−0.0089
200	50%	KM	0.9666	0.0355	−0.0132
200	50%	BART	0.9383	0.0348	−0.0144

4.2 Performance in regression scenarios with and without proportional hazards

While the performance of the BART method is favorable with respect to the KM method in one sample studies, its efficiency in real life applications lies in more complex regression scenarios. Semiparametric methods such as the CPH model and the other parametric survival models aims to model a particular functional relationship between the covariates and some survival outcomes. However BART and RSF offers a more flexible approach allowing nonparametric functional relationships. In this subsection, we demonstrate such ability of the BART method via two simulation studies designed in Sparapani et al. (2016)

Two simulation studies are used, one having proportional hazards and the other having non proportional hazards. It is presumed that the model having non-proportional hazards should pose significant challenges to the semi parametric CPH model. Nine independent binary covariates $x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$ are generated from the Bernoulli distribution with probability 0.5. They are then related to the Weibull event time t with survival function through Equation 2.9, using different rate and scale parameters for the proportional and non proportional hazards model. For the proportional hazards model

$$\alpha = 2, \lambda = exp(3 + (0.1(x_1 + x_2 + x_3 + x_4 + x_5 + x_6) + x_7) \tag{4.1}$$

For the non proportional hazards model

$$\alpha = 0.7 + 1.3x_7, \lambda = 20 + 5(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + 10x_7) \tag{4.2}$$

Censoring times were generated independently from an exponential distribution with parameters selected to induce 20% censoring. Samples of size 200 were considered and 100

datasets were generated, for each of the models. Performances of the CPH, RSF and BART methods were compared based on measures of accuracy and bias for both the model settings. The root mean square error and bias are calculated for the CPH, RSF and BART survival prediction estimates. The expected value of the simulated estimates is then used as a measure of performance. The performance measures used for comparison and their formulas have been shown in Table. Figure 4.1 and Figure 4.3 shows box plots of bias and RMSE for the proportional hazards model and Figure 4.2 and Figure 4.4 shows box plots of bias and RMSE for the non proportional hazards model measured at the 25th, 50th and 75th percentiles of the overall survival distribution. It is clear from these plots that the BART method performs closely to the CPH and RSF models in the proportional hazards case, however non parametric methods of BART and RSF performs significantly better than the semi parametric CPH model in the non-proportional hazards scenario.

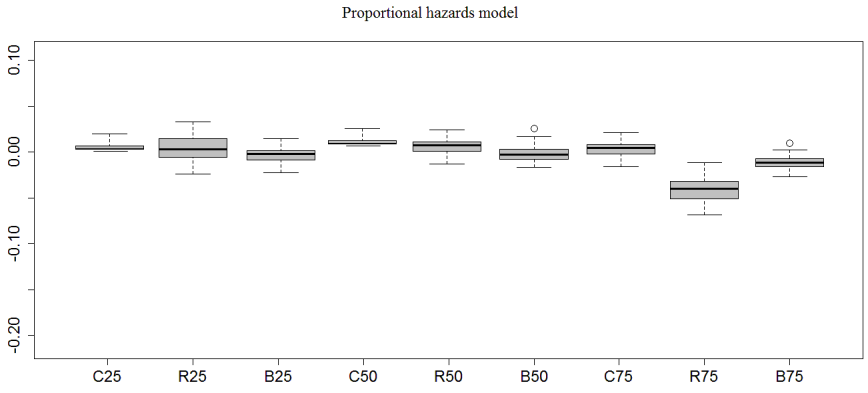

Figure 4.1: Box plots for bias for a PH model

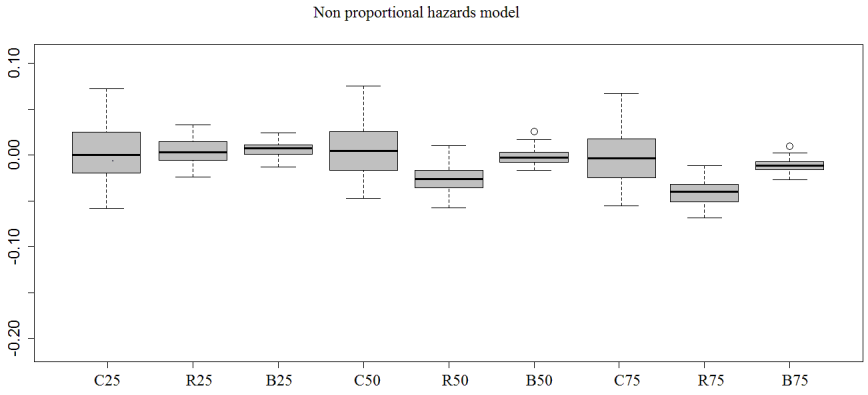

Figure 4.2: Box plots for bias for a NPH model

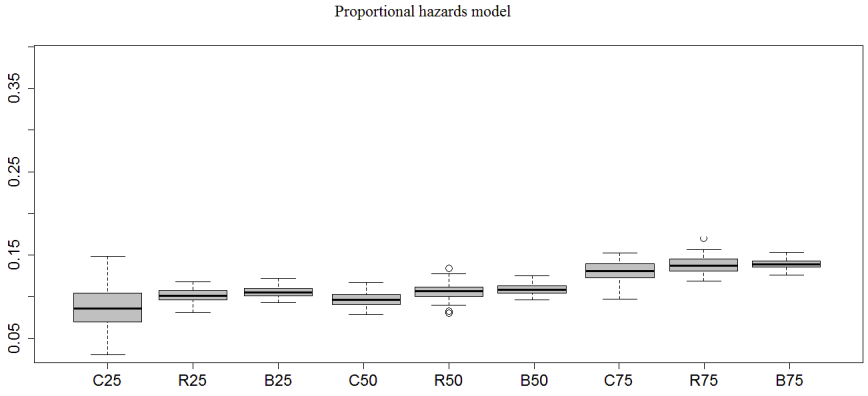

Figure 4.3: Box plots for RMSE for a PH model

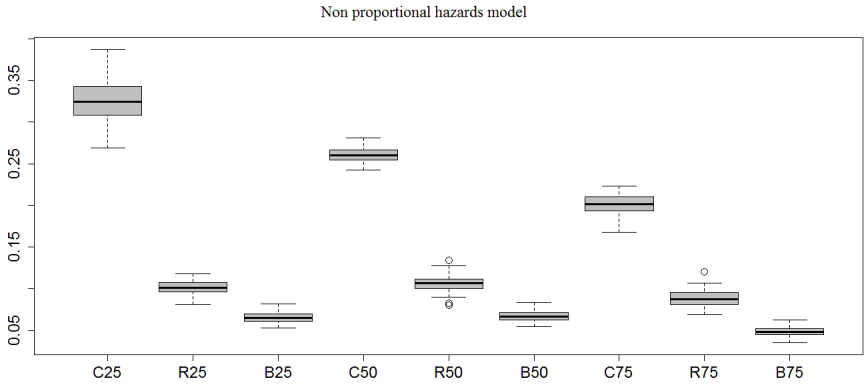

Figure 4.4: Box plots for RMSE for a NPH model

CHAPTER 5

APPLICATION

In this section, we have tried to apply the CPH, RSF and BART models to a random sample of 1500 female patients between ages 24-90 having invasive ductal carcinoma as obtained from the U.S. SEER database for the year 2005. Samples with missing data were not incorporated in the study, to facilitate the demonstration of methods. Ten covariates were considered in the analysis namely Age (in years), Race (White, Black, Others), disease stage (In-situ, localized, regional, or distant), tumor grade (well-differentiated, moderately differentiated, poorly differentiated, or undifferentiated), tumor size (in cms), estrogen receptor status (postive, negative or borderline), progesterone receptor status (positive, negative or borderline), radiotherapy (received or denied), surgery (received or denied) and the number of lymph nodes. All the covariates selected for the analysis are important in breast cancer studies. While most of them are self explanatory, some of them require necessary elucidation. Stage of the tumor determines whether the cancer is limited to one area in the breast, or it has spread to healthy tissues inside the breast or to other parts of the body. In disease stage in-situ, the atypical cells have not spread outside of the ducts or lobules into the surrounding breast tissue. In Stage localized, cancer is evident, but it is contained to only the area where the first abnormal cells began to develop. In stage regional, the cancer is extensive in the underarm lymph nodes, or has spread to other lymph nodes or tissues near the breast and finally in stage distant, the cancer has spread beyond the breast, underarm and internal mammary lymph nodes to other parts of the body near to or distant from the breast. Cancers grade determines how different the tumor cells are from normal breast cells and how quickly they are growing. In cancer grade 1 (well differentiated), the cancer cells look most

like normal cells and are usually slow-growing, for grade 2 (moderately differentiated) the cancer cells look less like normal cells are growing faster, for grade 3 (poorly differentiated), the cancer cells look most changed and are usually fast-growing and for grade 4 the tumor tends to grow and spread at the highest rate. Normal breast cells and some breast cancer cells have receptors, that attach to estrogen and progesterone and depend on these hormones to grow. Breast cancer cells may have neither, one, or both of these receptors. ER-positive cancers are breast cancers that have estrogen receptors. PR-positive cancers are breast cancers with progesterone receptors. Certain drugs are used to treat breast cancers that have one or both of these receptors. Most types of hormone therapy for breast cancer either lower estrogen levels or stop estrogen from acting on breast cancer cells. This kind of treatment is helpful for hormone receptor-positive breast cancers, but it doesnt work on tumors that are hormone receptor-negative(both ER- and PR-negative). Finally our response variable for the study was disease-specific survival (in months) based on the SEER cause-of-death code. Death from other causes was treated as censoring (non-informative censoring). The censoring times were assumed to be independent of the failure times. For evaluating the performance of the three models, the dataset was split randomly in the ratio 2:1 into a training set and a validation set.

5.1 Exploratory Analysis

Training Sample. Of the 1000 patient cases, 721 patients were white, 202 black, and the rest were people from other different origins. A total of 796 deaths occurred in the cohort of 1000 patients. The number of survival months (our outcome of study) ranges from 1-106 months. The mean follow up time was 34.75 months and the median follow-up months was 30 months. Most of the tumors were staged regionally (38.9%). Most tumors were graded as

poorly differentiated (56.5%). The mean age was 60.33 years with an SD of 14.59 years. The median tumor size was 29 mm with an IQR of 33 mm. 65.4% of the tumors were estrogen positive. 43.7 % of the patients recieved both surgery and radiation.

Test Sample. Out of the 500 patient cases, 360 were white (72%), 96 black (19.2%), and the rest were from other different origins. A total of 395 deaths (79.5%) occurred in the cohort of 500 patients. The number of survival months (our outcome of study) ranges from 1-106 months. The mean follow up time was 35.21 months and the median follow-up months was 30 months. Most of the tumors were staged regionally (36.8%). Most tumors were graded as poorly differentiated (55.5%). The mean age was 61.26 years with an SD of 14.73 years. The median tumor size was 28 mm with an IQR of 35 mm. 60.6% of the tumors were estrogen positive. 45.8 % of the patients recieved both surgery and radiation.

5.2 Cox Proportional Hazards Regression Analysis

Based on the p values obtained in Table 5.1, covariates, number of lymph nodes and tumor grade does not impact overall survival. Considering reference categories, borderline estrogen and progesterone status and tumor grade does not seem to impact overall survival as well (Table 5.1). For the remaining significant covariates, $\exp(\beta)$ can be interpreted as a multiplicative effect on the hazard of death. For example, holding all covariates constant, an additional year of age increases the monthly hazard of death by a factor of 1% ($\exp(\beta)$=1.01). Similarly a black patient has an increased hazard of 52% compared to its reference category (Table 5.1). On application of the backward variable selection mechanism described in subsection 3.2.1, race of the patient, age at diagnosis, tumor grade, tumor size, tumor stage, radiation therapy, surgery and estrogen receptor status are chosen as the most important variables. However, the accuracy with which the model estimates the hazard ratios depends

upon the assumption of proportional differences between the hazards of different groups of the covariates, which implies that the covariates do not vary over time. So in order to determine the correctness of the hazard ratios obtained in (Table 5.1) we need to check whether our dataset satisfies the PH assumption. For our training dataset the PH assumption is violated for variables age, radiation therapy, number of lymph nodes and er status. Also since the p value for the global index is significant, the entire model is assumed to violate the PH assumption (Table 5.2). This can be easily verified from the plot of the Cox-Snell residuals (Cox and Snell, 1968) shown in Figure 5.2 which deviates widely from the 45 degree line, the expected line of perfect fit. Since the covariates in the model donot satisfy the PH assumption, the general CPH model does not seem to be a very good fit for our dataset.

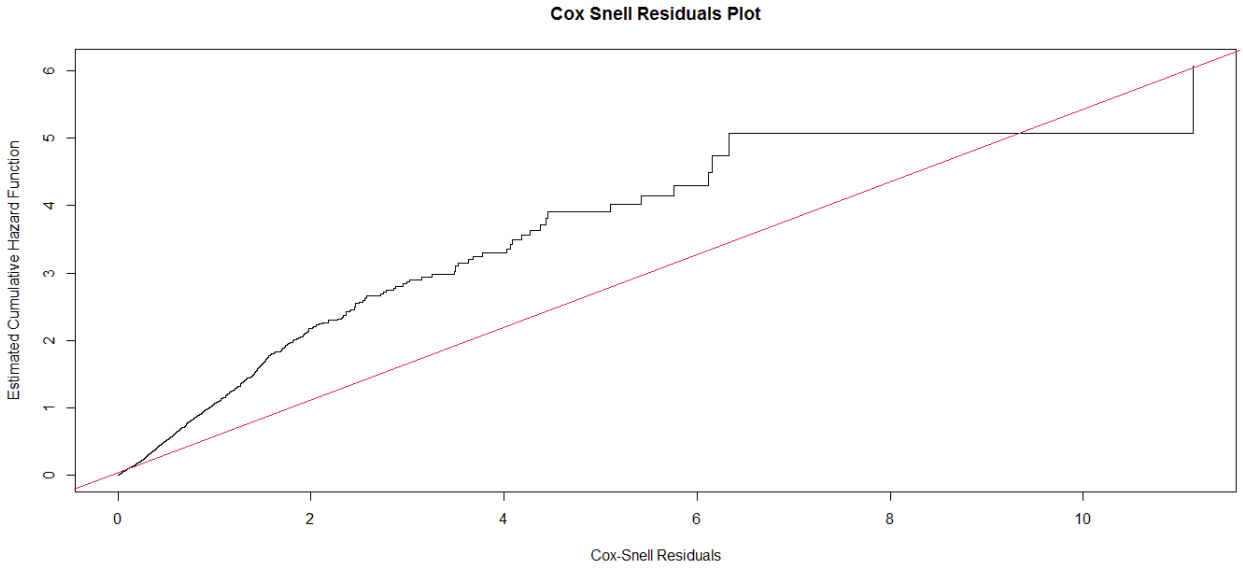

Figure 5.1: Cox Snell residuals plot for the CPH model

Table 5.1: Results of the CPH model

Covariate	β	$\exp(\beta)$	SE(β)	Z-value	p value	95% CI
Race						
Others	-	1.000	–	–	–	–
White	0.222	1.249	0.145	1.531	0.125	$0.940 - 1.660$
Black	0.010	1.528	0.160	2.653	< 0.01	$1.117 - 2.090$
Age	0.013	1.013	0.003	4.902	< 0.01	$1.008 - 1.018$
Grade						
Undifferentiated	-	1.000	–	–	–	–
Well-differentiated	-0.881	0.414	0.365	-2.412	0.015	$0.203 - 0.848$
Moderately-differentiated	-0.596	0.551	0.341	-1.746	0.081	$0.282 - 1.076$
Poorly-differentiated	-0.292	0.748	0.336	-0.871	0.384	$0.387 - 1.441$
Surgery						
No-Surgery	-	1.000	–	–	–	–
Surgery	-0.866	0.421	0.105	-8.221	< 0.01	$0.342 - 0.517$
Radiation						
No-Radiation	-	1.000	–	–	–	–
Radiation	-0.372	0.689	0.074	-5.037	< 0.01	$0.596 - 0.797$
Tumor Size	0.006	1.006	0.001	4.636	< 0.01	$1.003 - 1.008$
Stage						
In-situ	-	1.000	–	–	–	–
Localized	1.454	4.279	0.432	3.362	< 0.01	$1.834 - 9.986$
Regional	1.884	6.578	0.433	4.356	< 0.01	$2.818 - 15.353$
Distant	2.319	10.170	0.440	5.273	< 0.01	$4.294 - 24.086$
Lymph Nodes	0.00	1.00	0.0001	-0.048	0.962	$0.999 - 1.000$
estrogen-receptor status						
Negative	-	1.000	–	–	–	–
Borderline	1.089	2.972	0.602	1.807	0.071	$0.912 - 9.688$
Positive	-0.674	0.510	0.131	-5.961	< 0.01	$0.408 - 0.636$
progesterone-receptor status						
Negative	-	1.000	–	–	–	–
Borderline	0.624	1.867	0.372	1.677	0.093	$0.900 - 3.872$
Positive	-0.036	0.964	0.104	-0.351	0.725	$0.787 - 1.182$

Table 5.2: Assessing the PH assumption

Covariate	rho	chisq	p-value
Race	−0.009	0.077	0.781
Age	−0.105	9.898	< 0.01
Grade	0.065	3.118	0.07
Surgery	0.054	2.404	0.121
Radiation	0.105	8.855	< 0.01
Tumor size	−0.061	2.797	0.09
Stage	−0.079	5.333	0.029
Lymph-nodes	0.092	7.641	< 0.01
er status	−0.095	8.077	< 0.01
pr status	0.003	0.009	0.924
Global	−	51.192	< 0.01

5.3 Random Survival Forest Analysis

Figure 5.2, presents a graphical output of how the RSF models using logrank splitting and logrankscore splitting, ranks its covariates by level of OOB-importance, based on 1000 trees as described in subsubsection 3.2.2. The five most important covariates in both the RSF approaches are surgery, tumor size, tumor stage, tumor grade and er status with a slightly different ranking. The bottom five covariates based on importance values are ranked similarly for both the RSF models. The top five predictors having maximum variable importance values were also selected by the Cox model. However predictors race, age at diagnosis and radiation therapy, selected by the Cox model, have very low variable importance values for both the RSF models and are therefore considered unimportant for prediction purposes. The individual impact of the covariates Stage, Grade, Surgery and Radiation therapy on mortality can be seen in Figures 5.3 and 5.4. The plots for the rest of the covariates can be found in the Appendix (Figures A.1 ,A.2, A.3) . For the tumor stage variable, localized, regional and distant tumors have greater estimated mortality, in that order, with reference to insitu

staged tumors, which is the same sentiment as expressed by the hazard ratio estimates of the CPH model. Also patients not receiving surgery or radiation therapy are exposed to greater risks as compared to patients receiving surgery or radiation therapy. This result also matches with the CPH model results since patients recieving surgery have a 57.1% less hazard rate as compared to patients not receiving surgery and patients receiving radiation therapy have a 31.1% less hazard rate as compared to patients not receiving radiation therapy.

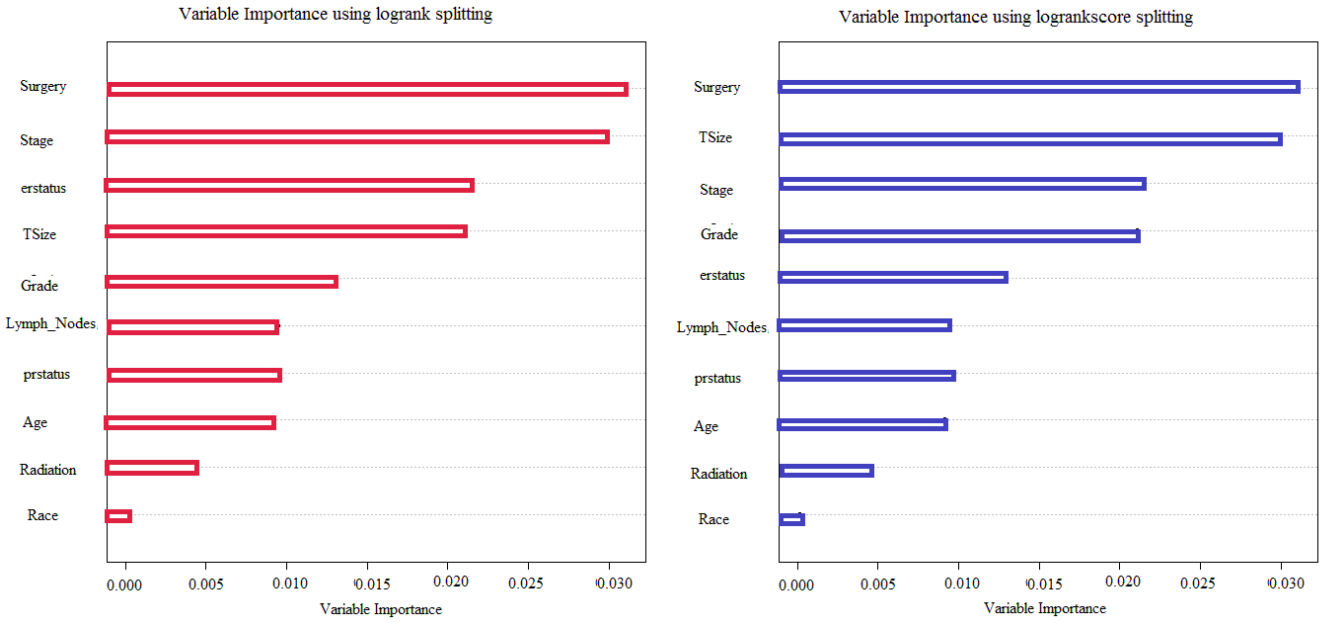

Figure 5.2: Variable Importance plots using "logrank" and "logrankscore" splitting.

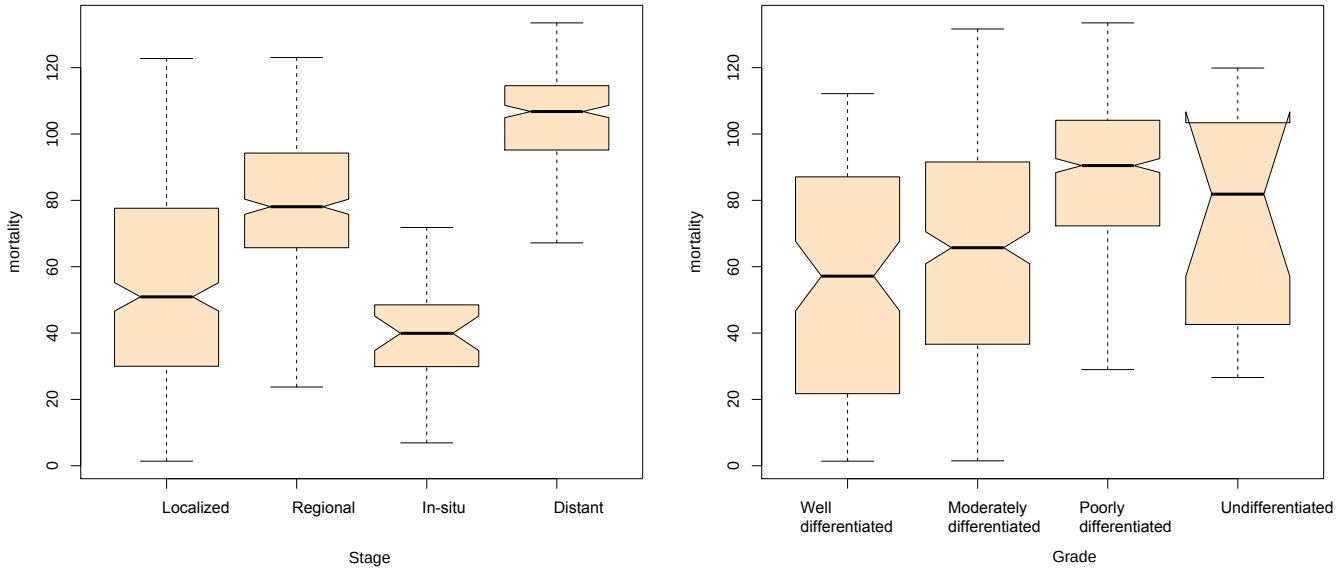

Figure 5.3: Plots of the marginal effect of covariates Stage and Grade on estimated mortality

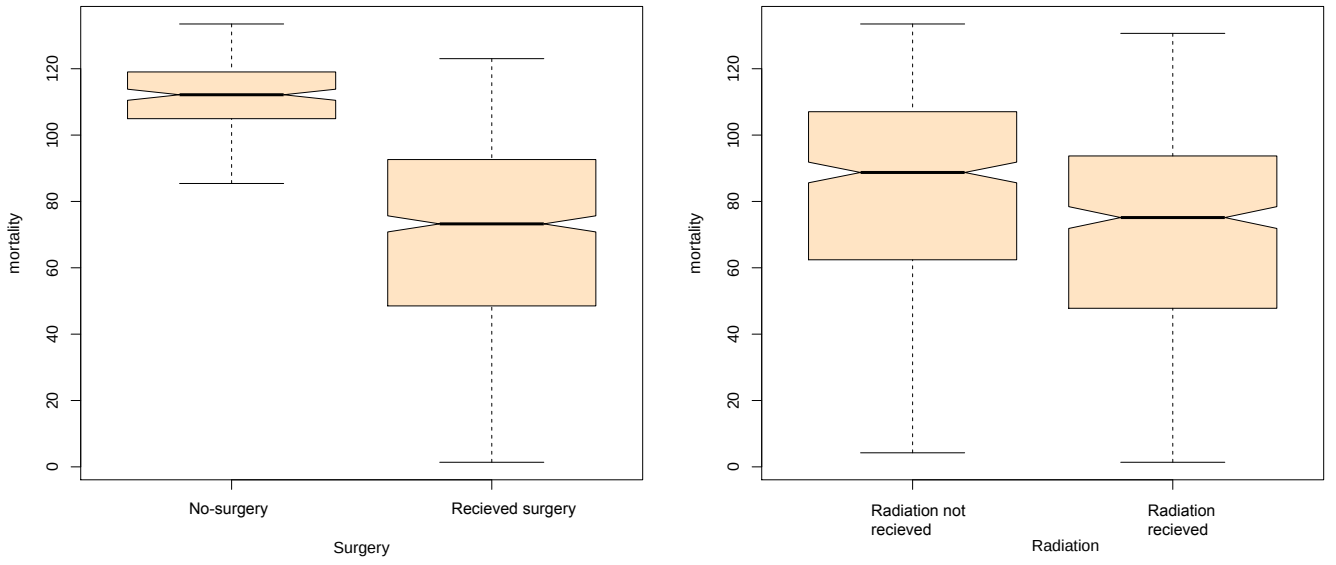

Figure 5.4: Plots of the marginal effect of covariates Surgery and Radiation on estimated mortality

5.4 BART Model Analysis

The BART survival model was fit to the training dataset with 50 trees in the sum and the default prior, having a burn-in of 1000 draws and 7000 draws from the posterior distribution after the burn-in, for estimating the survival function given the predictors. We obtained the partial dependence survival functions using equation for a particular subcategory of predictors. These functions can be explained as a marginal survival function for a single predictor level, averaged across the distribution of the remaining predictors. In Figure 5.5 and Figure 5.6 we have plotted the partial dependence functions for four different tumor sizes and five different ages. From Figure 5.5 it can be seen that survival probability drops rapidly with an increase in tumor size. For example the five year survival probability for a patient with tumor size 20 mm is 0.93 as compared to 0.71 for a patient with tumor size 120 mm. From Figure 5.6 it can be seen that the 5 year survival probability for a 50 year old breast cancer patient is 0.87 compared to 0.80 for a 70 year old patient. These results have already been similarly expressed by the CPH model which predicted a 0.6% increase in monthly hazard of death for an additional mm increase in tumor size and a 1% increase in monthly hazard of death for an additional year of age.

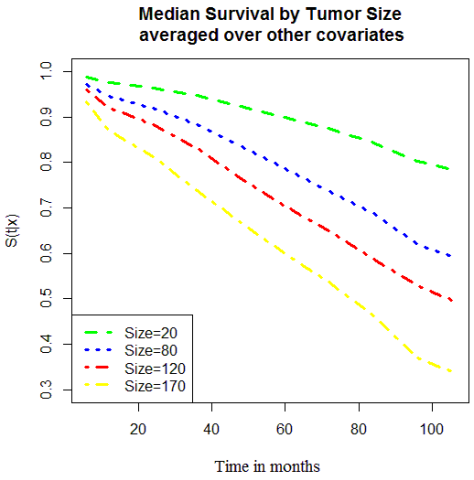

Figure 5.5: Marginal median survival function for Tumor Size

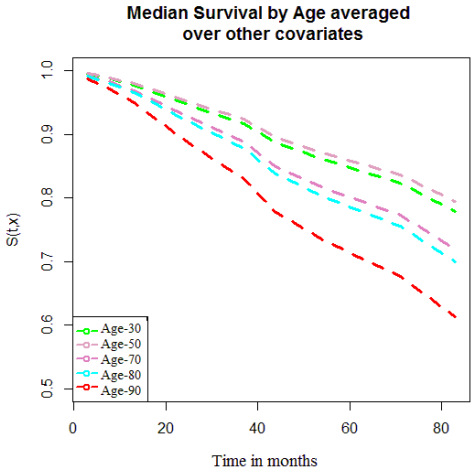

Figure 5.6: Marginal median survival function for Age

The BART survival model can also be used to study the effect of interactions between covariates on the survival outcome. Figure 5.7 studies the effect of the interaction between estrogen and progesterone receptor status on the survival probability. It can be seen from the plot that er and pr positive breast cancer patients have a higher survival probability as compared to er and pr negative patients. Since Age and Stage variables were selected by both the CPH and RSF models we wanted to check whether there exists any interaction between them. There is no evidence of interaction as the plots in Figure 5.8 are nearly parallel, but there may be an indication of a nonlinear relationship between median survival probability and age, since the median survival dips slightly at age 70. BART can also be conveniently used, to draw inference on various aspects of the survival distribution (obtained by regressing on all or a subset of covariates), directly from the posterior samples . This property has been illustrated in some supplementary figures found in the Appendix (Figures A.6 , A.7, A.4 and A.5).

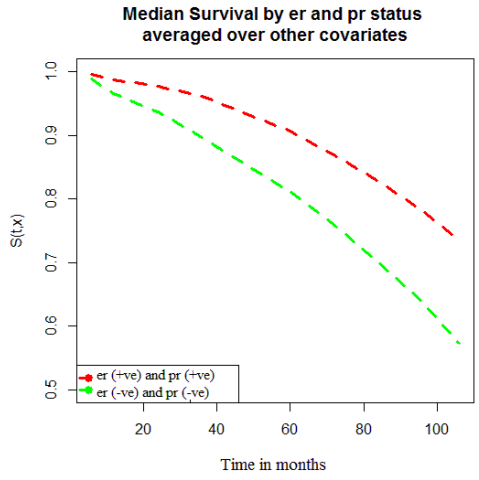

Figure 5.7: Marginal median survival function for Tumor Size

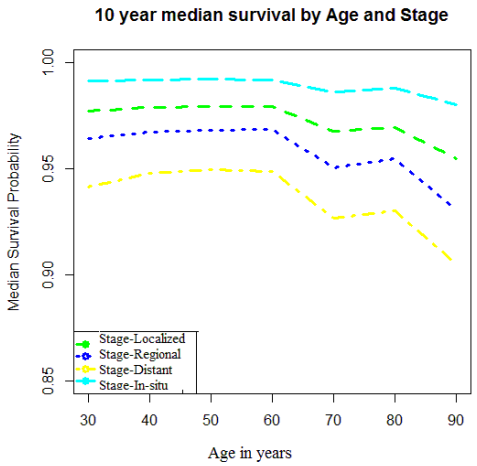

Figure 5.8: Marginal median survival function for Age

As another illustration on exploring significant interactions between the covariates, we explored the difference in the partial dependence survival function at five years between patients having Stage 1 tumor and Stage 4 tumor, separately by tumor size, age, tumor grade, surgery status, radiation status and estrogen receptor status. These variables were selected as they have been considered important by both the CPH and the RSF models. Results obtained are shown as a forest plot in Figure 5.9. One of the results indicated by the plot would be that surgery decreases the 5-year survival across the disease stages, although the magnitude of the effect may vary slightly.

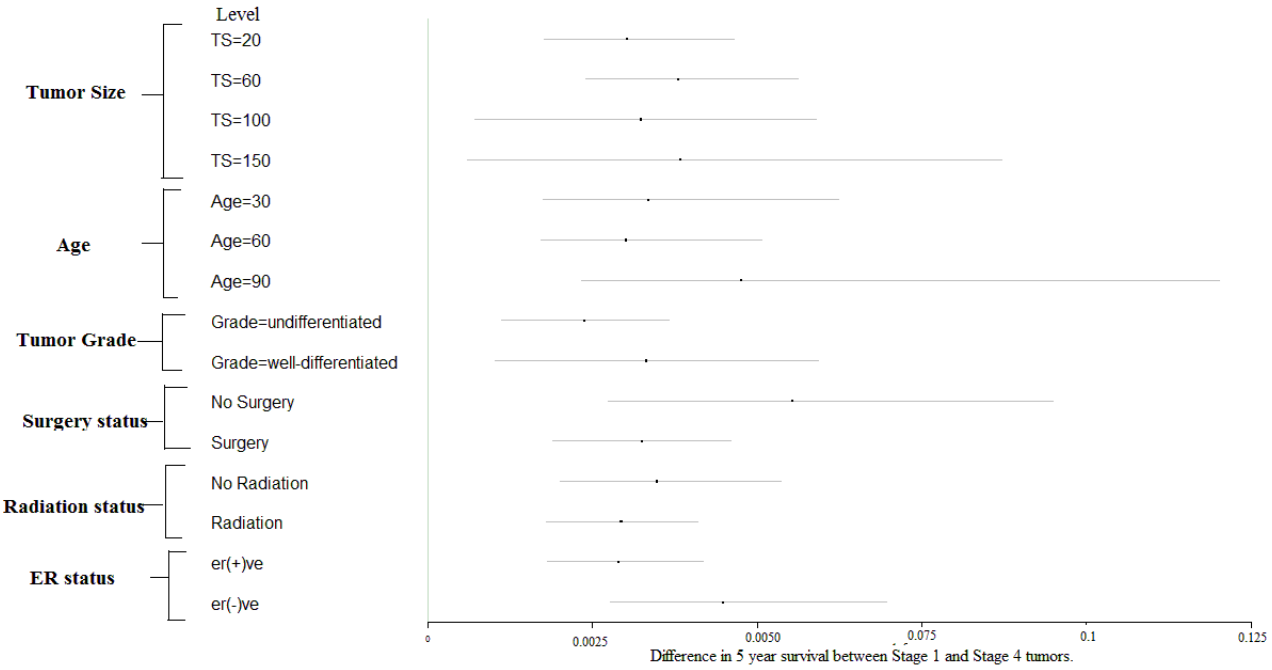

Figure 5.9: Forest plot of the difference in 5 year survival between Stage 1 and Stage 4 by several covariates.

Finally, we carried out variable selection as described in 3.2.3 by examining the average frequency per splitting rule for all 11 predictor variables (time post diagnosis plus 10 predictors), the model being run using different numbers of trees (m =100,m=50,m=20). As can be seen from Figure 3.2.3, covariate time naturally is the most selected covariate across the different number of trees. Besides time, the model identifies stage, tumor size, age, erstatus and surgery as the five most important covariates impacting overall survival .

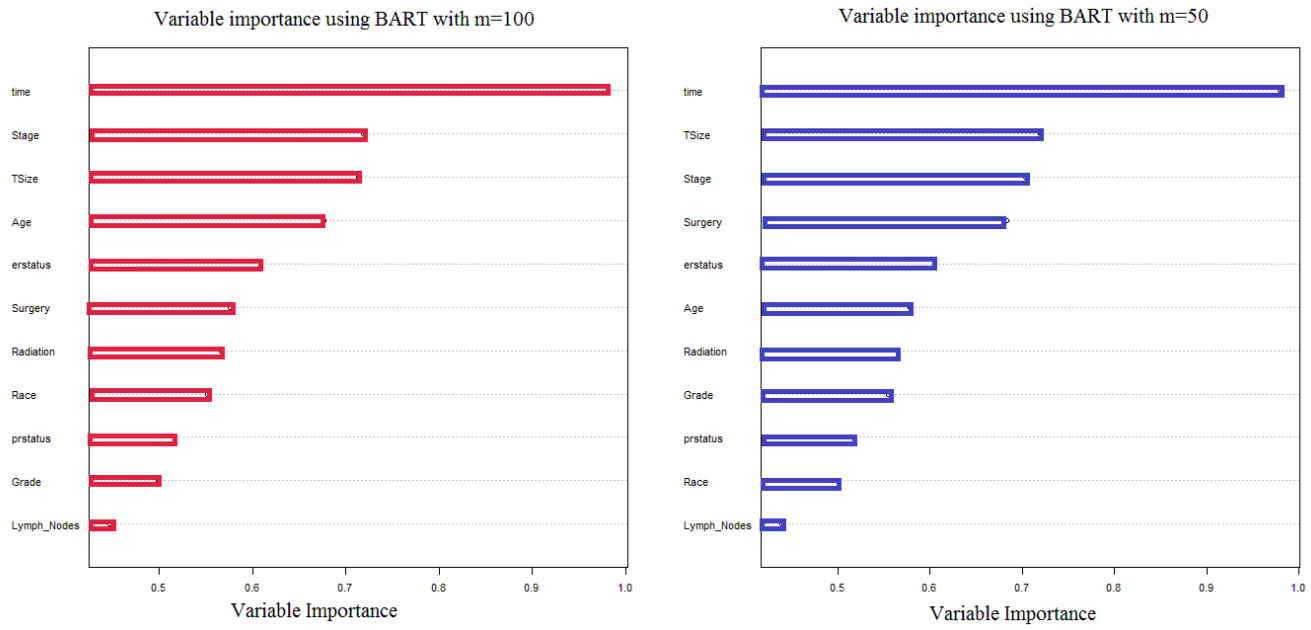

Figure 5.10: Variable importance using Bart with 100 and 50 trees.

5.5 Comparing model performances of CPH, RSF and BART

The RSF model using logrank splitting showed a better performance than all the other modelling approaches, with a Harrells Concordance error rate (1 minus c-index) (Section 3.3.3) of 0.144 for the training set. It was closely followed by the RSF model using logrankscore splitting, having an error rate of 0.175 (Table 5.3). For the test set the performance of the RSF model using logrank splitting was marginally better as compared to its counterparts (Table 5.4). Significant p-values for c-index calculated for all the models indicated that all the estimates were different from 0.5, implying greater capacity of predicting higher probabilities of survival for higher observed survival times.

Table 5.3: Performance for risk score prediction for training set

Model	c-index	SE(c-index)	95 % CI	p-value
CPH	0.730	0.092	0.711-0.749	0.000
BART	0.761	0.008	0.743-0.772	0.000
RF	0.856	0.006	0.843-0.868	0.000
RF-LS	0.825	0.007	0.810-0.839	0.000

Table 5.4: Performance for risk score prediction for test set.

Model	c-index	SE(c-index)	95 % CI	p-value
CPH	0.722	0.015	0.693-0.751	< 0.01
BART	0.702	0.014	0.672-0.736	< 0.01
RF	0.731	0.014	0.707-0.763	< 0.01
RF-LS	0.726	0.015	0.701-0.755	< 0.01

From the plotted time dependent ROC curves, we can see that over the first 12 (Figure 5.11) and 24 (Figure 5.12) months of follow up, the BART model has the highest AUC. At time=36 (Figure 5.13) and time=48 (Figure 5.14) months the RSF model with logrank splitting has the highest AUC . The substantiative interpretation for RF-AUC at time=12 months is: at any month, t, between 0-12 months the probability that a subject who dies on month t having a model score greater than a subject who survives beyond month t is 0.807.

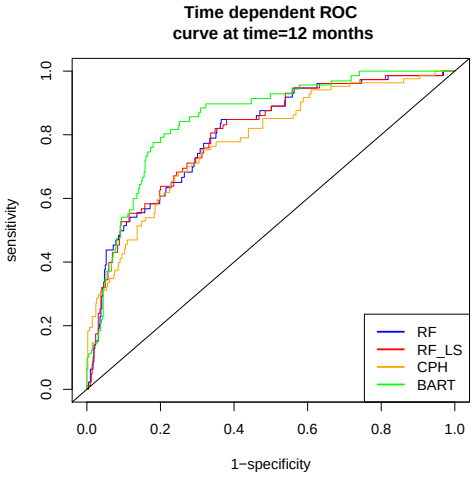

Figure 5.11: ROC curve at time = 12 months

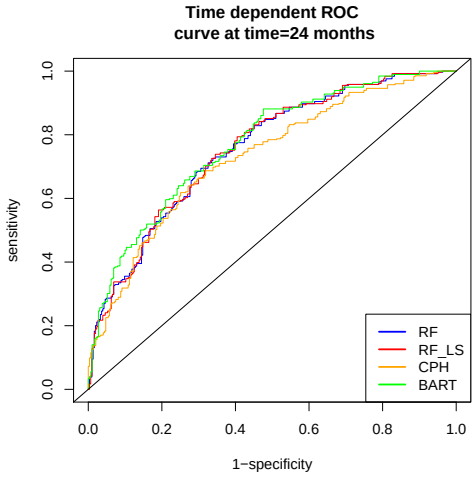

Figure 5.12: ROC curve at time = 24 months

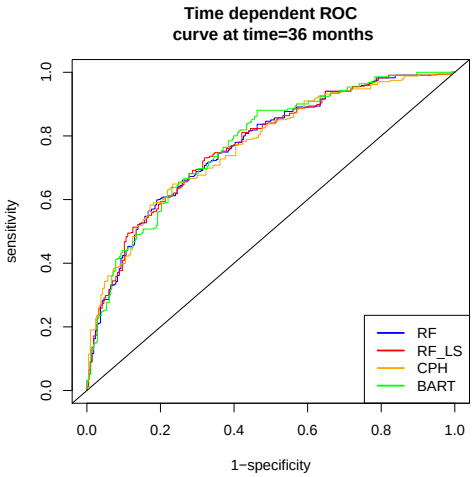

Figure 5.13: ROC curve at time = 36 months

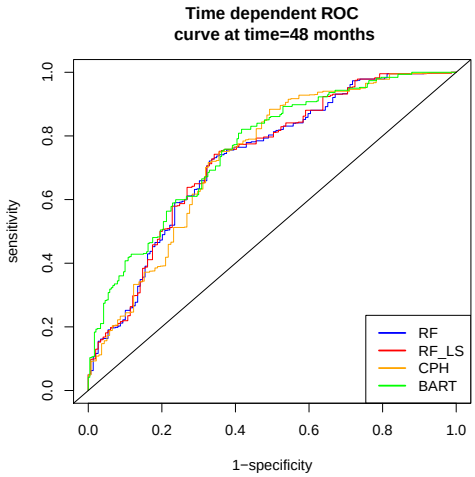

Figure 5.14: ROC curve at time = 48 months

The estimated AUC value for the RSF model with logrank splitting tends to decline over time to 0.785 for $12 < t \leq 24$ (Table 5.5). Thus the estimated AUC values suggests good short-term discriminatory potential of the model score. The model score has very good discriminatory capacity for distinguishing those patients who die at time t from those who

live beyond time t. This accuracy is especially good for follow-up times $24 < t \leq 36$ and $36 < t \leq 48$ months, with close AUC(t) estimates of 0.821 and 0.822 (Table 5.5). Estimates of AUC(t) also become increasingly variable over time due to the diminishing size of the risk set.

Table 5.5: Performance for AUC for the test set

Model	AUC			
	Time=12 months	Time=24 months	Time=36 months	Time=48 months
CPH	0.766	0.773	0.806	0.810
BART	0.839	0.789	0.811	0.813
RF	0.807	0.785	0.821	0.822
RF-LS	0.782	0.771	0.813	0.816

Using a follow up of 106 months yields a IAUC estimate of 0.839 for the RSF model with logrank splitting (Table 5.6). This implies that conditional on one event occurring within 106 months, the probability that the model score is larger for the subject with the smaller event time is 83.9%. The integrated Brier score values between 0 and 106 months, for the test set, are lowest for random survival forest with logrankscore splitting. The CPH, BART and RSF with logrank splitting models have approximately the same IBS values (Table 5.6). All three models perform substantially better than Kaplan-Meier having an IBS estimate of 0.150. Based on all these evaluation measures, it can be inferred that the BART method improves survival prediction accuracy in some cases or has comparable performance to the other two methods. We believe that this improvement is attributable to the method's flexibility in

Table 5.6: Performance assessment for risk prediction for each of the train and test set

Model	c-index		IAUC		IBS	
	Train	Test	Train	Test	Train	Test
CPH	0.730	0.722	0.850	0.839	0.112	0.119
BART	0.761	0.702	0.895	0.876	0.063	0.115
RF	0.856	0.731	0.893	0.852	0.065	0.113
RF-LS	0.825	0.726	0.889	0.849	0.065	0.112

modeling non-linear and additive covariate effects. For reference purposes a comparison of the survival probability curves, predicted using the three models under review, for certain covariate combinations, have been provided in the Appendix (Figure A.8 and A.9)

CHAPTER 6

CONCLUSION

The thesis focuses on the comparison of BART with the CPH and RSF models in analyzing survival data. It reviews three modeling approaches and compares them in terms of interpretative competence, prediction accuracy and variable selection methods. Simulation studies are performed to judge the competence of the BART estimator with the KM estimator in one sample studies and of the three aforementioned models in regression scenarios. The BART model has a similar performance to the KM estimator in terms of prediction accuracy and bias in one sample studies. In regression scenarios the three models perform closely when the proportional hazards assumption is met. However the performance of the CPH model depreciates with respect to the other two models when the proportional hazards assumption is violated.

We then apply the three models to analyze a real life breast cancer dataset. The covariates selected for our analysis of the breast cancer survival times fail to follow the proportional hazards assumption of the CPH model. Thus we apply RSF and BART to our dataset to deal with its complex structure and attain increased accuracy in predicting survival times. We chose BART because of its flexibility to accommodate high dimensional datasets and account for non-linearity and interactions present in covariates. Additionally working under a Bayesian paradigm allowed for natural quantification of uncertainty,that helped in construction of credible and prediction intervals. Thus, regressing on selected predictors we could estimate the median survival time and credible intervals, for a given patient, using the posterior distributions of the process parameters obtained using the BART model. Alternatively survival curves along with confidence bounds for the population could be plotted

using all or a subset of covariates. The RSF model was chosen as a competing method to the BART model because similar to BART, it is a decision tree structured black box model having high prediction accuracy and an efficient variable selection mechanism. Being black box models RSF and BART lack interpretative capacities. It cannot directly quantify the risks presented by individual covariates to the overall hazard like the CPH model does in terms of hazard ratios. However the partial dependence survival functions do give us an idea about how each of the covariates individually and jointly affect the overall survival risk. Additionally the performance of all the three models were compared using several assessment measures. BART's and RSF's comparable values of c-index, IAUC and IBSC further validates BART's predictive ability. BART's lack of interpretability as compared to the CPH model can thus be counterbalanced by its gains in prediction accuracy and the ability to incorporate complex interaction effects among the covariates.

Our primary motivation in using the breast cancer dataset was that we were more interested in identifying a statistical model that predicts overall survival effectively based on a set of covariates. We also wanted to understand the impact of these clinical covariates on the survival of breast cancer patients; and that was carried out successfully by the variable selection methods of BART. In addition, the variable selection procedure, partial dependence functions imbue the BART model with a high level of interpretability. We discovered important associations between stage of the tumor, tumor size, er status, surgery status and long-term survival using the BART model. The RSF model additionally considered tumor grade as important. Age at diagnosis was considered an important predictor by the CPH model. There are many studies which have estimated the risk factor importance of breast cancer using CPH and machine learning models. In line with our findings for the CPH model, Rosenberg et al. (2005) concluded that tumor size, tumor grade and race, all have significant constant effects on disease-specific survival in breast cancer, while the effects of age at diagnosis and disease stage have significant effects that donot follow the proportional

hazards assumption. Also similar to our results for the RSF and BART models D'Eredita et al. (2001) reported tumor size and tumor grade as the most informative medical factors using a RSF model and Delen et al. (2005) affirmed the effectiveness of tumor size and tumor stage through the sensitivity analysis of Artificial Neural Network. In contrast with our findings Omurlu et al. (2009) reported the importance of pr status and number of lymph nodes using a CPH and RSF model for analysis.

One of the serious disadvantages of BART in comparison to RSF was its computational time. BART is highly computationally demanding because the model requires expanding the data at a grid of event times. This problem is aggravated in case of large datasets. The authors mention using parallel processing and time scale coarsening as possible remedies. Both the BART and RSF models have been incorporated as R packages survbart and randomForestSRC respectively and the function computing the RSF model algorithm is a lot faster.

6.1 Future work

A lot of survival studies suffer from missing data problems. We have not addressed this problem in our study example, however there is an R package bartMachine which allows the user to directly handle missing covariate data within the BART framework. Reducing the computation time of BART for efficient analysis of large datasets is also a further challenge. In our study, we computed the performance assessment measures of our selected models based on a single validation dataset. There is a high chance that the results obtained could be strictly restricted to our chosen test set and could give absolutely dissimilar results for different test sets with varying levels of censoring. We dealt with different sample sizes and censoring rates in our simulation studies but haven't dealt with it sufficiently for our real data

application. We wish to ascertain the performance of BART as a prediction model based on several other varied datasets. We also plan to generate a sufficient number of bootstrap samples from our data and then repeat our data splitting procedure, while controlling the training and test set censoring rates, for all the generated bootstrap samples. The models could then be fit on all the bootstrapped training samples and the performance measures evaluated from the respective test samples. Obtaining a mean score for the different performance evaluation measures based on all the test samples could lead to results with increased viability.

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Blanchard, G. (2004). Un algorithme accéléré d'échantillonnage bayésien pour le modèle cart. *Revue d'intelligence artificielle*, 18(3):383–410.

Bleich, J., Kapelner, A., George, E. I., Jensen, S. T., et al. (2014). Variable selection for bart: an application to gene regulation. *The Annals of Applied Statistics*, 8(3):1750–1781.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011). Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Statistical Modelling*, 11(5):429–446.

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

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

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

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. wadsworth & brooks. *Monterey, CA*.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57.

Brier, G. (1950). Verification of forecasts expressed in term of probabilities. *Monthly weather review*, 78:1–3.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292.

Chambless, L. E. and Diao, G. (2006). Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in medicine*, 25(20):3474–3486.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.

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

Ciampi, A., Negassa, A., and Lou, Z. (1995). Tree-structured prediction for censored survival data and the cox model. *Journal of clinical epidemiology*, 48(5):675–689.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275.

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

Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127.

Denison, D. G., Mallick, B. K., and Smith, A. F. (1998). A bayesian cart algorithm. *Biometrika*, pages 363–377.

D'Eredita, G., Giardina, C., Martellotta, M., Natale, T., and Ferrarese, F. (2001). Prognostic factors in breast cancer: the predictive value of the nottingham prognostic index in patients with a long-term follow-up that were treated in a single institution. *European Journal of Cancer*, 37(5):591–596.

Fahrmeir, L. (1998). Discrete survival-time models. *Encyclopedia of biostatistics*.

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

Faradmal, J., Soltanian, A. R., Roshanaei, G., Khodabakhshi, R., and Kasaeian, A. (2014). Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse. *Asian Pacific journal of cancer prevention: APJCP*, 15(14):5883–5888.

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gao, F., Manatunga, A. K., and Chen, S. (2004). Identification of prognostic factors with multivariate survival data. *Computational statistics & data analysis*, 45(4):813–824.

Gao, F., Manatunga, A. K., and Chen, S. (2006). Developing multivariate survival trees with a proportional hazards structure. *Journal of Data Science*, 4(3):343–356.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

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

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

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137.

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in medicine*, 23(1):77–91.

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

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, pages 841–860.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Krketowska, M. (2004). Dipolar regression trees in survival analysis. *Biocybernetics and biomedical engineering*, 24:25–33.

Krketowska, M. (2006). Random forest of dipolar trees for survival prediction. *Artificial Intelligence and Soft Computing–ICAISC 2006*, pages 909–918.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23.

Ma, S. and Huang, J. (2007). Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics*, 23(4):466–472.

Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, pages 77–83.

Mallick, B. K., Denison, D. G., and Smith, A. F. (1999). Bayesian survival analysis using a MARS model. *Biometrics*, 55(4):1071–1077.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50(11):1.

Naftel, D., Blackstone, E., and Turner, M. (1985). Conservation of events. *Unpublished notes*.

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263.

Omurlu, I. K., Ture, M., and Tokatli, F. (2009). The comparisons of random survival forests and cox regression analysis with simulation and an application related to breast cancer. *Expert Systems with Applications*, 36(4):8582–8588.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13):2109–2123.

Pittman, J., Huang, E., Nevins, J., Wang, Q., and West, M. (2004). Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. *Biostatistics*, 5(4):587–601.

Rao, M. (1998). Survival analysis, techniques for censored and truncated data.

Rosenberg, J., Chia, Y. L., and Plevritis, S. (2005). The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the us seer database. *Breast cancer research and treatment*, 89(1):47–54.

Ross, J. S. (2009). Multigene classifiers, prognostic factors, and predictors of breast cancer clinical outcome. *Advances in anatomic pathology*, 16(4):204–215.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, pages 35–47.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in medicine*.

Tibshirani, R. et al. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537.

Xu, R. and Adak, S. (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics*, pages 305–315.

Xu, R., Adak, S., et al. (2001). Survival analysis with time-varying relative risks: a tree-based approach. *Methods Archive*, 40(2):141–147.

Yin, Y., Anderson, S. J., Parran Hall, G., Street, D., et al. (2002). Nonparametric tree-structured modeling for interval-censored survival data. In *Joint Statistical Meeting*.

Zhang, H. and Singer, B. (2013). *Recursive partitioning in the health sciences*. Springer Science & Business Media.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhou, Q. and Liu, J. S. (2008). Extracting sequence features to predict protein - DNA interactions: A comparative study. *Nucleic Acids Research*, 36(12):4137–4148.

APPENDIX A

LIST OF SUPPLEMENTARY FIGURES

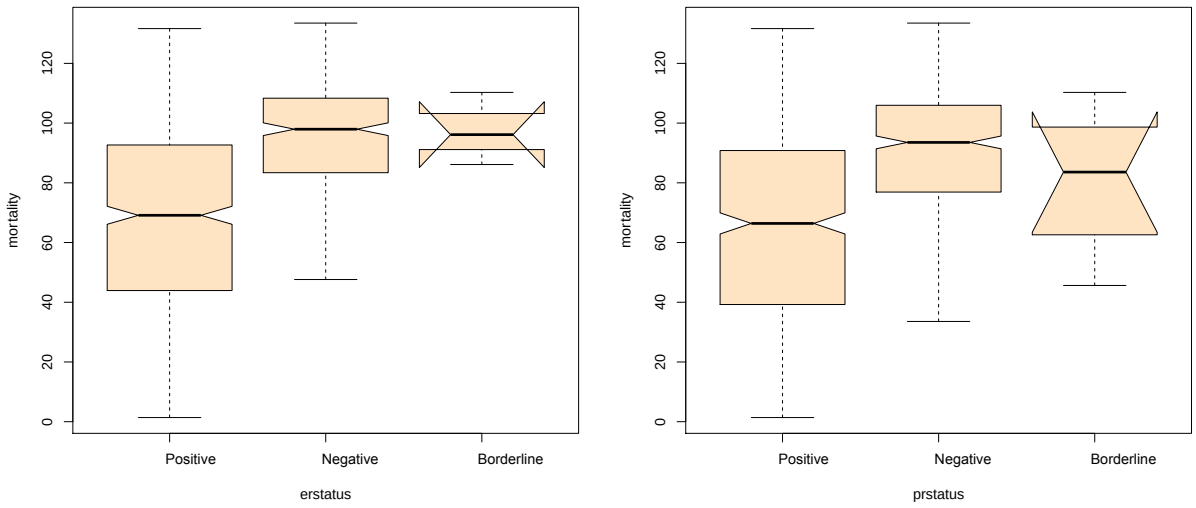

Figure A.1: Plots of the marginal effect of covariates erstatus and prstatus on estimated mortality computed using RSF

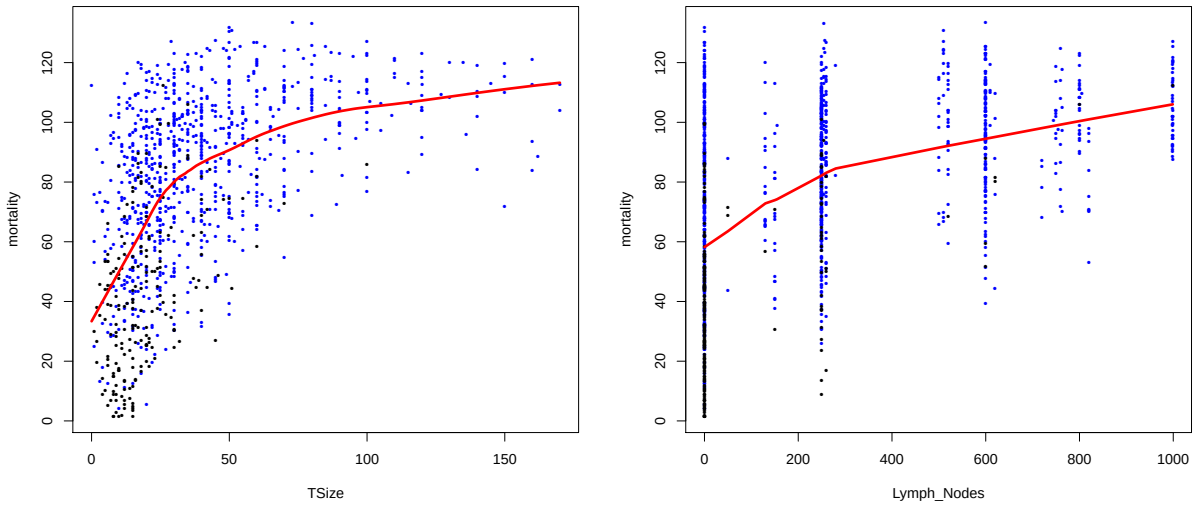

Figure A.2: Plots of the marginal effect of covariates Tumor size and Number of lymph nodes on estimated mortality computed using RSF

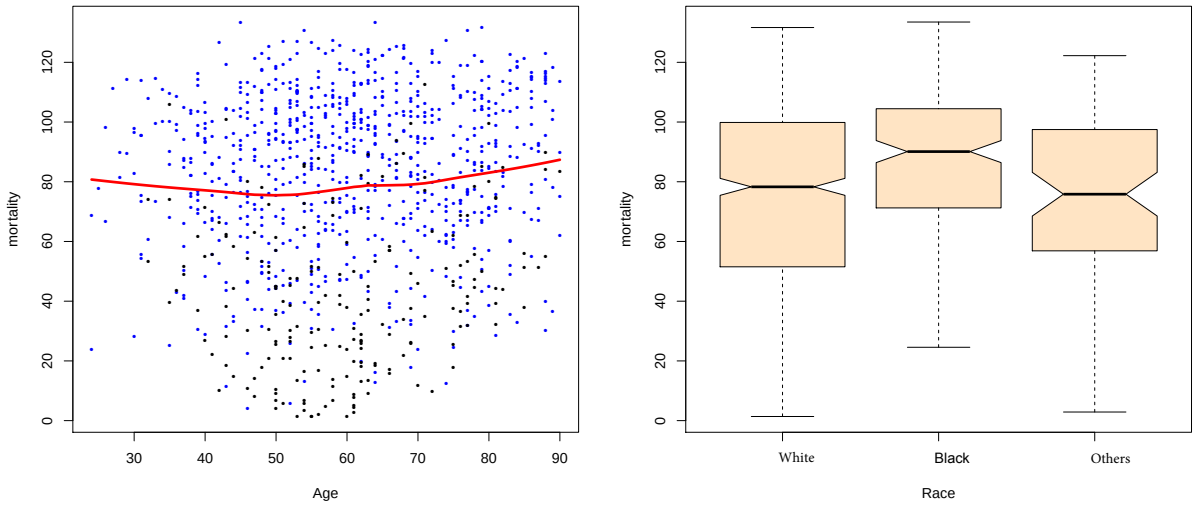

Figure A.3: Plots of the marginal effect of covariates Age and Race on estimated mortality computed using RSF

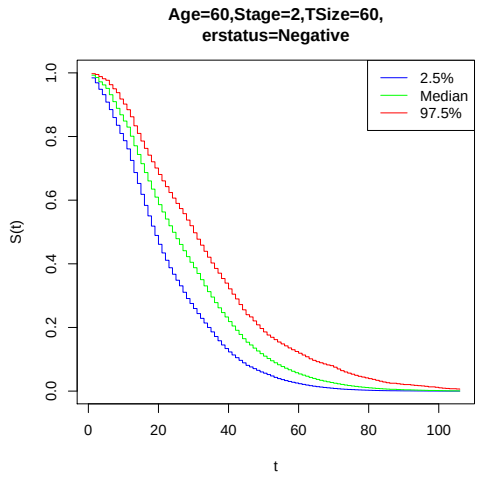

Figure A.4: Median survival probablity along with 95% confidence intervals computed using BART

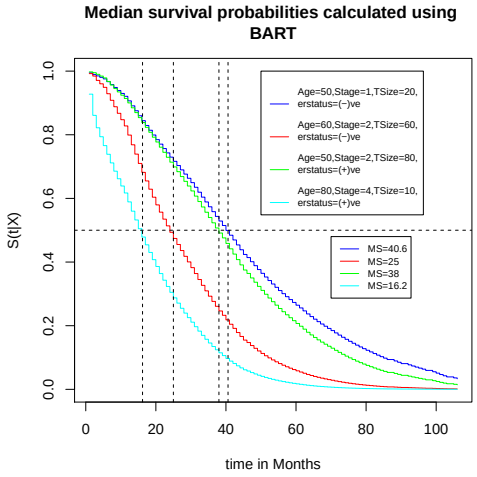

Figure A.5: Median survival probabilities for several covariate combinations computed using BART (MS:Median Survival in months)

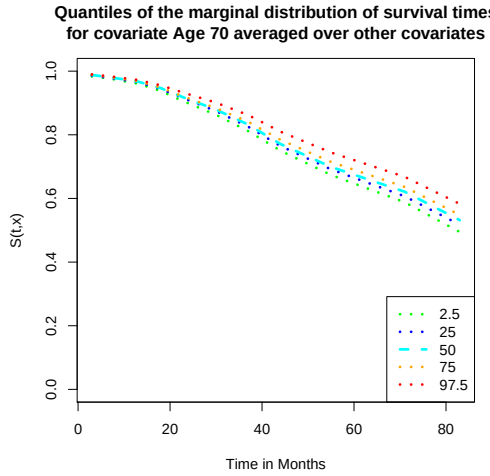

Figure A.6: Quantiles of the marginal distribution of survival times for covariate Age=70 averaged over other covariates

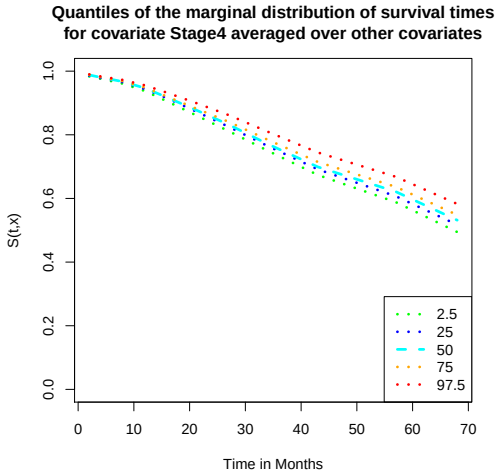

Figure A.7: Quantiles of the marginal distribution of survival times for covariate Stage=4 averaged over other covariates

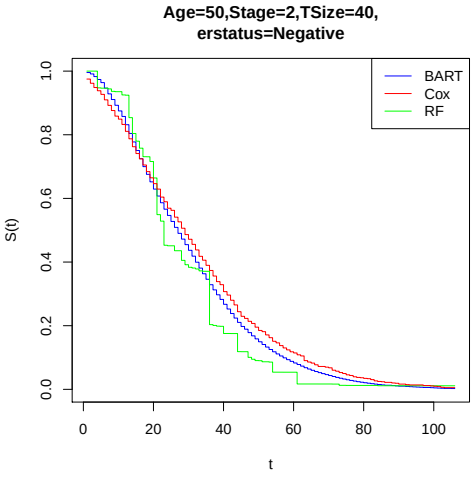

Figure A.8: Comparison of survival probability curves predicted using CPH,RSF and BART at Age=50, Stage=2, Tumor Size =40 and erstatus=Negative

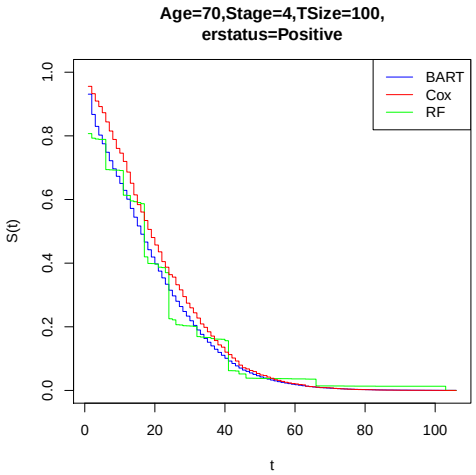

Figure A.9: Comparison of survival probability curves predicted using CPH,RSF and BART at Age=70, Stage=4, Tumor Size =100 and erstatus=Positive