UNIVERSITY OF CALIFORNIA,
IRVINE

An Ensemble Prognostic Model for Metastatic, Castrate-Resistant Prostate Cancer

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Yeeleng Scott Vang

Thesis Committee:
Professor Xiaohui Xie, Chair
Associate Professor Alexander Ihler
Associate Professor Zhaoxia Yu

2016

DEDICATION

To my daughter, wife, and families for your unconditional love and to my eldest brother for his selfless and exemplary role in my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Prof. Xiaohui Xie for his guidance and unwavering support through this effort. I would also like to thank my committee members Prof. Alexander Ihler and Prof. Zhaoxia Yu for their suggestions and feedback on this work.

I would also like to thank Yifei Chen for his help getting GBMCI compiled and running on the latest version of R. Furthermore, I would like to thank my manager Sheri Festa for her flexibility and continued support and in my graduate education.

Lastly, I cannot thank my wife and families enough for their unwavering support without which this thesis would not have been possible.

ABSTRACT OF THE THESIS

An Ensemble Prognostic Model for Metastatic, Castrate-Resistant Prostate Cancer

By

Yeeleng Scott Vang

Master of Science in Computer Science

University of California, Irvine, 2016

Professor Xiaohui Xie, Chair

Metastatic, castrate-resistant prostate cancer (mCRPC) is one of the most prevalent cancers and is the third leading cause of cancer death among men. Several treatment options have been developed to combat mCRPC, however none have produced any tangible benefits to patients' overall survivability. As part of a crowd-sourced algorithm development competition, participants were asked to develop new prognostic models for mCRPC patients treated with docetaxel. Such results could potentially assist in clinical decision making for future mCRPC patients.

In this thesis, we present a new ensemble prognostic model to perform risk prediction for mCRPC patients treated with docetaxel. We rely on traditional survival analysis model like the Cox Proportional Hazard model, as well as more recently developed boosting model that incorporates smooth approximation of the concordance index for direct optimization. Our model performs better than the the current state-of-the-art mCRPC prognostic models for the concordance index performance measure and is competitive with these models on the integrated time-dependent area under the receiver operating characteristic curve.

Chapter 1

Introduction

Survival analysis is a branch of statistics used to analyze data sets where the dependent variable is the time to an event of interest. In the clinical setting, survival analysis is often used to study the time until death of patients in longitudinal studies. Complicating these longitudinal studies is the problem of right censoring where patients may drop out before the conclusion of the test or patients does not experience the event before the end of the test, thereby resulting in an unobservable survival time for the patient. The tools to tackle these difficulties necessitates the use of algorithms from survival analysis.

The scope of this thesis encompasses using tools from survival analysis to develop a prognostic model for predicting risk in metastatic, castrate-resistant prostate cancer (mCRPC) patients treated with docetaxel. Prostate cancer happens to be the third leading cause of cancer death and is the most common cancer among men in developed countries [1]. mCRPC is inevitably a fatal disease with current treatment options offering minimal, if any, benefit to mCRPC patients' overall survival outcome. We will be be answering the question of what benefits, if any, mCRPC patients treated with docetaxel see in their overall survival. Since the historical median survival time for mCRPC patient is less than 2 years [10], an accurate

prognostic model indicating whether doxcetaxel provides any tangible improvement would be of potential benefit in the decision making of treatment options.

We provide a brief description of the structure of the thesis along with our contributions in the following section.

1.1 Thesis Outline

The thesis beings with providing relevant background information in Chapter 2. Here we provide an overview of prostate cancer, some survival statistics, the specific mCRPC condition, and current and future treatment options. We then introduce the task description and provide context to its motivation. We conclude this chapter with formulation of the performance measures, give an intuitive interpretation of them, and derived their relationship to one another.

Chapter 3 provides a survey of existing state-of-the-art prognostic models developed for mCRPC patients risk prediction. Focus will be given to the entire development phase from pre-processing steps to statistical models used. This will be used to highlight each algorithms strengths and weaknesses as well as range of models used.

In Chapter 4, we provide a short introduction to the survival distribution, derive important characterizing functions, and point out their relationship. We highlight some of the difficulties working with survival data. We then introduce three models used for survival analysis that will serve as the basis of our ensemble prognostic model.

Chapter 5 looks at some of the major steps in our model development process. We also validate the proportional hazard assumption. We then provide information on our experiment setup and present the result of our experiment. We conclude this section with a detail

analysis of the strength and weakness of our model compared to current state-of-the-art models.

Chapter 6 summarizes our contribution and this thesis and also looks at some possible direction for future research.

Chapter 2

Overview of mCRPC and Prediction Task

2.1 mCRPC Background

Prostate cancer is the third leading cause of cancer mortality, behind lung and colorectal, and is the most frequently diagnosed cancer for men [1, 26, 39]. Nearly 15% of the 2 million men diagnosed in the US with prostate cancer had metastatic disease (Stage IV) at the time of diagnosis [2]. It has been demonstrated since 1941 that deprivation of androgen hormone (e.g. testosterone) leads to regression of prostate cancer and alleviation of pain in patients [27].

Androgen deprivation therapy (ADT) [7] has become the de facto initial treatment for patients with prostate cancer. ADT can involve either medical or surgical castration (e.g. bilateral orchidectomy)[5]. The effect of castration effectively reduces the serum testosterone level to a very low level known as the castration level. Medical castration is usually achieved by Gonadotropin-releasing hormone (GnRH) agonist which inhibits the pituitary

gland from releasing luteinizing hormone necessary for testicular androgen production. A popular treatment in the 1990's consisted of combining GnRH agonist and androgen blockers, such as flutamide or bicalutamide, in a treatment called total androgen blockade (TAB) [20] which seek to completely block androgen activity in the body.

Despite these blockage, prostate cancer is known to progress to a state called castration-resistant prostate cancer (CRPC) between 18 to 48 months. This state is characterized by elevated level of prostate specific antigen (PSA) in spite of very low level of testosterone. Prostate cancer cell has been found to be able to maintain dihydrotestosterone (DHT) concentration in excess of serum concentrations to support cell growth and proliferation [35] and may, in some instances, also synthesize DHT de-novo [32].

The continued progression of CRPC leads to metastatic, castrate-resistant prostate cancer (mCRPC) which accounts for one third of all patients with metastatic disease [1, 26]. More than 90% of mCRPC patients develops bone metastases which results in increased risk of morbidity and mortality [13] as a result of skeletal-related events. mCRPC is inevitably fatal with the historic median survival of patients being less than 2 years [10]. Docetaxel, an anti-mitotic chemotherapy, is usually the standard first-line treatment option for mCRPC patients [20]. Although secondary treatment, such as cabazitaxel, and tertiary treatment, such as active cellular immunotherapy, exists, they have not been found to produce any major improvement in mCRPC patient overall survivability [46].

The current focus in mCRPC treatment is shifted to inhibitors of steroid biosynthesis [20]. This involves treating patients with ketoconazole to inhibit 17α-hydroxylase, which is a key ingredient to the production of precursors for androgen production. High doses have been used to suppress tumor activity but no surivval benefit has been shown. It is also associated with elevated and significant adverse events including bone fragility, hypotension, and fatal hepatic dysfunction. The outcome observed after treatment with ketoconazole has lead to investigation of stronger and more selective inhibitors that also has more favorable toxicity

profile. Two such inhibitors recently approved by the US Food and Drug Administration (FDA) are abiraterone acetate (Zytiga) and enzalutamide (Xtandi). Preliminary results from phase I studies of these newer drugs are promising while further phase II studies are on-going, with phase III trials in the planning stage.

2.2 Task Description

Halabi et al. [21] showed that by using the most up-to-date mCRPC patient data, better prognostic models can be developed. They extended their study to investigate the site of metastatic disease as being informative prognostic indicators. Their finding demonstrated the importance of prognostic research to include current clinical trials and patient health status in determining the best treatment choices.

The results of the Halabi study motivated the organizers at Sage Bionetworks to create the DREAM 9.5 - Prostate Cancer DREAM Challenge [2], henceforth known as DREAM Challenge. Organizers hoped a crowd-sourced competition could lead to new models for predicting survival in mCRPC patients treated with docetaxel. With better models, the goal is to allow clinical researchers to decide if docetaxel is a viable first treatment option or not. It is this question posed by the DREAM Challenge that this thesis will address.

Participants are asked to build a prognostic model to predict a global risk prediction and optionally 3 separate optimized risk predictions at 12, 18, and 24 months for patients in the test data set. For the global risk prediction, a concordance index (c-index) and an integrated time-dependent AUC(iAUC), from 6 to 30 months, using Hung and Chiang's estimator of *cumulative* AUC [28], are calculated for each model. For the time specific predictions (12, 18, 24 months), an AUC score using Hung and Chiang's estimators are calculated for these times. Details of these performance measures are discussed in the following subsections.

2.3 Data Set

The data set used is provided as part of the DREAM Challenge. This data set contains 2070 patients from four mCRPC Phase III clinical trials study broken down as 476 patients from Memorial Sloan Kettering Cancer Center (ASCENT2), 526 patients from Celgene (CELGENE), 598 patients from Sanofi (EFC6546), and 470 patients from AstraZeneca (AZ). The DREAM Challenge used the 1600 patients from the first three trials as their training data and divided the AZ set into 157 patients for the leaderboard round and 313 patient for the final round. AZ ground truths were not made available to the participants.

The four clinical trials were pre-processed and aggregated by the DREAM Challenge organizer and provided as one CSV table representing patient level data consisting of the dependent variables and the clinical covariates such as patient demographics, lesion measure, prior medicine, vital sign, ect. Additional raw longitudinal data is provided for possible further exploration. The dependent variables are $DEATH$, as the binary survival outcome variable, and $LKADT\_P$ (last known alive day - in days) as the time to event.

One of the main issue with clinical trial data set is the issue of missing data. We consider various methods for missing data imputation including using the mean, median, or mode value of the covariates across the four data sets. We settle on using the training table preprocessed by Laajala et al. [30] as part of their entry into the DREAM Challenge which is discussed further in Section 3.3.

2.4 Performance Measures

The performance measures, concordance index and integrated time-dependent AUC, set forth by the organizer for the DREAM Challenge are both commonly used in survival analysis in

the context of discerning discriminatory power of prognostic models. Discrimination reflects the ability of a prognostic model to correctly identify a clinical status and is of particular importance for clinician as they are interested in how a predicted score is able to distinguish between individuals labeled as high risk or a given event from those not. In the following subsections, we provide background information for both performance measures and show the relationship between the two.

2.4.1 Concordance Index

The concordance index (c-index) is the standard performance measure for model assessment in survival analysis with right censoring [17, 42]. This measure was developed by Harrell [24] and measures the separation of two survival distribution. It is given as:

$$c - index = \frac{1}{|E|}\left(\sum_{(i,j)\in E} \mathbb{1}(f(x_i) < f(x_j)) + 0.5 \cdot \sum_{(i,j)\in E} \mathbb{1}(f(x_i) = f(x_j)) \right) \tag{2.1}$$

where E is the set of validly orderable pairs, $|E|$ is the number of pairs in E, $\mathbb{1}$ is the indicator function whether the condition is satisfied or not, and $f(x_i)$ is the prediction of survival time for patient i. The set of validly orderable pairs consists of pairs where the shorter survival time is not censored. In the case of ties, they are given half weights.

Intuitively, the c-index is the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all pairs that can be ordered. The survival analysis problem involving censored data can be cast as a ranking problem, and in this sense, c-index is a natural measure to quantify the quality of rankings [42].

2.4.2 Integrated Time-dependent AUC

The time-dependent area under the receiver-operating characteristic (ROC) curve, AUC(t), is another frequently used measure for survival analysis. The time-dependent ROC, ROC(t), follows from the standard definition of ROC curve which depends on true positive rates (TPR) and false positive rates (FPR). To begin with, we first define the following notations.

Following the derivation of [6], let T_i and C_i denote the survival and censoring times respectively for patients $i = 1, \ldots, n$. Furthermore, let $Z_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{1}(T_i \leq C_i)$ denote the observed survival time and the indicator of death respectively. We denote $D_i(t)$ the time-dependent outcome status for patient i at time t, for $t \geq 0$. Define $D_i(t) = 1$ if patient i is considered a case (experience death) at time t and $D_i(t) = 0$ if patient i is considered a control (did not experience death) at time t.

By casting the predictions from a prognostic model as a binary problem, we can extend the definition of ROC to cover survival analysis. For the general setting, for any threshold c, let TPR(c, t)=Pr$(X > c|D(t) = 1)$ and FPR(c, t)=Pr$(X > c|D(t) = 0)$. By varying c and plotting TPR(c, t) against FPR(c, t), the ROC(t) curve can be obtained. The time-dependent AUC at time t is defined as the area under this curve:

$$\text{AUC}(t) = \int_{-\infty}^{\infty} \text{TPR}(c, t) \left| \frac{\partial \text{FPR}(c, t)}{\partial c} \right| dc \tag{2.2}$$

For survival analysis, there are several potential extensions of cross-sectional TPR and FPR, leading to different formulation of the time-dependent AUC [6, 25]. For the DREAM Challenge, the Hung and Chieng's estimator [28] for *cumulative* AUC is used and is formulated as follows.

First define cases as *cumulative* if $T_i \leq t$ and controls as *dynamic* if $T_i > t$. Then cumulative

TPR and dynamic FPR are respectively defined as:

$$\text{TPR}^{\mathbb{C}}(c,t) = \Pr(X_i > c | T_i \leq t) \quad \text{and} \quad \text{FPR}^{\mathbb{D}}(c,t) = \Pr(X_i > c | T_i > t). \tag{2.3}$$

At any fixed time t, all the patients are classified as either a case or a control on the basis of vital status at time t. Each patient is a control for times $t < T_i$ and then is a case for $t \geq T_i$. *Cumulative* specificity and *dynamic* sensitivity accuracy summaries are important when interest lies in discriminating between subjects who die prior to a given time T_i and those that survive beyond T_i [25].

Hung and Chieng then estimated these two functions using non-parametric method based on Inverse Probability of Censoring Weighting (IPCW) as:

$$\widehat{\text{TPR}}^{\mathbb{C}}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i > c, Z_i \leq t) \frac{\delta_i}{n\hat{S}_C(Z_i)}}{\sum_{i=1}^{n} \mathbb{1}(Z_i \leq t) \frac{\delta_i}{n\hat{S}_C(Z_i)}} \tag{2.4}$$

and

$$\widehat{\text{FPR}}^{\mathbb{D}}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i > c, Z_i > t)}{\sum_{i=1}^{n} \mathbb{1}(Z_i > t)} \tag{2.5}$$

where $\hat{S}_C(\cdot)$ is the Kaplan-Meier estimator of the survival function of the censoring time C.

It can then be shown (see [28]) that an estimator of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ is given by:

$$\widehat{\text{AUC}}^{\mathbb{C},\mathbb{D}}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(Z_i \leq t) \mathbb{1}(Z_j > t) \mathbb{1}(X_i > X_j) \frac{\delta_i}{\hat{S}_C(Z_i)\hat{S}_C(t)}}{n^2 \hat{S}(t)[1 - \hat{S}(t)]} \tag{2.6}$$

The integrated time-dependent AUC (iAUC) is the area under the time-dependent AUC for all time t and is obtained using Simpson's rule to numerically integrate over relevant time

points.

2.4.3 Relationship between c-index and integrated time-dependent AUC

Both c-index and AUC are frequently used to assess the discriminatory power of a prognostic model [17]. In fact, Harrell's c-index was developed specifically to estimate the concordance probability for right censored survival analysis. A c-index of 1 indicates the model has perfect discrimination, the same as when AUC is equal to 1. When c-index is 0.5, the model is doing no better than random guessing (no discrimination), equivalent to when AUC is 0.5. This implies that both measures operate over the same range and should naturally share a relationship.

To see this relationship, we start with the definition of concordance probability. The concordance probability for a pair of patients (X_1, T_1) and (X_2, T_2) is defined as:

$$C = \Pr(X_2 > X_1 | T_2 > T_1). \tag{2.7}$$

Define controls as *dynamic* as in subsection 2.4.2 and now define cases as *incident* if $T_i = t$. The *incident* TPR is then defined as:

$$\mathrm{TPR}^{\mathbb{I}}(c, t) = \Pr(X_i > c | T_i = t). \tag{2.8}$$

Each patient is a control for times $t < T_i$ and then is a case when $t = T_i$. *Dynamic* specificity and *incident* sensitivity are defined by dichotomizing the risk set at time t into those observed to die and those that survive and is therefore a natural companion to hazard models [25].

Van et al. [44] showed that the *incident/dynamic* $\mathrm{AUC}(t)$ can be formulated as:

$$\mathrm{AUC}^{\mathbb{I},\mathbb{D}}(t) = \frac{\#\{j \in R(t); x_j < x_i\} + 0.5 \cdot \#\{j \in R(t), j \neq i; x_j = x_i\}}{|R(t)| - 1} \tag{2.9}$$

where $R(t)$ is the risk set and $|R(t)|$ is the size of the risk set at time t.

Then Van et al. also showed that computing some weighted average proportional to $|R(t)|-1$ leads to:

$$\begin{aligned} C^{\tau} &= \sum_{i \in \mathcal{D}} \Big[(|R(t_i)| - 1) \cdot \mathrm{AUC}(t_i) \Big] \Big/ \sum_{i \in \mathcal{D}} (|R(t_i)| - 1) \\ &= \frac{\sum_{i \in \mathcal{D}} \Big(\#\{j \in R(t_i); x_j < x_i\} + 0.5 \cdot \#\{j \in R(t_i), j \neq i; x_j = x_i\} \Big)}{\sum_{i \in \mathcal{D}} (|R(t_i)| - 1)}. \end{aligned} \tag{2.10}$$

This formulation is precisely Harrell's c-index as defined in Section 2.4.1.

Chapter 3

Related Work on mCRPC Prognostic Models

3.1 Introduction

Prostate cancer remains the most prevalent malignancy in men with development of mCRPC as the major cause of death in these patients [21]. Three prognostic models were developed incorporating some prognostic markers of overall survival in pre-chemotherapy patients [4, 23, 40]. As treatment advanced to include various chemotherapy, Halabi et al. developed a more modern prognostic model to account for these new trial design [21, 22]. Recently, as part of a crowd-sourced DREAM Challenge, teams were invited to develop prognostic models for patients treated with docetaxel [2] with the aim of besting the model by Halabi et al. used as the baseline model in this competition. Along with the baseline model, the top five performing models from the Dream Challenge are discussed in detail in the following sections.

3.2 Halabi et al. Model

Halabi et al. [21], the baseline model, considered eight previously defined predictors of overall survival for mCRPC patients. These eight covariates were the binary covariates ANALGESICS and BONE,VISCERAL,LYMPH_NODES, the categorical covariate ECOG_C, and the numerical covariates LDH, ALB, HB, ALP, and PSA. In their model, missing data were imputed with the median of non-missing values across all four longitudinal Phase III clinical trials. The model used was a penalized Cox Proportional Hazards model with LASSO-regularization. The advantage of using the LASSO-regularization was to produce sparse regression coefficients which can further reduce the predictor space.

3.3 Laajala et al. Model

Laajala et al. [30], the first place team in the DREAM Challenge, used the most sophisticated pre-processing steps. They used Gaussian generalized linear model to impute missing-at-random and structured missing data and log-transformed numerical covariates to fit a normal distribution. They further used principal component analysis to reveal that the distribution of patients from the ASCENT2 study did not match that of the AZ study and therefore was excluded completely from training purpose.

Their final prognostic model was an ensemble of three highly optimized elastic-net regularized GLM-based Cox model trained using batch data available only from the EFC6546 study, only from the CELGENE study, and the combination of both EFC6546 and CELGENE studies simultaneously. The final prognostic model equally weighted prediction from all three models.

3.4 Xiao et al. Model

Xiao et al. [47], the second place team in the DREAM Challenge, used 20 candidate covariates identified as informative by the Halabi study [21]. Missing data was imputed using the Multivariate Imputation by Chained Equations (MICE) algorithm and the Nelson-Aalen estimator. Each patients were assigned a weight based on the local density in the covariate space. The final prognostic model was a Ridge-regularized GLM-based Cox model trained over 100 repeated-subsampling from a random 2/3 portion of the training set and validated on the remaining 1/3 samples.

3.5 Song and Wang Model

Song and Wang [41], the third place team in the DREAM Challenge, imputed missing data with the median value of each covariates. They further eliminated covariates that had little variation and those with large percentage of missing data which ultimately left them with 65 covariates. From here, they performed a round of univariate filtering based on p-value, selecting to keep those only with p-value less than 0.05. This left them with just the AST and LIVER covariates. Based on Halabi study, Song and Wang also included the ECOG_C, LDH, ALB, HB, PSA, and ALP covariates. No normalization or log transformation was used. Their final prognostic model consisted of the non-regularized Cox Proportional Hazards model.

3.6 Wolfinger and Chu Model

Wolfinger and Chu [45], the fourth place team in the DREAM Challenge, imputed missing data using the Full Conditional Specification (FCS) method. This was achieved by averaging 10 iteration of FCS. Continuous covariates were log transformed and each covariates were

standardized to have mean zero and standard deviation one.

Wolfinger and Chu used three classes of models to generate each individual predictor comprising their final ensemble model. The first class of models was Cox Proportional Hazard models and using variation in forward, backard, and stepwise covariate selections. The second class of model is the Life regression models with fitted standard distribution for time-to-event data. These distribution models included Weibull, Gamma, Logistic, and Log-normal. The third class of models uses Buckley-James approach to wrap standard methods including linear regression (lasso and elastic net), k-nearest neighbors, trees (boosting and forests), partial least squares, and kernel methods. Their final ensemble prognostic model consisted of 351 individual models derived with slight variation of the three classes of models mentioned above.

3.7 Greiner et al. Model

Greiner et al. [19], the fifth place team in the DREAM Challenge, used a variety of pre-processing steps. The first involved removing covariates that did not have values for the AZ test data set. Categorical covariates were replaced with one-hot encoding version. Two copies of numerical covariates were included. One copy was quantile normalized, whose missing value was imputed using the median. The other copy was first linearly normalized such that the minimal value is 0 and maximum value is 1. Missing value was then imputed using the mean, followed by another quantile normalization of these values. No covariate selection step was used.

Greiner et al.'s final model used an algorithm called Patient Specific Survival Prediction (PSSP)[31] which is based on multi-task logistic regression (MTLR) which seeks to learn patient-specific survival distribution. This method differs from traditional survival analy-

sis method as it models the survival function directly by combing multiple local logistic regression models instead of modeling hazard function as is commonly done with Cox Proportional Hazard models. Essentially the model learns a logistic regression function for each time point, then combine these functions in such a way that guarantees the resulting curve is monotonically decreasing from a probability of 1 at $t = 0$.

Chapter 4

Material and Methods of Survival Analysis

4.1 Introduction

We begin this chapter with a short introduction to the relevant functions used in survival analysis. The survival distribution is characterized by the following four functions.

The probability density function is given as:

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(t \leq T < t + \Delta t) \tag{4.1}$$

The survival function defines the probability of the event is later than time t and is defined as:

$$S(t) = \Pr(T > t) = 1 - \Pr(T \leq t) = 1 - F(t) = 1 - \int_0^t f(s)ds \tag{4.2}$$

The survival function is non-increasing with respect to time.

A related concept is the hazard function defined as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{4.3}$$

which gives the instantaneous death rate at time t given the patient is alive up to time t. The hazard function is not a probability as it is not bounded above by 1 [16].

The cumulative hazard function is defined as:

$$\Lambda(t) = \int_0^t \lambda(s) ds = -\log S(t) \tag{4.4}$$

Intuitively, when the hazard function is high, the cumulative hazard increases faster and the survival decreases faster with time.

Knowledge of f, S, λ, or Λ is enough to specify the survival distribution and the other remaining three functions [16].

4.2 Censoring

The difficulty in modeling survival data is the problem of censoring. There are various categories of censoring such as right censoring, left censoring, and interval censoring. Since right censoring is the only kind present in this data set, only it is explored further in this thesis. For other kinds of censoring, we refer readers to Kleinbaum [29].

Right censoring occurs when patients have not experienced the event of interest by the time the study have concluded, patient are lost to follow-up, termination of treatment due to

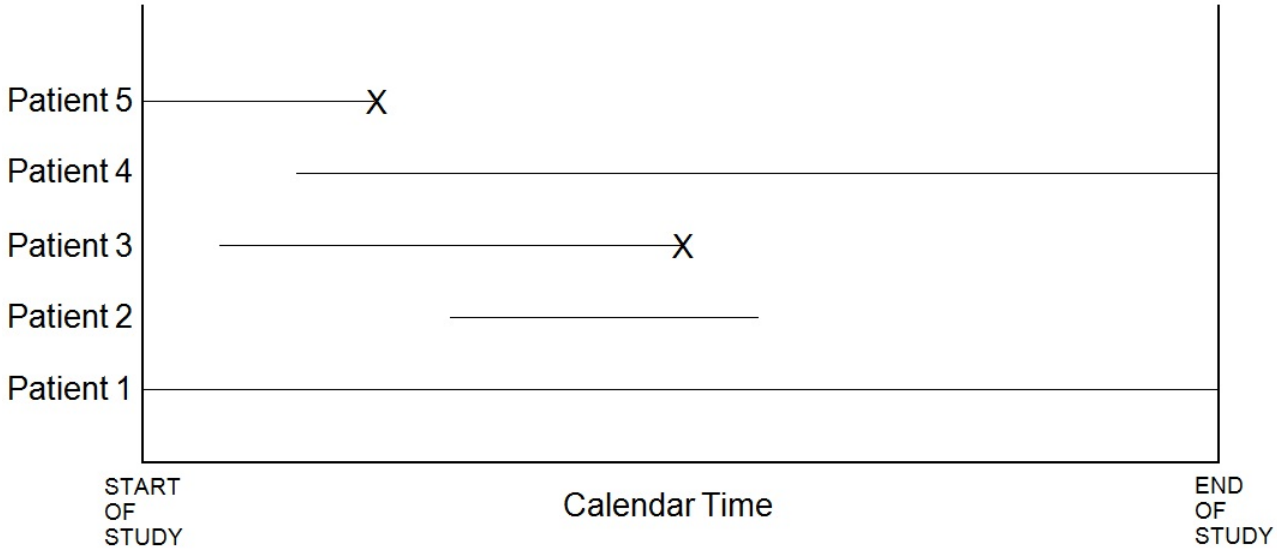

Figure 4.1: Lifetime curve for patients. General right censoring of patients with possible different starting time.

negative side effects of treatment, or death due to an unrelated cause [16]. The concept of right censoring can be explained using Figure 4.1. In this example, only patients 3 and 5 experienced the event of interest, denoted with the "X" at the right of their lifetime curves. The remaining patients did not experience the event and were right censored either by dropping out of the test (patient 2) or were still alive at the end of the study (patients 1 and 4). In this example, patients were able to have different start time. Censored patients are included in the risk set up to their censor times but are excluded thereafter. The risk set at time t(j) is defined as the set of all individuals whose survival times are at least t(j) or larger. They play a pivotal role in the partial likelihood function defined in the following section.

4.3 Survival Models

4.3.1 Model #1 - Cox Proportional Hazard

The Cox Proportional Hazard (CoxPH) model is arguably the most popular survival model in practice today. In his seminal work [11, 12], Cox introduced the proportional hazard model:

$$\lambda(t|x_1, \ldots, x_n) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_n x_n} \tag{4.5}$$

where $\lambda_0(t)$ is the baseline hazard function, and $e^{\beta_1 x_1 + \cdots + \beta_n x_n}$ is the relative hazard, which summarizes the effects of the covariates.

An important property of the CoxPH model is that the baseline hazard is unspecified. This makes the Cox model robust as estimating parameters for the CoxPH model does not require estimating the baseline hazard function. Estimating parameters for the Cox model is done by maximizing the partial likelihood function, which only considers probabilities for those subjects who fail and not explicitly for those subjects who are censored [29]. The partial likelihood is given as:

$$L(\beta|x) = \prod_j \frac{e^{\beta^T x_{(j)}}}{\sum_{i \in R_j} e^{\beta^T x_{(i)}}} \tag{4.6}$$

where R_j is the set of subjects at risk for failure at time $t_{(j)}$.

The CoxPH model has been extended to model time-varying covariates and time-varying coefficients as well [3, 33]. This CoxPH framework has been implemented in the open-source R package "survival" [43].

4.3.2 Model #2 - Gradient Boosting Machine

Gradient boosting machine (GBM) is an ensemble model by additive expansion of sequential "weak learners" [14, 15]. In the problem formulation, Friedman [14] was interested in learning a functional mapping $y = F(x; \beta)$ from data $\{x_i, y_i\}_{i=1}^{N}$ that minimizes some loss function $L(y, F(x; \beta))$.

More concretely, $F(x)$ takes the form of :

$$F(x) = \sum_{m=0}^{M} \rho_m f(x; \tau_m) \tag{4.7}$$

where ρ is the weight, f is the weak learner, and τ is the parameter set. Therefore β consists of the parameter sets $\{\rho_m, \tau_m\}_{m=1}^{M}$. These parameter are learned in the follow "stage-wise" greedy process:

(1) set an initial estimator $f_0(x)$.

(2) for $m \in 1, 2, , M$

$$(\rho_m, \tau_m) = \arg\min_{\rho,\tau} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \rho f(x_i; \tau)) \tag{4.8}$$

Step (2) is approximated by GBM in the following two steps:

First, fit $f(x; \tau_m)$ by

$$\tau_m = \arg\min \sum_{i=1}^{n} (g_{im} - f(x_i; \tau))^2 \tag{4.9}$$

where

$$g_{im} = -\Big[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\Big]_{F(x)=F_{m-1}(x)} \tag{4.10}$$

Second, learn ρ by

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \rho f(x_i; \tau_m)) \tag{4.11}$$

Then it updates $F_m(x) = F_{m-1}(x) + \rho f(x; \tau)$. To control overfitting, shrinking is often introduced to give the following form of the update equation:

$$F_m(x) = F_{m-1}(x) + \gamma\rho f(x; \tau) \qquad \text{where} \quad 0 \leq \gamma \leq 1 \tag{4.12}$$

If regression trees are used as the weak leaner, the tree parameters, e.g. tree depth, determine the complexity of $f(x)$. The performance of GBM can be improved by a method called subsampling, where a random subset of the training data is used to fit each weak learner. This leads to a method called *stochastic gradient boosting* [15]. This GBM framework is implemented in the open-source R package "gbm" [38].

Boosting Proportional Hazard Regression

Ridgeway [37] extended boosting for the Cox proportional hazard model by using the following negative log partial likelihood cost function:

$$L(y, F) = -\sum_{i=1}^{n} \delta_i \left\{ F(x_i) - \log\left(\sum_{j:t_j \geq t_i} e^{F(x_j)} \right) \right\} \tag{4.13}$$

Eqn (4.9), (4.10), (4.11) can then be applied to learn for each weak learner. This extension

is implemented in the "gbm" package by specifying "coxph" as the distribution to use [38].

4.3.3 Model #3 - Gradient Boosting Machine for Concordance Index Optimization

Gradient boosting machine for concordance index (GBMCI) learning was developed by Chen et al. [9] with the intent of optimizing a smooth approximation of the concordance index as proposed in [48]. This led to the following approximation termed the smoothed concordance index (SCI):

$$SCI = \frac{1}{|E|} \sum_{(i,j)\in E} \frac{1}{1 + e^{\alpha(F(x_i) - F(x_j))}} \tag{4.14}$$

where α controls the steepness of the sigmoid function (accordingly, the approximability of SCI to CI), and $F(x)$ is the prediction of survival time. Letting the loss function, $L(y, F) = -SCI$, the weak learner can be fitted using Eqn (4.9) and $\{g_{im}\}_{i=1}^{n}$, where

$$\begin{aligned} g_{im} = \left[\frac{dSCI}{dF(x_i)}\right]_{F(x)=F_{m-1}(x)} = \frac{\alpha}{|E|} \Bigg\{ & \sum_{(k,i)\in E} \frac{e^{\alpha(F_{m-1}(x_k) - F_{m-1}(x_i))}}{\left[1 + e^{\alpha(F_{m-1}(x_k) - F_{m-1}(x_i))}\right]^2} \\ & - \sum_{(i,j)\in E} \frac{e^{\alpha(F_{m-1}(x_i) - F_{m-1}(x_j))}}{\left[1 + e^{\alpha(F_{m-1}(x_i) - F_{m-1}(x_j))}\right]^2} \Bigg\} \end{aligned} \tag{4.15}$$

Next,

$$\rho_m = \arg\min_{\rho} \frac{1}{|E|} \sum_{(i,j)\in E} \frac{1}{1 + e^{\alpha(F_{m-1}(x_i) + \rho f(x_i;\tau_m) - F_{m-1}(x_j) - \rho f(x_j;\tau_m))}} \tag{4.16}$$

SCI is neither convex nor concave therefore a global minimum is not guaranteed. The algorithm's performance is dependent on initialization. The heuristic used was to set the

initial estimation of $\{f_o(x_i)\}_{i=1}^n$ to the prediction from a fitted PH model and line search was used to detect a local optimal for ρ_m. This framework is implemented as "gbmsci" [8] in R.

4.4 Ensemble Model

We will consider the scenario of both batch and non-batch effects.

As suggested by Laajala et al. [30], for our batch ensemble model, we train three models separately using batch data available only for the EFC6546 study, only from the CELGENE study, and the combination of both EFC6546 and CELGENE studies simultaneously.

The risk scores for each batch model j is base on a weighted combination of the individual CoxPH, GBM, and GBMCI models.

$$Risk_j = \alpha_1 * CoxPH + \alpha_2 * GBM + \alpha_3 * GBMCI \tag{4.17}$$

where $\sum_{i=1}^3 \alpha_i = 1$ and $\alpha_i \geq 0$.

For simplification, one standard set of weights were used for all three batch models. A fine grid search was used over the combined EFC6546 and CELGENE dataset to find the optimal weights and are given as:

$$\alpha_1 = 0.60, \alpha_2 = 0.20, \alpha_3 = 0.20$$

Our batch ensemble prognostic model weighs equally the three batch models to come up with our final risk prediction.

$$Risk_{batch} = \frac{\sum_{j=1}^3 Risk_j}{3} \tag{4.18}$$

For our non-batch ensemble prognostic model, we consider both EFC6546 and CELGENE data sets simultaneously.

$$Risk_{non-batch} = \alpha_1 * CoxPH + \alpha_2 * GBM + \alpha_3 * GBMCI \tag{4.19}$$

where the α_i weights are as given above in our batch model.

Chapter 5

Experiment and Result

5.1 Introduction

In this section, we go over the method used to select a subset of the covariates to use for model training and prediction. We further explain how hyperparameters were chosen for the individual models. Finally we give background information of the experiment setup and present the results.

5.2 Covariate Selection

Due to the large number of total clinical covariates available in the design matrix, a covariate selection step is perform to reduced the number to a manageable size for actual model training. Each covariate is model as a univariate survival problem using the CoxPH model in R with the *Surv* package to calculate its c-index value. These concordance indices are used to rank the covariates in descending order. The top 40 covariates are shown in Table 5.1. 40 is somewhat arbitrary as it represents a value we felt would prevent the model from

Rank	Covariate	c-index	Rank	Covariate	c-index
1	LDH	.654227	21	TARGET	.541053
2	ALP	.632614	22	PHOS	.540026
3	HB	.623218	23	CA	.538559
4	HEMAT	.622910	24	LIVER	.536700
5	RBC	.613883	25	CCRC	.533530
6	AST	.589591	26	MG	.529337
7	PSA	.587145	27	BONE	.527648
8	ALB	.573132	28	RegionEastEuro	.526697
9	PULSE	.571807	29	MHCARD	.524862
10	NA.	.558981	30	LYMPH_NODES	.524571
11	ECOG_C	.558148	31	MHGEN	.523388
12	SYSTOLICBP	.557846	32	RegionWestEuro	.523202
13	ANALGESICS	.557252	33	LUNGS	.522093
14	DIASTOLICBP	.555605	34	NEU	.520842
15	LYMperLEU	.555510	35	GLU	.520204
16	BMI	.549258	36	AGEGRP2	.520070
17	NEUperLEU	.548435	37	MHVASC	.519193
18	LYM	.546818	38	GONADOTROPIN	.518933
19	WEIGHTBL	.544780	39	MHNEOPLA	.518097
20	CREACL	.543328	40	POT	.517963

Table 5.1: Top 40 Ranked Covariates ranked in descending order of c-index

overfitting as well as what we believe is the cutoff ceiling for numbers of covariates that would be useful to clinician in the field.

From these top 40 covariates, 10-fold cross validation was performed using the training data and the non-batch ensemble model to further down select the number of covariates and determine the optimal number. From Figure 5.1, the optimal model was found by using the top 29 covariates.

5.3 Assessment of proportional hazard assumption

After obtaining the 29 covariates to be used, we perform two assessments to check for the validity of the proportional hazard assumption. The first is a statistic test for the corre-

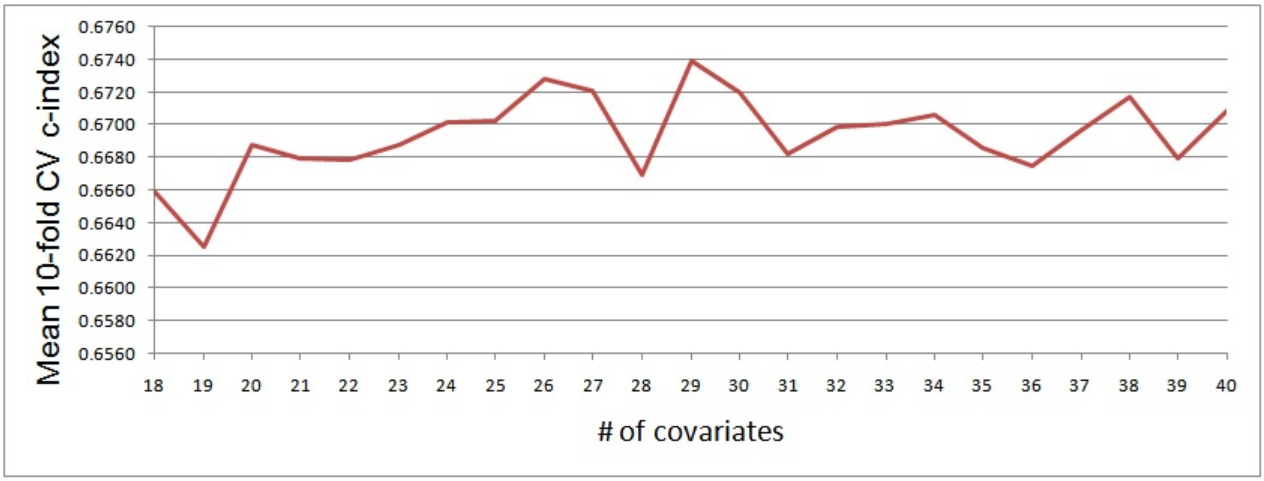

Figure 5.1: Mean c-index of 10-fold cross validation as a function of number of covariates

lation between the Schoenfeld residual and ranked survival time. Schoenfeld residual can be thought of as the observed minus the expected values of the covariates at each failure time. Per [29], a correlation ρ of zero supports the proportional hazards assumption (the null hypothesis). Table 5.2 clearly indicates all 29 covariates have correlation nearly zero. The last row reported as 'GLOBAL' represented the entire model with all 29 covariates, therefore a correlation value is inappropriate and shown as 'NA'. In addition, the p-value for all covariates are greater than 0.05 and therefore are statically insignificant, providing further evidence in support of the proportional hazard assumption.

Aside from the statistical test, Grambsch and Therneau [18] provided a graphical means to discern the validity of the proportional hazard assumption. An advantage of this graphical approach is that it provides insight into the temporality and extent of non-proportionality that otherwise is impossible to glean from a statistical test. Grambsch and Therneau showed that these statistical tests are essentially a test of non-zero slope in the graph of Schoenfeld residuals versus time. Figures 5.2 shows the scaled Schoenfeld residuals versus log time for all the 29 covariates. We can see the solid line in each window is often close to being flat. Also, the random pattern of the residual at each failure time indicates covariate effect is not changing with respect to time, which is the definition of the proportional hazard assumption.

Covariate	ρ	chi-sq	p-value
LDH	-3.14E-02	5.10E-01	.475
ALP	5.24E-02	1.46E+00	.226
HB	8.63E-03	4.19E-02	.838
HEMAT	-2.67E-02	3.83E-01	.536
RBC	5.12E-02	1.58E+00	.208
AST	-2.41E-02	3.57E-01	.550
PSA	4.26E-02	1.12E+00	.290
ALB	4.04E-02	9.65E-01	.326
PULSE	7.24E-02	2.22E+00	.136
NA.	-1.42E-03	1.40E-03	.970
ECOG_C	1.36E-02	1.13E-01	.737
SYSTOLICBP	2.08E-03	1.13E-01	.960
ANALGESICS	-2.20E-02	2.86E-01	.593
DIASTOLICBP	-.13E-05	4.62E-06	.998
LYMperLEU	-4.12E-02	9.79E-01	.322
BMI	1.074E-02	6.11E-02	.805
NEUperLEU	-4.53E-02	1.21E+00	.271
LYM	-4.02E-02	9.11E-01	.340
WEIGHTBL	-2.39E-02	3.14E-01	.575
CREACL	-3.35E-03	6.08E-03	.938
TARGET	-2.87E-02	4.54E-01	.501
PHOS	7.88E-03	2.95E-02	.864
CA	-2.68E-02	4.48E-01	.503
LIVER	-2.81E-02	4.19E-01	.517
CCRC	8.41E-03	3.98E-02	.842
MG	-6.76E-03	2.32E-02	.879
BONE	-5.21E-02	1.46E+00	.228
RegionEastEuro	-3.86E-02	8.21E-01	.365
MHCARD	-1.82E-02	1.87E-01	.665
GLOBAL	NA	1.83E+01	.939

Table 5.2: Statistical test for proportional hazard assumption

5.4 Experimental Setting

The hyperparameters of the models are determined as follows. For the CoxPH model, the Efron approximation is used for tied death times as it is more accurate and computationally more efficient [43]. For GBM, number of trees is 1500, tree depth is 5, bag fraction is .8, and shrinkage is .002. For GBMCI, number of trees is 2000, tree depth is 2, bag fraction is .8, and shrinkage is 1. Initial hyperparameter settings for GBM and GBMCI were based on [9] and fine tuned for this data set. Risk prediction from our ensemble models are then submitted to the DREAM Challenge website for scoring, where a random subset of the predictions are used to calculate the score c-index and iAUC to prevent overfitting.

The non-batch ensemble model is denoted "CBCL". The batch ensemble model is denoted "CBCL-Batch". All simulation was carried out with the R System for Statistical Computing (version 3.2.1, [36]).

5.5 Results

We evaluate our models against one baseline model and the top five performing models of the DREAM Challenge described in Sections 3.2 through 3.7. All five top performing models from the DREAM Challenge outperformed the baseline model handily.

Each model is tested using the 313 patients from the final scoring round, with the results for the comparative models taken directly from the published DREAM Challenge's final result webpage (available online at `https://www.synapse.org/#!Synapse:syn2813558/wiki/232674`). Results of our model are shown in Table 5.3. As the random subsampling of the predictions used to evaluate the scores is hidden from participants, it is impossible to know which prediction was correct and which were not.

Method	c-index	AUC(12month)	AUC(18month)	AUC(24month)	iAUC
CBCL-Batch	**.7337**	**.8082**	.7715	.7874	.7710
CBCL	.7321	.7961	**.7734**	.7796	.7696
Laajala et al.	.7307	.7918	.7674	**.8388**	**.7915**
Xiao et al.	.7263	.7708	.7663	.8147	.7789
Song and Wang	.7157	.7492	.7645	.8369	.7778
Wolfinger and Chu	.7212	.7713	.7553	.8085	.7758
Greiner et al.	.7205	.7577	.7620	.8258	.7743
Halabi (baseline)	.6989	.7418	.7375	.7634	.7429

Table 5.3: Result comparison with the state-of-the-art models from DREAM Challenge.

5.6 Discussion

As can be seen from Table 5.3, our ensemble models out performs the previous generation of models for c-index, AUC at 12 months, and AUC at 18 months. The top performing model from the DREAM Challenge, Laajala et al., remains the best performer under AUC at 24 month and iAUC.

In terms of the c-index performance measure, even without considering batch effect, our ensemble framework offers a 4.75% improvement over baseline and .20% improvement over the top performing model of the DREAM Challenge. When considering batch effect, the performance improves to 5.04% and .41% over the baseline and top performing model respectively. As our model is the only one that took advantage of direct optimization of the c-index loss function, this result is not at all surprising. This shows that our ensemble framework generally performed well in terms of relative ranking of patient.

The time dependent AUC scores at 12 and 18 months are scored highest by our batch and non-batch model respectively. The risk predictions from our ensemble models at these time specific date are based on our global risk prediction. By using the proportional hazard assumption, we assume the relative risk would remain the same irrespective of time and thus the concordance between patients would be the same at all time points. This assumption seems to be supported by the results at 12 and 18 months.

With regard to the AUC at 24 months and iAUC, the Laajala et al. model continues to perform best. This was rather surprising as we had expected our model to perform better given that we had better results under c-index and the clear relationship between c-index and time dependent AUC under a certain formulation as shown in subsection 2.4.3.

Since we are not provided access to the ground truth data, we will compare our prediction to those of Laajala et al.'s to further gauge insight into this surprising result. A plot of our CBCL-batch model's normalized prediction vs Laajala's normalized prediction is shown in Figure 5.3. The Pearson correlation is calculated as $r = .5171$. It is evident there is a moderate linear correlation between the predictions of these two models from the Pearson correlation. Whereas Laajala's predictions tend to be spread across the normalized range evenly, our batch model's predictions are mostly concentrated at the lower end with few outliers towards the middle and top end of the range.

Since ranking and concordance are invariant to affine transformation, the Spearman correlation may better offer insight into the result discrepancy. The Spearman correlation is calculated as $r_s = .7799$. The Spearman correlation, which looks at the monotonic relationship between the two predictions, shows that we have concordance at about 78% of the sets. This provides strong evidence why the result of c-index and iAUC between our batch model and Laajala et al.'s model are close to each other.

From the formulation of *cumulative/dynamic* iAUC and *incident/dynamic* c-index in Section 2.4, it is clear that AUC and iAUC scores would be higher than c-index as the cumulative TPR considers all patients to be cases when $t \geq T_i$ instead of only when $t = T_i$ as with incident TPR. We believe our batch model scoring highest on the c-index measure implies our model achieved the most accurate concordance ranking with respect to the ground truths. As Heagerty [25] pointed out, *cumulative/dynamic* accuracy summaries are appropriate when a specific time t is important, that is discriminating between subjects who die prior to a given time t and those that survive beyond time t. Figure 5.3 shows our predictions tend to be

concentrated at the low end of the normalized risk scale. We hypothesize this concentration leads to more patients being mis-classified as cases earlier than what the ground truth would dictate. This hypothesis is supported by the results from Table 5.3 where CBCL-Batch scores highest at c-index (essentially at time 0 month) and 12 month on the AUC measure. Then at 18 month, it is overtaken by the CBCL model and ultimately by the Laajala et al. model at 24 month and iAUC (which measured the integrated AUC from 6 to 30 months).

5.7 Conclusion

We presented an ensemble prognostic model for the task of predicting risk for mCRPC patients. This was achieved by combining a Cox proportional hazard model, gradient boosting machine, and gradient boosting machine with direct c-index optimization. We demonstrated that our model outperformed all current mCRPC prognostic models on the c-index performance measure as well as AUC at 12 and 18 month, and performed competitively on the AUC at 24 month and iAUC measures.

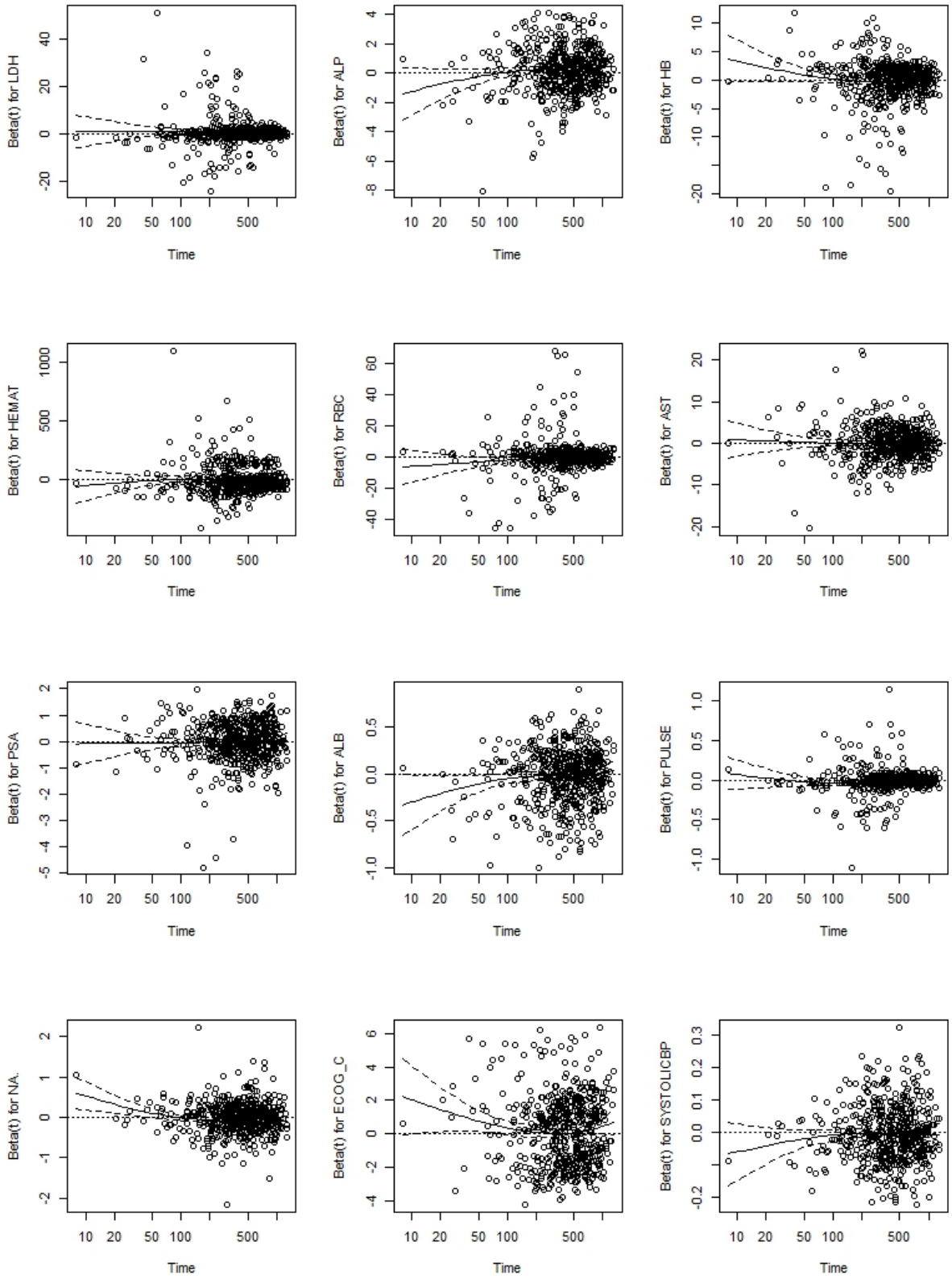

Figure 5.2: Scaled Schoenfeld residual vs log survival time

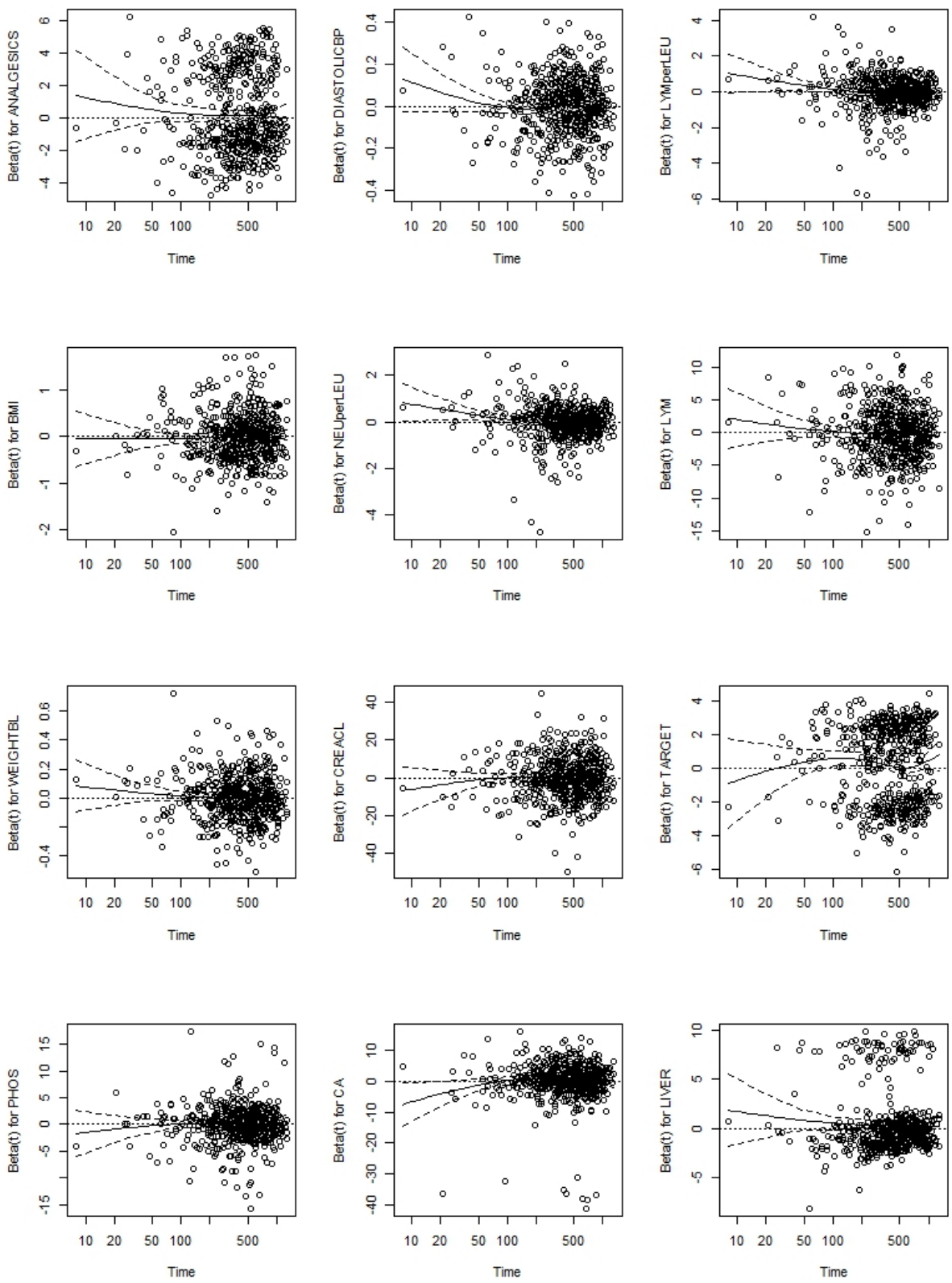

Figure 5.2: Scaled Schoenfeld residual vs log survival time (con't)

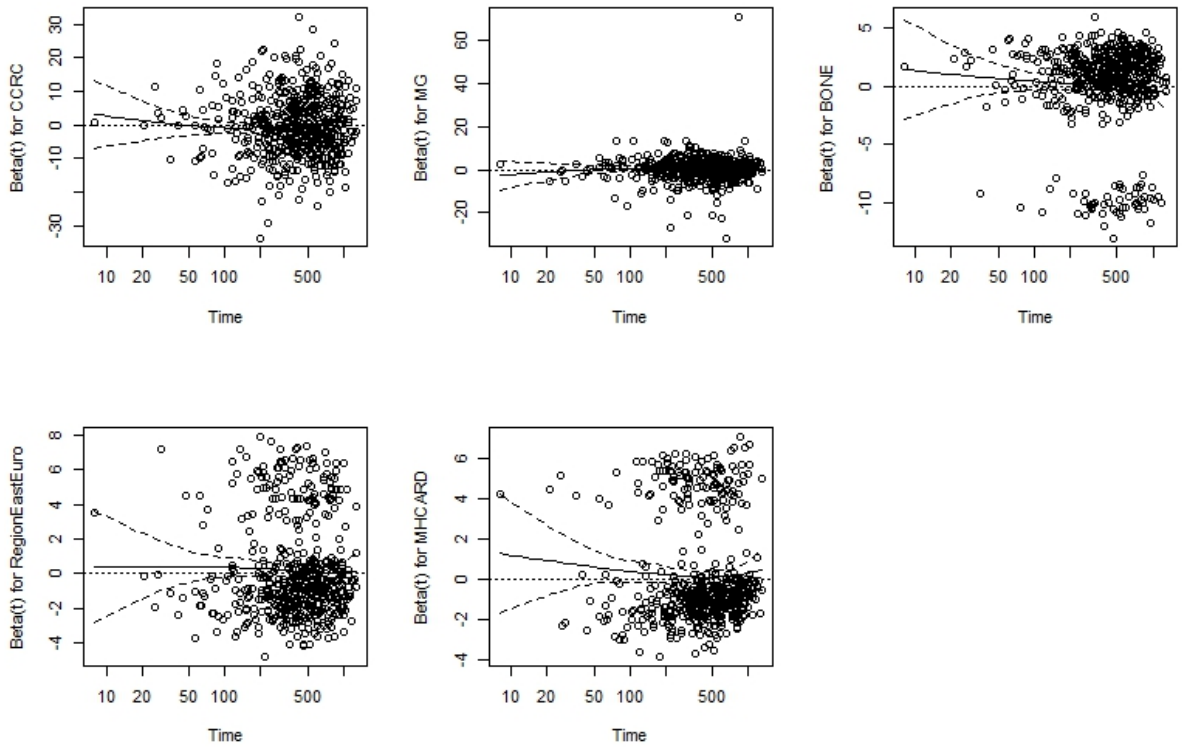

Figure 5.2: Scaled Schoenfeld residual vs log survival time (con't)

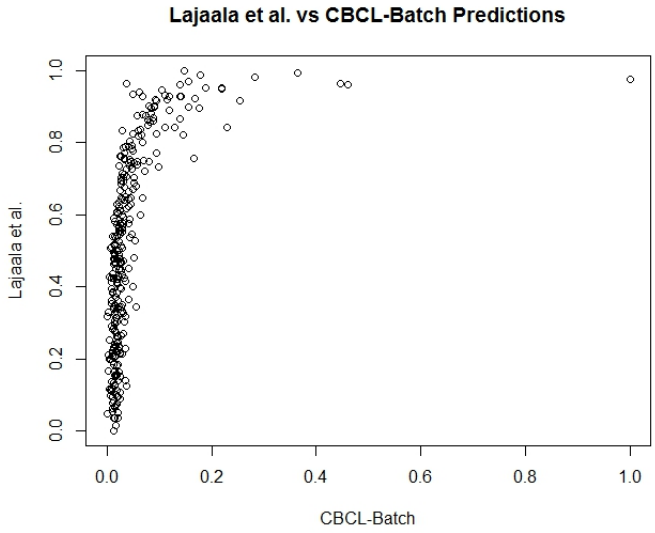

Figure 5.3: Comparison between normalized predictions from CBCL-Batch and Lajaala et al. models

Chapter 6

Conclusion

This thesis proposed an ensemble prognostic model to predict risk for mCRPC patients treated with docetaxel. Our approach outperformed the baseline and current state-of-the-art models on the c-index metric and AUC at 12 and 18 month, and is competitive on the iAUC metric and AUC at 24 month. This chapter summarizes our contributions and suggests future directions for the ideas presented in this thesis.

6.1 Summary of Contribution

We utilized the concept of ensemble methods to synthesize a prognostic model for mCRPC risk prediction. Starting with the simple Cox Proportional Hazard model, we improved on it by incorporating gradient boosting machine and a variant of gradient boosting machine that directly optimizes a smooth approximation of the c-index loss function. As c-index is one of the most common metrics for assessing discriminatory power of prognostic models, it naturally made sense to incorporate c-index as a loss function over which our model was optimized. This allowed our ensemble model to beat the current state-of-the-art mCRPC

prognostic models under the c-index measure and AUC at 12 and 18 month, while performing competitively on the iAUC measure and AUC at 24 month.

6.2 Future Work

With the current framework performing as well as it is, there is still room for additional improvement to pursue in the future. One of the major areas where we can improve on is the covariate selection step. While the proposed method used in this thesis has the desirable characteristic of automatic covariate selection, it only considers one individual covariate at a time, thus neglecting any potential confounding factors. In the paper by Mayr et al. [34], they proposed a boosting framework to estimate the optimal set of covariates that maximizes a smooth approximation of the c-index. Essentially they minimize the empirical risk by using gradient descent in the function space where the function space is spanned by weak-learners, of which each weak learner is a simple linear model containing one possible covariate. The final ensemble learner effectively constitutes an optimal combination of these covariates. By incorporating this method into the covariate selection step, we can extend our model to encompass an end-to-end framework based on c-index optimization and really aim to improve over the state-of-the-art models on the c-index performance metric.

Another possible direction is to optimize the prognostic model for *cumulative/dynamic* AUC directly which has not been explored. Zou et al. [49] proposed an *incident/dynamic* AUC optimization framework by posing the survival time prediction problem as a binary classification problem. They used a non-parametric estimation of the *incident/dynamic* AUC and further formulated the constraints into a linear programming problem, of which is solvable by any of the many existing efficient algorithms. As the definition of *cumulative/dynamic* AUC is quite similar to *incident/dynamic* AUC, it seems plausible that an extension can be made to Zou et al.'s framework to directly optimize a non-parametric estimation of

cumulative/dynamic AUC. If this is possible, this framework could theoretically offer improvement over existing models on the iAUC performance metric.

Bibliography

[1] CANCERMondial. http://www-dep.iarc.fr/, 2010.

[2] DREAM9.5 - Prostate Cancer DREAM Challenge. https://www.synapse.org/#!Synapse:syn2813558/wiki/70844, 2015.

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

[4] Armstrong, A. J., Garrett-Mayer, E. S., Yang, Y.-C. O., de Wit, R., Tannock, I. F., and Eisenberger, M. A contemporary prognostic nomogram for men with hormone-refractory metastatic prostate cancer: a tax327 study analysis. *Clinical Cancer Research 13*, 21 (2007), 6396–6403.

[5] Becker, L., and Birzgalis, E. Orchiectomy in the management of adenocarcinoma of the prostate. *Surgery, gynecology & obstetrics 122*, 4 (1966), 840.

[6] Blanche, P., Latouche, A., and Viallon, V. Time-dependent auc with right-censored data: a survey. In *Risk Assessment and Evaluation of Predictions*. Springer, 2013, pp. 239–251.

[7] Borgmann, V., Hardt, W., Schmidt-Gollwitzer, M., Adenauer, H., and Nagel, R. Sustained suppression of testosterone production by the luteinising-hormone releasing-hormone agonist buserelin in patients with advanced prostate carcinoma: A new therapeutic approach? *The Lancet 319*, 8281 (1982), 1097–1099.

[8] Chen, Y. GBMCI: Generalized Boosted Regression Model. R package version 1.6-3.2. https://github.com/uci-cbcl/GBMCI, 2013.

[9] Chen, Y., Jia, Z., Mercola, D., and Xie, X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and mathematical methods in medicine 2013* (2013).

[10] Cookson, M. S., Roth, B. J., Dahm, P., Engstrom, C., Freedland, S. J., Hussain, M., Lin, D. W., Lowrance, W. T., Murad, M. H., Oh, W. K., et al. Castration-resistant prostate cancer: Aua guideline. *The Journal of urology 190*, 2 (2013), 429–438.

[11] Cox, D. R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society 34* (1972), 187–220.

[12] Cox, D. R. Partial likelihood. *Biometrika 62*, 2 (1975), 269–276.

[13] El-Amm, J., Freeman, A., Patel, N., and Aragon-Ching, J. B. Bone-targeted therapies in metastatic castration-resistant prostate cancer: evolving paradigms. *Prostate Cancer 2013* (2013).

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

[15] Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis 38*, 4 (2002), 367–378.

[16] Gillen, D. Survival analysis. Lecture notes, 2016.

[17] Gönen, M., and Heller, G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika 92*, 4 (2005), 965–970.

[18] Grambsch, P. M., and Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*, 3 (1994), 515–526.

[19] Greiner, R., Gofer, E., Kumar, L., Vega, R., Patel, J., and Tahmasebi, N. Learn-pc report (dream challenge). `https://www.synapse.org/#!Synapse:syn3822697/wiki/231039`, 2015.

[20] Gupta, E., Guthrie, T., and Tan, W. Changing paradigms in management of metastatic castration resistant prostate cancer (mcrpc). *BMC urology 14*, 1 (2014), 1.

[21] Halabi, S., Lin, C.-Y., Kelly, W. K., Fizazi, K. S., Moul, J. W., Kaplan, E. B., Morris, M. J., and Small, E. J. Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *Journal of Clinical Oncology 32*, 7 (2014), 671–677.

[22] Halabi, S., Lin, C.-Y., Small, E. J., Armstrong, A. J., Kaplan, E. B., Petrylak, D., Sternberg, C. N., Shen, L., Oudard, S., De Bono, J., et al. Prognostic model predicting metastatic castration-resistant prostate cancer survival in men treated with second-line chemotherapy. *Journal of the National Cancer Institute* (2013), djt280.

[23] Halabi, S., Small, E. J., Kantoff, P. W., Kattan, M. W., Kaplan, E. B., Dawson, N. A., Levine, E. G., Blumenstein, B. A., and Vogelzang, N. J. Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. *Journal of Clinical Oncology 21*, 7 (2003), 1232–1237.

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

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

[26] Hotte, S., and Saad, F. Current management of castrate-resistant prostate cancer. *Current Oncology 17* (2010), S72–S79.

[27] Huggins, C., and Hodges, C. V. Studies on prostatic cancer: I. the effect of castration, of estrogen and of androgen injection on serum phosphatases in metastatic carcinoma of the prostate. *The Journal of urology 167*, 2 (2002), 948–951.

[28] Hung, H., and Chiang, C.-T. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics 38*, 1 (2010), 8–26.

[29] Kleinbaum, D. G., and Klein, M. *Survival analysis: a self-learning text*. Springer Science & Business Media, 2006.

[30] Laajala, T. D., Khan, S., Airola, A., Mirtti, T., Pahikkala, T., Gopalacharyulu, P., and Tero, A. Predicting patient survival and treatment discontinuation in DREAM 9.5 mCRPC challenge. `https://www.synapse.org/#!Synapse:syn4227610/wiki/233734`, 2015.

[31] Lin, H.-c., Baracos, V., Greiner, R., and Chun-nam, J. Y. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems* (2011), pp. 1845–1853.

[32] Locke, J. A., Guns, E. S., Lubik, A. A., Adomat, H. H., Hendy, S. C., Wood, C. A., Ettinger, S. L., Gleave, M. E., and Nelson, C. C. Androgen levels increase by intratumoral de novo steroidogenesis during progression of castration-resistant prostate cancer. *Cancer research 68*, 15 (2008), 6407–6415.

[33] Martinussen, T., and Scheike, T. H. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007.

[34] Mayr, A., and Schmid, M. Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. *PloS one 9*, 1 (2014), e84483.

[35] Montgomery, R. B., Mostaghel, E. A., Vessella, R., Hess, D. L., Kalhorn, T. F., Higano, C. S., True, L. D., and Nelson, P. S. Maintenance of intratumoral androgens in metastatic prostate cancer: a mechanism for castration-resistant tumor growth. *Cancer research 68*, 11 (2008), 4447–4454.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[37] Ridgeway, G. The state of boosting. *Computing Science and Statistics* (1999), 172–181.

[38] Ridgeway, G. Generalized boosted models: A guide to the gbm package. *Update 1*, 1 (2007), 2007.

[39] Siegel, R., Naishadham, D., and Jemal, A. Cancer statistics, 2013. *CA: a cancer journal for clinicians 63*, 1 (2013), 11–30.

[40] Smaletz, O., Scher, H. I., Small, E. J., Verbel, D. A., McMillan, A., Regan, K., Kelly, W. K., and Kattan, M. W. Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration. *Journal of Clinical Oncology 20*, 19 (2002), 3972–3982.

[41] Song, L., and Wang, D. Write-up for prostate cancer dream 9.5 challenge. `https://www.synapse.org/#!Synapse:syn4732218/wiki/234510`, 2015.

[42] Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems* (2008), pp. 1209–1216.

[43] Therneau, T. A package for survival analysis in s. r package version 2.37-4. *URL http://CRAN. R-project. org/package= survival. Box 980032* (2013), 23298–0032.

[44] van Houwelingen, H., and Putter, H. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

[45] Wolfinger, R., and Chu, T.-M. Ensemble selection modeling of the dream prostate cancer challenge data. `https://www.synapse.org/#!Synapse:syn4732827/wiki/234504`, 2015.

[46] Wu, J. N., Fish, K. M., Evans, C. P., deVere White, R. W., and Dall'Era, M. A. No improvement noted in overall or cause-specific survival for men presenting with metastatic prostate cancer over a 20-year period. *Cancer 120*, 6 (2014), 818–823.

[47] Xiao, J., Wang, S., Shang, J., Han, J., and Peng, J. Applying weights to Cox model improves reliability of prostate cancer survival prediction. `https://www.synapse.org/#!Synapse:syn4732274/wiki/234503`, 2015.

[48] Yan, L., Verbel, D., and Saidi, O. Predicting prostate cancer recurrence via maximizing the concordance index. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 479–485.

[49] Zou, M., Liu, Z., Zhang, X.-S., and Wang, Y. Ncc-auc: an auc optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data. *Bioinformatics 31*, 20 (2015), 3330–3338.