# SURVIVAL ANALYSIS FOR BIG DATA

A Dissertation

by

KAHKASHAN AFRIN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Satish T. S. Bukkapatnam |
| Co-Chair of Committee, | Bimal Nepal |
| Committee Members, | Yu Ding |
| | Bani K. Mallick |
| Head of Department, | Lewis Ntaimo |

August 2020

Major Subject: Industrial Engineering

# ABSTRACT

Survival analysis has emerged as a promising tool in biostatistics for life expectancy prognosis and personalized healthcare. However, its accuracy and potential are limited by the modern big data challenges to which the traditional survival analysis models have not yet adequately adapted. This refers to the data laced with challenges of volume, variety, velocity, and veracity. In this dissertation, we are concerned with the challenges of data imbalance—veracity and multi-view data—variety and volume. To achieve the overarching goal of improving prognosis accuracies, this dissertation was aimed at proposing methodological improvements and leveraging statistical advancements for solving the big data challenges in survival analysis and addressing the limiting assumptions of the most commonly used proportional hazard models.

Firstly, we address the data imbalance issue by proposing a balanced random survival forest (*BRSF*) model that integrates a synthetic minority over-sampling technique with random survival forests for improved mortality prediction. Secondly, for the multi-view survival learning challenge, we proposed an integrated non-parametric survival (*iNPS*) learning method that captures the joint and individual structures in different data types and models their non-linearity and interactions by using a non-parametric survival learning method. Theoretical results and extensive empirical comparisons using complex cancer and cardiovascular data sets suggests major improvements in the survival prognosis accuracy due to the methods presented in this dissertation.

Finally, we extend non-parametric survival learning to multiple recurring events for continuous prediction of epileptic seizures and to provide probabilistic estimates for seizure onset over a broad prediction horizon. These estimates are essential for developing individualized quantitative risk measures and management plans for epilepsy patients and potential application in a wearable seizure alert system.

We believe that that the methodological advancements and their clinical applications presented in this study may provide a foundation for further knowledge discovery and subsequent improvement in survival analysis—a healthcare domain of immense importance.

# DEDICATION

To my family and my best friend, Ashif, for their unconditional love, support, and sacrifice. It is their light that is shining through my world.

*"And the Sun burns to light up the Earth. It never stops, for his every burn shines a light somewhere."*

—Kahkashan

# ACKNOWLEDGMENTS

I would like to thank my advisors, Dr. Satish Bukkapatnam and Dr. Bimal Nepal, for their help and guidance for my research. They provided me with unprecedented opportunities to explore several research projects, teaching, and mentoring, that greatly helped me to expand my knowledge base in the past four years. I also want to thank my committee members, Dr. Yu Ding and Dr. Bani Mallick, for their continuous support and valuable time towards my research. It was the incredible support of my advisors and committee members that enabled me to finish this dissertation amid a pandemic!

My sincerest gratitude goes to Dr. Natrajan Gautam at Texas A&M for his nonjudgmental mentorship and positive words of encouragement that have, at times, provided me the emotional support to keep going and to Dr. Manoj Kumar Tiwari, Director of the National Institute of Industrial Engineering, Mumbai for introducing me to research and encouraging me to pursue Ph.D. in the United States. PhDs in this world need more mentors like them, I certainly do!

This acknowledgment would be incomplete without thanking my former and current lab members —Dr. Vu Nguyen, Dr. Trung Le, Dr. Hoang Tran, Dr. Zimo Wang, Bhaskar Botcha, Akash Tiwari, and T. Ganatma Nakkina for being there in the good, the bad, and the ugly. Throughout my Ph.D. journey, my friends, both here in the U.S. and back in India, have helped me in keeping my spirits high and anxiety low. They have equally supported me, shared my joys and sorrows, and made my Ph.D. a hell lot more fun.

It goes without saying, but I would like to thank my family —my parents and my uncle for going against the societal norms to let me study and choose my own path and for loving me so dearly. Despite our difficult socio-economic conditions, they have always put my studies on top of everything. Every bit of any success that I ever had is a manifestation of their prayers. I am thankful for the love of my beautiful little sisters and my brother, Altamash, who are always at the core of my heart. Finally, I would like to express my gratitude and love for my childhood best friend and partner, Ashif, who has been my biggest support system and a great mentor. He has not only inspired me to be a better researcher but also a better person.

Last but not least, I would like to thank all anonymous patients and research staff who participated in the clinical studies and data collection, both at the Texas A&M University and the online data sets used in this study. It is their contribution, often unacknowledged, that is enabling the transformation of healthcare, medicine, and bioinformatics.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supervised by a dissertation committee consisting of Dr. Satish Bukkapatnam of the Department of Industrial and Systems Engineering, Dr. Bimal Nepal of the Department of Industrial Distribution, Dr. Yu Ding of the Department of Industrial and Systems Engineering, and Dr. Bani Mallick of the Department of Statistics.

A part of the initial analyses depicted in Chapter 2 and Chapter 4 were conducted by Gurudev Illangovan and Revanth Dusi, respectively, of the Texas A&M Department of Industrial and Systems Engineering.

All other work conducted for the thesis was completed by me independently.

## Funding Sources

Part of this graduate study was supported by the National Science Foundation's PFI-AIR-TT 1543226 grant. I would also like to acknowledge the assistantship provided by the Industrial Distribution Program and the Department of Engineering Technology & Industrial Distribution to support my graduate studies at Texas A&M.

## TABLE OF CONTENTS

# LIST OF FIGURES

FIGURE Page

# LIST OF TABLES

TABLE Page

# 1. INTRODUCTION

## 1.1 Motivation

The field of *survival analysis* encompasses methods and tools to predict the *time-to-event* of a process, which is the length of time until the occurrence of an event [1]. Historically, the term “event” used to refer to the health outcome or change of status such as death, owing to its root in the clinical trial literature and life table analysis [2, 3]. Since then, survival analysis has emerged as an effective prognosis tool in numerous domains besides medical research and trials, e.g., machine failure prediction in complex manufacturing systems [4], market failure and crash prediction in business intelligence [5, 6], and assessment of healthcare utilization [7]. Specifically, in the biostatistics/bioinformatics domain, establishing the survival prognoses for patients is essential. It is utilized for varied important tasks such as mortality prediction, risk stratification, and key biomarker identification in acute and high-risk patients [8, 9]. Their importance is further accentuated while dealing with life-threatening conditions such as acute cardiovascular diseases, neurological disorders, cancer, and diseases that may extend for several years [10].

This broad and crucial utilization has been facilitated by the development of various survival analysis models. All parametric, semi-parametric, as well as non-parametric survival analysis models, are available in the literature [11, 12]. Parametric survival regression models are used to study the effect of covariates/features on the event time. Let $\boldsymbol{x}_i$ be a set of features for individual $i$; it’s effect on the event time $T$ can be modeled by using a hazard rate, $\lambda_i(t)$ at time t. The parametric survival regression model then reduces the estimation of this hazard function, $\lambda_i(t)$ at time $t$ into a regression problem of the form $\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{x})$, where $\boldsymbol{\beta}$ is an unknown vector of the regression coefficients and the features have a multiplicative effect. The baseline hazard function,

$\lambda_0(t)$ within parametric models, attaches to the event time T through the underlying parametric distribution of the event time, for example, if $\lambda_0(t)$ is a constant $c$, then the resulting survival regression model becomes exponential.

Parametric models are sometimes preferred for their ease of interpretability. However, they can be used only when the underlying distribution of the event time data can be correctly identified apriori. This has been increasingly infeasible with the increased data complexity. Incorrect distributional assumption can make the parametric model fitting, and hence their interpretation, biased.

The semi-parametric survival analysis models such as Cox proportional hazard (CPH) alleviates the need for specifying distributional assumption, i.e., CPH does not require a distributional assumption of $\lambda_0(t)$ to estimate $\boldsymbol{\beta}$ [13]. Since its proposal in 1972, CPH has been used in almost all application areas and remains the gold standard for survival analysis [14], so much so that survival analysis has now almost become synonymous to CPH. However, this domination, perhaps disproportionately, by the celebrated CPH model, is not warranted without limitations. The CPH model makes certain restrictive assumptions [15], many of which do not hold in real-life scenarios. Two such assumptions are:

1. Linearity assumption: here the log-risk is assumed to be a linear combination of the features, i.e., $log(\lambda_i(t)) = log(\lambda_0(t)) + \boldsymbol{\beta}^\top \boldsymbol{x}$. Furthermore, it also does not take into account the missing features or their interaction effects. Interaction can be introduced by using brute force to incorporate the interaction as a new feature. With the increasing number of features, neither linearity can be satisfied, nor interactions can be incorporated successfully as the number of high-level interaction terms grows. When dealing with large data, these cause a severe decrease in accuracy, basically rendering the CPH model useless.

2. Proportional hazard assumption: here, the hazard ratio between any two patients is assumed to be constant at every time instant, $t$. The growing complexity of

application, such as prognosis of survival in cancer patients certainly makes this assumption overly simplistic for real-life applications.

Despite the tremendous growth of survival analysis applications, it's methodological advancement has been relatively slow compared to the machine learning regression and classification problems. Nonetheless, there are two key machine learning survival models (referred to as *survival learning* henceforth in this dissertation) developments that have addressed some of the traditional survival model's limitations. These methodologies for predicting time-to-event, based on advanced machine learning techniques, have opened new possibilities to overcome the limitations of CPH models [15, 16].

To solve the problem of linearity, non-linear survival learning models have resulted from the application of deep learning for survival analysis. Nonetheless, their CPH based activation function can still be a limiting factor [16]. Another method is random survival forests (RSF) [15], a non-parametric approach for producing ensemble estimates of the hazard, based on a Breiman's ensemble tree, random forests model. Since RSF is virtually free from the limiting assumptions of CPH and is determined to be the current state of the art in survival learning, we later explore and compare it with CPH for complex applications with real-world data.

Although the new survival learning models address the limitations of traditional proportional hazard models, the Big Data related challenges in the biological survival data sets have been largely ignored, resulting in decreased accuracy of prognosis [17]. Since survival analysis models are one of the key tools which provide decision support to the physicians, not only for intervention selection but also for identifying high-risk patients and for patient counseling/consent for intervention [18, 19, 20], the limited accuracy of current survival models which guide critical life and death decisions is a major concern.

Accurate prognosis of the event and the effect of features can allow risk-calibrated interventions for better management of the outcomes. Thus the overarching goal of this dissertation is to improve the healthcare data analysis outcomes by focusing on two aims: (1) increasing the accuracy of survival prognosis by exploring the survival learning methods and proposing methodological development to address the data challenges for survival analysis and (2) its application in critical healthcare areas such a cardiac disease and cancer to improve mortality prediction. Application is also extended to beyond the time-to-death prediction by applying survival learning for epileptic seizure prediction where the events are multiple/recurring seizure episodes.

## 1.2 Challenges

A key limitation to the survival model's accuracy is the challenges posed by the exponential increase of Big Data in healthcare. During the last decade, we have seen tremendous improvements in sensing and data collection technologies. These improvements have led to the exponential growth in data collection, especially in biomedical research, where it has facilitated an easy collection of high-throughput data through technologies such as genome sequencing, imaging, multichannel physiological monitoring, and continuous longitudinal data using wearables. It has consequently led to growing volume, complexity, veracity, and variety of the survival data. It is fair to assume that most of the survival data now fits the characteristics of Big Data [21] and can be referred to as *healthcare big data.*

When analyzing healthcare big data, it becomes challenging to filter the relevant knowledge and build a useful model. Prior efforts in building a useful model from Big Data to address the challenge of volume, veracity, variety, etc. have engendered some innovative and indeed useful methodologies [22, 23]. However, the most primarily benefit the traditional regression, classification, and clustering domains where survival analysis has been largely ignored. In this dissertation, we explore and address two specific problems related healthcare big data to build more accurate survival models: (1) Data imbalance (related to veracity) and (2) Multi-view survival data (related to variety and volume). Below we describe these problems with respect to the survival data and our research objectives to address them.

### 1.2.1 Data imbalance

The first challenge we explore and address in this dissertation is data imbalance—related to the veracity of the healthcare big data. The survival learning techniques used for survival status and time-to-event prognosis can become biased and undermine hazard estimation when there is an imbalance in the survival data. Here imbalance refers to a significantly unequal representation or dominance of one class over the other. For example, the majority of the data samples are either in censored/survival class or death/mortality class. For example, the imbalance shown in Figure 1.1 is representative of one of the real dataset used in this paper (the STEMI dataset) which consists of tracking 267 patients for their mortality over a period of 1 year, out of which 62 (only 23%) belong to the minority class (i.e., suffered mortality). Such situations are common and often beyond control in clinical studies and bioinformatics. The accuracy of survival prognosis further decreases with a high-dimensional low sample size (HDLSS) data. However, obtaining real clinical data with large sample size is again beyond control in survival analysis, especially when the study focuses on patients with a specific disease [10]. The conjugation of these two factors significantly worsens the survival learning method's performance.

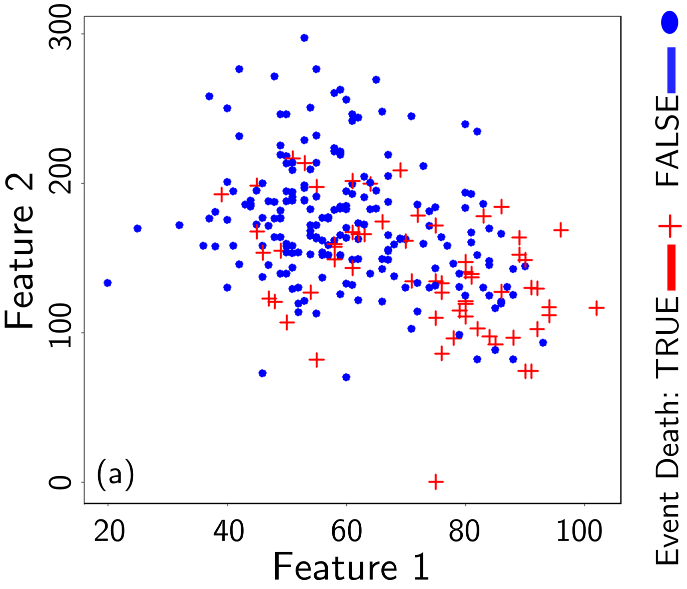

Figure 1.1: Representation of the class imbalance in the feature space (minority class in red).

Data imbalance problem has been recognized in the machine learning literature for almost three decades now [24]. Due to its importance in increasing classification and regression model's accuracy and robustness, it has been studied extensively with many techniques being proposed to balance the class sizes with most having a primary focus on classification tasks [25, 26]. However, it's exploration and application in the survival domain is highly scarce, and more importantly, most of the balancing techniques proposed for the classification literature does not directly apply to the survival domain. For example, one of the most common balancing strategies is to pre-process and manipulate the class distributions to have equal representation by undersampling the majority class and drive the learning algorithm to focus equally on both classes, however with HDLSS survival data, it is infeasible.

Chapter 2 of this dissertation presents theoretical results on the effect of data imbalance on the prognosis accuracy and hazard estimation of survival models. We also obtain empirical performance of balanced and unbalanced survival models for a variety of benchmark and real-life survival data sets. Consequently, a balanced random survival forests (BRSF) method is proposed to address the imbalance.

### 1.2.2 Multi-view data

Data growth in volume and complexity is not very new. However, beyond growing in volume and complexity, survival data has, very recently, also grown in variety. Through simultaneous advancements in a variety of sensing techniques, high dimensional data can now be recorded for the same set of subjects from multiple different sources. Coupled with decreasing data acquisition costs, multi-view data set collection have become a routine part of a patient's standard clinical workflow [27], for example, the 'omics' data such as genomics, transcriptomics, proteomics, etc. measured in cancer studies. Owing to the genome sequencing, this is predominantly known as *multi-omics* data in the bioinformatics literature. However, such data can exist in other biological scenarios such as Electroencephalography (EEG) and fMRI (Functional magnetic resonance imaging) data in neuroinformatics studies as well as multiple sensors data acquisition from wearable devices. Since these different varieties of data sets provide a different view of the same subject/patient, here we use the term *multi-view* to refer to such data and it's constituent data sets are referred to as *data type*.

There are different types of multi-view data [28], and two common types are: (1) vertical multi-view data: multiple data types for the same subject. For example, the breast invasive carcinoma dataset (BRCA) from *The Cancer Genome Atlas (TCGA)* project—a landmark cancer genomics program (https://www.cancer.gov/tcga) [29]. This dataset contains multiple data types, including gene expression, DNA methylation, and micro RNA or miRNA expression data for the same set of patients. We are calling it vertical data as for the same set of patients, different data types can be concatenated vertically, and (2) horizontal multi-view data: same data type measured for a different group of subjects. For example, *The Cancer Proteome Atlas (TCPA)* (http://tcpaportal.org) with protein expression measured for a set of patients with 32 different cancer types. Figure 1.2 presents a pictorial representation of the vertical and horizontal multi-view data. In this dissertation, we will be focusing on vertical multi-view data.

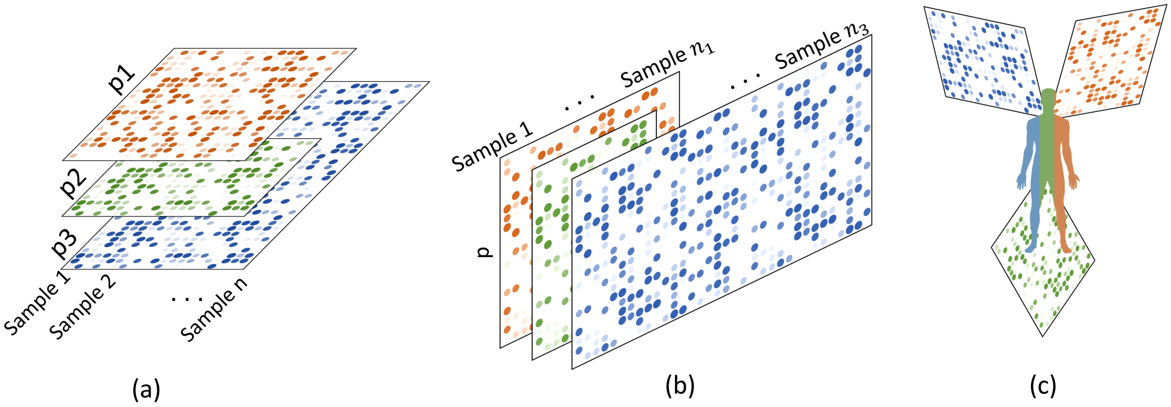

Figure 1.2: A pictorial representation of (a) vertical integration (b) horizontal integration and (c) multi-view representation of the same sample using different data types.

For vertical multi-view data, it is well accepted that analyzing all data types in an integrated form can reveal more information about the patients. By providing multiple views of the same subject from different yet significant biological processes/molecular level, they open up new opportunities for not only understanding biological pathways of disease but also develop personalized treatment for them [30]. Nonetheless, an integrated analysis of multi-view data poses significant challenges [31, 32]. The simplest way is to concatenate the data types. However, when each of the data types is high dimensional (hundreds and thousands of features), their simple concatenation increases the dimension immensely, and with this sheer immensity comes numerous challenges in the data analysis. Some of these challenges are: HDLSS data, redundancy, computational complexity, model overfitting, and multicollinearity- since each data type for the same patient can be inter-related [33]. Hence, an effective and robust data integration method is crucial.

Nonetheless, most of the key integrated multi-view analysis methods are exploratory and unsupervised with very few methods for supervised and even fewer for survival analysis [34]. Furthermore, the current multi-view survival analysis methods have, almost exclusively, utilized parametric or semi-parametric survival analysis methods such as accelerated failure time [35] and CPH, thus imposing limiting assumptions to extremely complex data sets [36, 37].

Hence, towards our overarching goal, in Chapter 3 of this dissertation, we aim to explore the multi-view data integration performance for survival prognosis and to propose a non-parametric multi-view survival learning method for overcoming the limited prognosis accuracy of parametric and semi-parametric models. In summary, this chapter accommodates various healthcare big data challenges, including high-dimensionality and multi-view data integration into one.

## 1.3 Survival Analysis for Prognosis using Smart Wearables

In this dissertation, addressing the healthcare big data challenges has been motivated by the proposed overarching goal of improving the survival analysis accuracy and healthcare outcomes. Here the survival analysis in healthcare is exclusively done in a traditional ‘in-clinic’ or ‘in-hospital’ patient care scenario where patients visit medical practitioners post experiencing disease symptoms and are followed through the duration of study or until an event to assess their survival or outcome of treatment. Nonetheless, in the last decade, we have experienced a dramatic shift in the status-quo of patient care from hospital to at-home, point-of-care (POC) setting to provide accessible quality care in rural and resource-limited areas. This remodeling of patient-care has been possible through the development of wearable sensors/devices. These wearable devices are transforming biomedicine by facilitating continuous, longitudinal health monitoring outside of the clinic [38]. Furthermore, this requires the integration of machine learning with wearable devices to make them *smart wearables* that are capable of not only physiological monitoring but also disease diagnosis, prognosis, and rehabilitation [39] (see fig. 1.3). However, very limited smart wearables are available for prognosis. The limited methods that do exist for prognosis in smart wearables are based on point-prediction and fail to provide a continuous prognosis, which is much needed for smart wearables that are made to be worn continually by the patients.

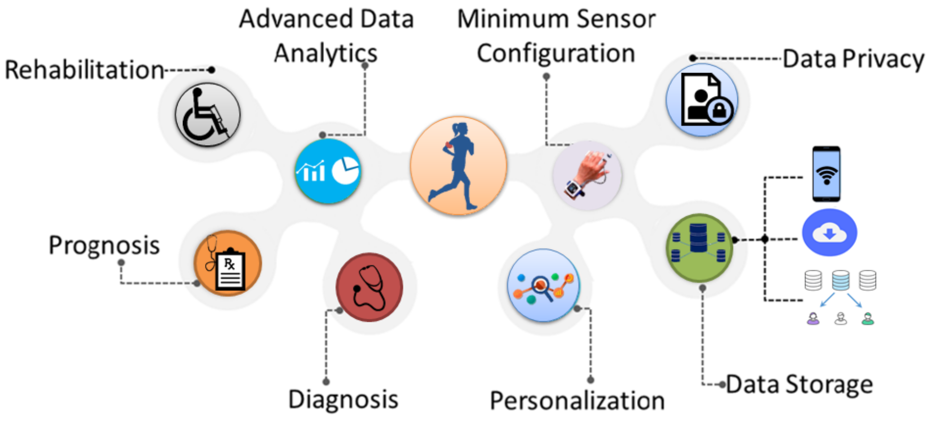

Figure 1.3: A pictorial representation of *smart wearables* with minimum-configuration wireless sensing and embedded advanced analytics. Their integration allows for point-of-care and accurate disease diagnosis, prognosis, and rehabilitation that can shift the status quo of patient care from hospital to in-home setting. Reprinted with permission from [39]. Copyright 2020 by Wiley Periodicals, Inc.

To address this, the final part of this dissertation aims at implementing survival learning for continuous prognosis to facilitate its application in smart wearables. More specifically, we used survival learning for continuous time-to-seizure event prediction using EEG data, where there are multiple/recurring seizure events. Further, this application of the proposed survival analysis method on one of the fastest-growing and one of the most transformative areas of healthcare, i.e., personalized and precision medicine using smart wearables [38], is aimed at significantly improving the broader impact of this dissertation.

## 1.4 Organization of this Dissertation

This dissertation contains four main parts/components:

1. Survival prognosis for imbalanced dataset—addressing data complexity and veracity

2. Survival prognosis for multi-view dataset—addressing volume and variety

3. Extension of survival prognosis for continuous time—series or longitudinal data

4. Conclusion

Each of these parts is discussed in different chapters and is organized as follows. Chapter 2 explores the effect of data imbalance in survival prognosis accuracy. Based on this exploratory insight, a data balancing approach is incorporated along with the survival models to improve the accuracy of prognosis. A comprehensive assessment of several benchmark and case study data sets illustrates a significant improvement in the prognosis accuracy of the balanced models over their unbalanced counterparts.

Chapter 3 explore the survival prognosis accuracy for multi-view data. This chapter consists of two main parts, the first part emphasizes on the effective integration of multi-view data and the second part deals with the use of non-parametric survival learning model to improve survival analysis modeling for multi-view dataset. A real-world dataset is used for comparative assessments to demonstrate the significance of effective multi-view data integration and the efficacy of a non-parametric survival learning model on the integrated data.

Chapter 4 proposes an extension of the use of survival learning beyond the retrospective analysis of right-censored clinical data with a single event, fixed follow-up time, and often time-invariant features to continuous time series data with multiple/recurring events. The significance of this extension is in leveraging the strength of the survival analysis model for continuous outcome prediction and laying the groundwork for event/episode prediction using smart wearable devices. Its efficacy is shown using epileptic seizure prediction on EEG data collected from small mammals.

Finally, chapter 5 summarizes the key findings and contributions and briefly describes the potential future work directions.

# 2. BALANCED RANDOM SURVIVAL FORESTS FOR MORTALITY PREDICTION FROM EXTREMELY UNBALANCED DATA

## 2.1 Introduction

Accuracies of survival models for life expectancy prediction as well as critical-care applications are significantly compromised both due to the use of traditional proportional hazard models and challenge associated with the survival data such as sparsity of samples and extreme imbalance between the survival (usually, the majority) and mortality class sizes. Motivated by the goal of improving survival prognosis accuracy, this chapter focuses on exploring the impact of data imbalance on survival model's performance and the necessary methodological development to address this challenges. As presented in section 1.2, the characteristics of the survival data pose significant challenges to survival models. The presence of extreme imbalance between the survival/censored and the death/mortality classes with as low as 2-10% data in the minority is a commonly occurring, yet often ignored aspect. Due to the contemporary clinical practice and infrastructure across the US, acute cardiac and other life-threatening diseases are mostly treated in small tertiary care hospitals and, as a result, the cohort size tends to be small, further exacerbating this challenge. Balancing is an essential step in maximizing the utility and improved mortality prediction performance. Although data balancing is important, only a few works focus on addressing class imbalance from a survival analysis context [40]. In this chapter, we propose a BRSF survival learning method by integrating a synthetic data balancing scheme with RSF model to improve its prognosis accuracy. We present some key theoretical results on the effect of data imbalance on improving model's predictive performance from a survival analysis context. Additionally, for empirical validation, the performance of the balanced survival models, i.e., the balanced CPH and BRSF are compared to their unbalanced counterparts. Further, the performance of BRSF is also compared to an optimized balanced CPH model. Here, optimized CPH refers to the CPH model where overfitting errors are minimized through predictor selection. All models are assessed on a set of 5 benchmark data sets each representing a different degree of class imbalance, as well as a dataset gathered at the Heart, Artery, and Vein center of Fresno from 267 acute cardiac STEMI (ST Elevated Myocardial Infarction) patients after they underwent cardiac revascularization therapy. In summary, this chapter reports the following three contributions: (1) we developed a BRSF approach to address the challenges with high class imbalance and small data size in survival analysis context, which to our knowledge, has not been previously addressed in the literature, (2) we established theoretical results on why and how balancing the class sizes can improve accuracy of survival prediction and provide results for estimating the relative improvement in survival of model's prediction after balancing the class sizes, and (3) we applied the proposed BRSF model in multiple real data with high imbalance and compared its performance to other contemporary survival models. These contributions collectively can enhance informed treatment decision for healthcare providers. This chapter is organized as follows. In the first section we give a detailed description of the RSF and the synthetic data balancing method used. Here we also discuss the effect of data imbalance on survival prognosis accuracy. The second section provide the details for the survival data sets used in this paper, performance evaluation metrics, and the comparative results obtained before and after addressing the data imbalance. Finally, the last section summarizes our work on addressing the data imbalance in survival analysis.

## 2.2 Balanced Random Survival Forests

Recent advent of innovative methodologies for predicting time-to-event, based on advanced machine learning techniques such as RSF [15] have opened new possibilities to overcome the limitations of CPH models [15, 41]. However, the data related challenges still undermine the performance of these methodologies. In this section, we review the RSF method and explain through theoretical assessment, how data imbalance in survival analysis context can be a detriment to prediction accuracies of RSF.

RSF is a non-parametric approach to right-censored survival analysis based on a Breiman's ensemble tree, random forests model. Growing a random survival forests, $\mathcal{F}$ can be thought of as a hierarchical procedure which initializes by randomly drawing $B$ bootstrap samples from the training data consisting of $N$ samples, each with $R$ predictors (here, features), and growing a survival tree $\{\mathcal{T}_b\}_{1\leq b\leq B}$ for each of the drawn samples (see Figure 2.1). The bootstrap samples are invariably extracted from right-censored survival data. For analyzing survival data, follow up time and associated right censoring are important considerations. Right-censored survival data of $N$ individuals is the collection of values in a set, $\Phi = \{(\boldsymbol{x}_i, T_i, \delta_i)\}_{1\leq i\leq N}$, where the subscript $i$ is the patient index, and $\boldsymbol{x}_i = (x_i^r), i = 1, \ldots, N; r = 1, \ldots, R$ are independent and identically distributed (i.i.d.) features of patient $i$. Let $T_i^0$ and $\mathcal{C}_i$ be the true event (death) and censoring times, respectively for subject $i$. The observed survival time is then given as $T_i = \min(T_i^0, \mathcal{C}_i)$, and $\delta_i := \mathbb{1}_{T_i^0 \leq \mathcal{C}_i}$ is the binary censoring status specified as follows: an individual $i$ is said to be right-censored if $T_i^0 > \mathcal{C}_i$, i.e., $\delta_i = 0$ or else the individual is said to have experienced death at time at time $T_i (\delta_i = 1)$.

Here, the construction of a survival tree, $\mathcal{T}_b$ from the $b^{th}$ bootstrapped data begins with a random selection of $p$ out of $R$ possible features in $\boldsymbol{x}$. Although we used the suggested, $p = \sqrt{R}$ [42, 43], the value of $p$ depends on the number of available features and is data specific. Previous studies have even shown good performance with $p = 1$, care must be taken as an increase in $p$ tend to result in correlated trees [44]. Next, all the $N$ bootstrapped samples are assigned to the root node, i.e., the topmost node of the tree. The root node is then split into two daughter nodes, and each of thus-generated daughter nodes is then recursively split with progressively increasing within-node homogeneity. Now, for any parent node with $p$ features, the split on a given feature, $x^v$ is of the form $x^v \leq \zeta_\gamma^v$ and $x^v > \zeta_\gamma^v; 1 \leq v \leq p$. Here, $\zeta_\gamma^v$ conventionally takes values at the midpoint of consecutive distinct observations of $x^v$ corresponding to the individuals in the parent node being split [45]. Thus, $\gamma$ has at most one less than the parent node size values.

Let $t_{1,q} < t_{2,q} < ... < t_{m,q}$ be $m$ unique event (death) times at a parent node, $q$, and $d_{l,q}$ denote the number of mortality samples in node $q$ at time $\{t_{l,q}\}_{1 \leq l \leq m}$ and $Y_{l,q}$ is the number of individuals who are alive (at risk) in node $q$ at time $\{t_{l,q}\}_{1 \leq l \leq m}$. Similarly, $d_{l,j}$ and $Y_{l,j}$ denote the number of deaths and individuals who are alive (at risk) in the daughter node $j \in \{1, 2\}$ at time $\{t_{l,q}\}_{1 \leq l \leq m}$. It follows that $d_{l,j}$ individuals had survival time less than $t_{l,q}$, and $Y_{l,j}$ individuals had a greater survival time. For a split using feature $x^v$ and its splitting values $\zeta^v_\gamma$, the goodness-of-split is measured using a log-rank statistic [45] represented as:

$$L(x^v, \zeta^v_\gamma) = \frac{\sum_{l=1}^{m} \left( d_{l,1} - Y_{l,1} \frac{d_{l,q}}{Y_{l,q}} \right)}{\sqrt{\sum_{l=1}^{m} \frac{Y_{l,1}}{Y_{l,q}} \left( 1 - \frac{Y_{l,1}}{Y_{l,q}} \right) \left( \frac{Y_{l,q} - d_{l,q}}{Y_{l,q} - 1} \right) d_{l,q}}} \tag{2.1}$$

The log-rank statistics in (2.1) compares the survival difference between the two daughter nodes at each distinct event time, $\{t_{l,q}\}_{1 \leq l \leq m}$. A larger difference between the two nodes represents a greater homogeneity within the node, hence, the best split at a node $q$ is determined by the feature $x^*$ and its value at the cut point $\zeta^*$ such that $|L(x^*, \zeta^*)| \geq |L(x^v, \zeta^v_\gamma)| \; \forall \; x^v$ and $\zeta^v_\gamma$. Algorithm 1 presents the procedure to select $x^*$ and $\zeta^*$ for any given parent node with $\kappa$ distinct values of $\gamma$.

RSF inherits the robustness and desirable properties (increased accuracy, minimized bias, and variance) of random forests model to the survival analysis. Recent works using RSF for survival data have shown improved results as compared to the CPH models and are getting popular as a survival analysis tool [46, 14]. Additionally, RSF effectively imputes the missing data—a common problem in healthcare data sets. However, along with inheriting the merits of random forests, RSF also inherits random forest's curse of

**Algorithm 1** Growing an RSF

```
Initialize: $i \leftarrow 1, b \leftarrow 1, x^* \leftarrow 0, \zeta^* \leftarrow 0$
Select $B$, $s_0$, $\Phi_{train}$
while $b \leq B$ do
    Grow $\mathcal{T}_b$
    while unique deaths in $\mathcal{L}(\mathcal{T}_b) \geq s_0$ do
        Find $x^*$, $\zeta^*$
        Perform node split
    end while
end while
Calculate CHF($\mathcal{F}$) for $\Phi_{OOB}$
```

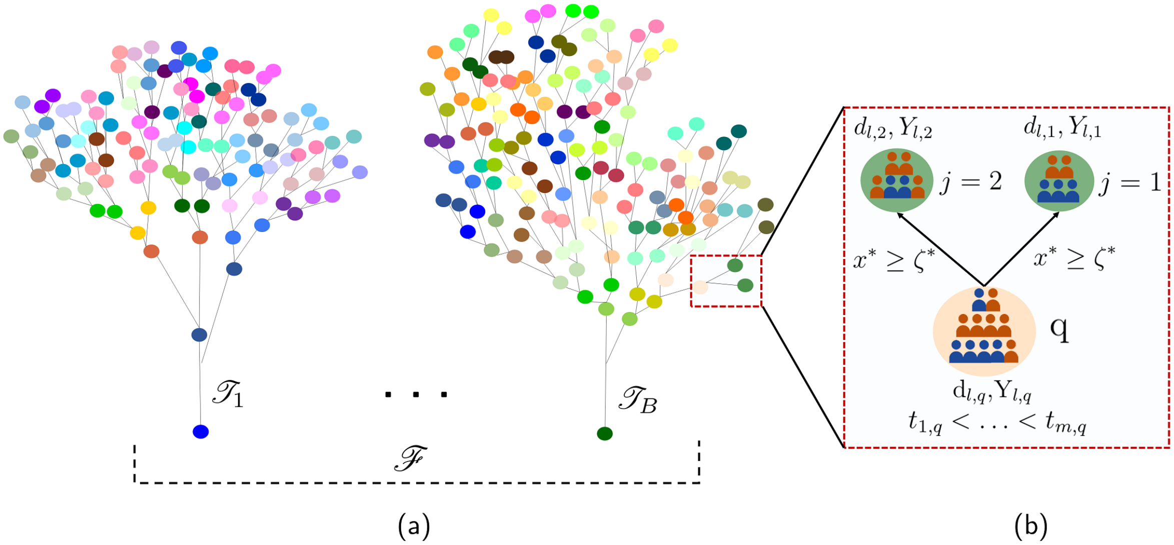
$\mathscr{T}_1$
. . .
$\mathscr{T}_B$
$\mathscr{F}$
$\mathrm{d}_{l,2}, \mathrm{Y}_{l,2}$
$\mathrm{d}_{l,1}, \mathrm{Y}_{l,1}$
$j = 2$
$j = 1$
$x^* \geq \zeta^*$
$x^* \geq \zeta^*$
q
$\mathrm{d}_{l,q}, \mathrm{Y}_{l,q}$
$t_{1,q} < \ldots < t_{m,q}$
(a)
(b)

Figure 2.1: A pictorial representation of (a) an RSF ($\mathcal{F}$) consisting of $B$ trees and (b) split of a parent node, $q$ into two daughter nodes using the feature $x^*$ at value $\zeta^*$.

learning from the imbalanced data which emphasizes on the minimizing the overall error rate and results in poor accuracy for the minority class [47]. The presence of extreme imbalance between the survival/censored and the mortality classes with as low as 2-10% data in the minority is a commonly occurring, yet often ignored aspect. Due to the contemporary clinical practice and infrastructure across the US, acute cardiac and other life-threatening, low-prevalence diseases are mostly treated in small tertiary care hospitals and, as a result, the cohort size tends to be small [40, 48], further exacerbating this challenge. Thus, data balancing is an essential step in maximizing the utility and improved mortality prediction performance. Despite this, only a few works focus on addressing class imbalance in the survival data [40] and none, to our knowledge, have done extensive theoretical and empirical studies on the effect of balancing in the survival analysis context.

Data balancing is an already developed area of research in the classification literature [49, 50] with several different balancing techniques being widely adopted to address the class imbalance problem. One popular balancing approach are data level sampling techniques that deals with modification of class distribution of the dataset before the learning algorithm is applied, such as, undersampling —randomly under sampling the majority class samples [51]; oversampling —randomly oversampling (with replacement) the minority class samples, until both classes have equal number of samples [26]. However, prior investigations suggest that over-sampling does not improve the minority class representation significantly and under-sampling is a better approach than over-sampling [52, 22]. Unfortunately, various real-life scenarios, including the present context where data is obtained from tertiary care hospitals, only limited samples are available. In such cases, under-sampling leads to an unwanted decrease in the training dataset and is not a feasible option.

This led us to explore a synthetic generation of minority class samples without resorting to excessive under-sampling. We adopt the synthetic minority over-sampling technique (SMOTE) proposed by [22]. This synthetic generation process proceeds in the feature space by selecting $k$ nearest neighbors of the minority class samples (we selected the default value of $k$ to be 5). Let $\boldsymbol{x}_i$ be the feature vector representing the features for the selected minority and $\boldsymbol{x}_j$ be the feature vector of a randomly chosen neighbor, then a new synthetic minority, $\boldsymbol{x}_s$ is generated in the feature (feature) space as follows:

$$\boldsymbol{x}_s = \boldsymbol{x}_i + \Gamma\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)$$

where, $\Gamma \sim \text{Uniform}(0,1)$ is a uniform random variable. Thus, the synthetically generated data can be interpreted as a randomly sampled point along the line segment between the minority samples and their nearest neighbors in the feature space. Based on the amount of oversampling needed, neighbors from the $k$ nearest neighbors are randomly chosen. For example, if the amount of over-sampling needed is 200%, then two of the $k$ nearest neighbors are chosen and one sample is generated in the direction of each (see [22] for further details). Representation of this scheme in two-dimensional feature space is shown in Figure 2.2 and the class ditributions before and after SMOTE's application is shown in Figure 2.3.

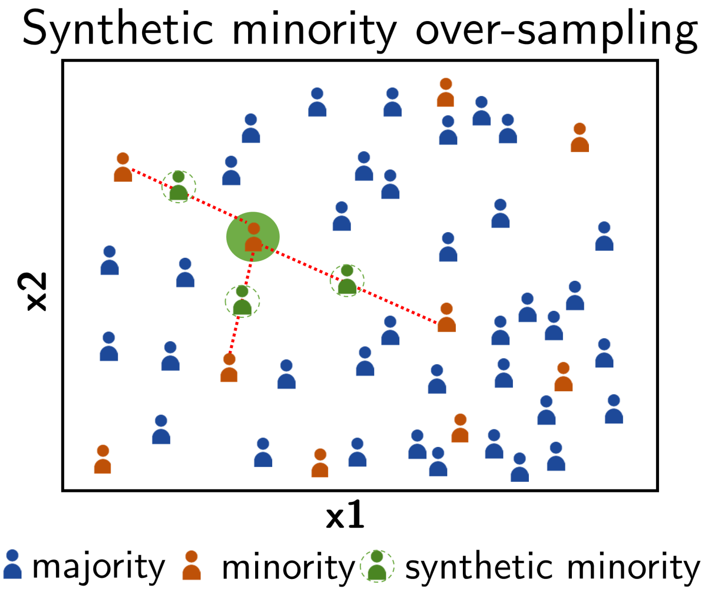

Figure 2.2: Representation of the imbalance in STEMI dataset and synthetically generated minority (in green) using SMOTE.

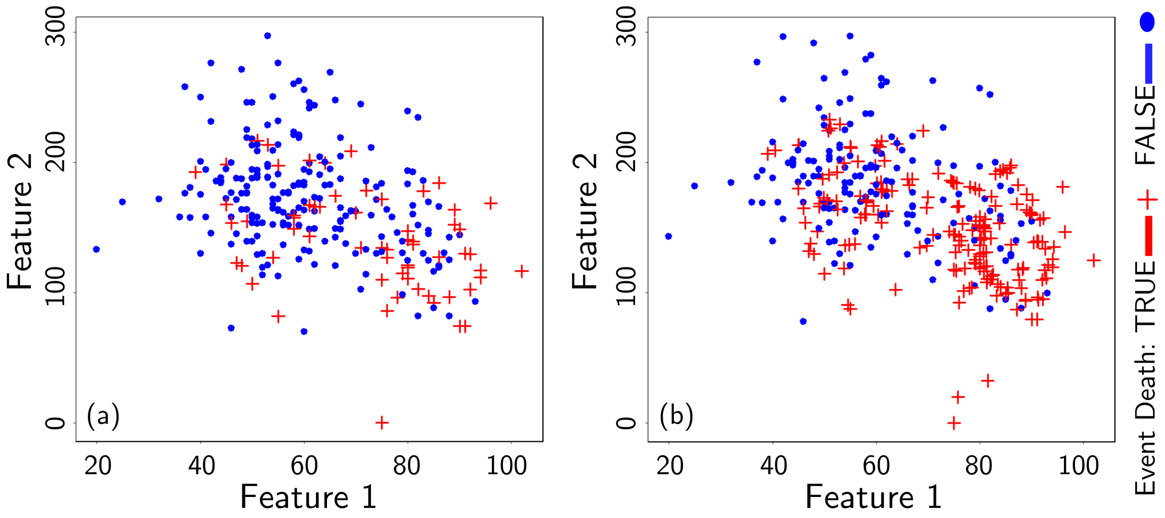

Figure 2.3: (a) Representation of the class imbalance in the feature space (minority class in red) and (b) representation of the balanced data using synthetically generated minority.

## 2.3 Results

In this section, we assess the performance RSF, BRSF, CPH, and balanced CPH through their application in mortality data sets with various level of imbalance. Along with the model accuracy results, a brief discussion of the data sets and details on the performance measures employed are also delineated.

### 2.3.1 Performance Measure

In the automated prognostics and decision support practice, where data drives the critical decision-making, the robustness of the model is of utmost importance. Recently, there have been vigorous debates on the effectiveness of the performance measures and on the efficacy of one measure over the other [42]. Here, we compare the performance of BRSF relative to contemporary survival models based on two of the most popular metrics in survival analysis literature—concordance index and Integrated Brier score. Further, we use 10 fold cross-validation (cv) scheme to calculate both of the measures to minimize bias for the test data, and to improve precision in the scenario of induced variance due to the data-driven steps in model building and validation measure.

#### *2.3.1.1 C-index*

Harrell's concordance index or C-index [53] is one of the most popular accuracy measure in the right-censored survival analysis literature. It assesses the model's discriminative strength by comparing the number of pairs of subjects where the model predicted a lower risk for the subject with higher survival time (concordant pairs), among all permissible pairs. In order to compute C-index, we first need to define permissible and concordant pairs. To account for the censoring, the set $\Theta$ of permissible pairs consists of all possible pairs of individuals, $i$ and $j$ in the data, but with two exceptions: 1) the ones in which shorter survival time is censored, and 2) when $T_i = T_j$, but neither of $i$ and $j$ has the event (death). Now, for any randomly selected pair out of the permissible cases, a pair can be concordant or partially concordant depending on their values of ensemble hazard, event time, and censoring status. For example, for a pair with distinct ensemble hazard and event times, a concordance value to 1 is assigned if the predicted risk (in terms of ensemble CHF) is greater for the individual that experiences death first i.e., $Pr\big(\sum_{l=1}^{n}\hat{H}_e(t_l^*|\boldsymbol{x_i}) > \sum_{l=1}^{n}\hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j > T_i\big)$. Here $t_1^*, ..., t_n^*$ denote all the unique event times in $\Phi$. For each pair in $\Theta$, the concordant pairs and their assigned concordance values can be given as:

$$\mathcal{I} = \begin{cases} 1, & \begin{cases} (\hat{H}_e(t_l^*|\boldsymbol{x_i}) > \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j > T_i) \\ (\hat{H}_e(t_l^*|\boldsymbol{x_i}) > \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j = T_i) \; \& \; (\delta_i = 1, \delta_j = 0) \\ (\hat{H}_e(t_l^*|\boldsymbol{x_i}) = \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j = T_i) \; \& \; (\delta_i = \delta_j = 1) \end{cases} \\ 0.5, & \begin{cases} (\hat{H}_e(t_l^*|\boldsymbol{x_i}) = \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j \neq T_i) \\ (\hat{H}_e(t_l^*|\boldsymbol{x_i}) \neq \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j = T_i) \; \& \; (\delta_i = \delta_j = 1) \\ (\hat{H}_e(t_l^*|\boldsymbol{x_i}) = \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j = T_i) \; \& \; (\delta_i = 1, \delta_j = 0) \\ (\hat{H}_e(t_l^*|\boldsymbol{x_i}) < \hat{H}_e(t_l^*|\boldsymbol{x_j})|T_j = T_i) \; \& \; (\delta_i = 1, \delta_j = 0) \end{cases} \\ 0, & \text{otherwise} \end{cases}$$

Then the C-index can be expressed as the ratio of the sum of concordance values and the total number of permissible pairs as:

$$\mathcal{C} = \frac{\sum_{i,j \in \Theta} \mathcal{I}}{|\Theta|}$$

Since $\mathcal{C}$ represents the classification accuracy of the model and is equiavlent to the area under the curve, a higher value is desirable. For concordance index valued from 0 to 1, a value of 0.5 is essentially no better than random guessing.

#### *2.3.1.2 Integrated Brier Score (IBS)*

We use prediction error curve (PEC) to capture a model's prediction of the survival probability for the test data at different time points. In the absence of censoring, PEC for an individual $i$ in the test data is an expectation of the squared difference between the true survival status and predicted survival probability of $i$ at time $t$ with features $\boldsymbol{x}_i$. However, censoring introduces bias in the population average of PEC. The introduction of inverse probability of censoring weight (IPCW) by [54] provides a versatile measure to overcome this limitation by weighting the squared residuals using IPCW. Given the survival data $\Phi = \{(\boldsymbol{x}_i, T_i, \delta_i)\}_{1\leq i\leq N}$, let the test dataset $D_M$ contain $M$ independent and identically distributed replicates of $\Phi$, where $M < N$. With the observed status for subject $i$, $\tilde{\mathcal{Y}}_i(t) = \mathbb{1}_{T_i>t}$ and its predicted survival probability $\hat{S}(t|\boldsymbol{x}_i)$, the prediction error or Brier score at time $t$ is given as:

$$\rho(t) = \frac{1}{M}\sum_{i\in D_M} \hat{W}_i(t)\left\{\tilde{\mathcal{Y}}_i(t) - \hat{S}(t|\boldsymbol{x}_i)\right\}^2 \tag{2.2}$$

In (2.2), the inverse probability of the censoring weights is estimated as [55]:

$$\hat{W}_i(t) = \frac{(1-\tilde{\mathcal{Y}}_i(t))\delta_i}{\hat{G}(T_i - |\boldsymbol{x_i})} + \frac{\tilde{\mathcal{Y}}_i(t)}{\hat{G}(t|\boldsymbol{x_i})}$$

where $\hat{G}(t|x) \approx P(\mathcal{C}_i > t|x_i = x)$ denotes the estimated conditional survival function of the censoring time. The aim here is to give the averaged prediction error at every time point in the test data. We also use survival probability plots of individuals in the test data at all event time points to show the predicted survival probability of the balanced and unbalanced models.

IBS consolidates the PEC estimates over all time points and is defined as:

$$IBS(\rho, \tau) = \frac{1}{\tau}\int_0^\tau \rho(\mu)d\mu$$

Where $\tau$ is the total time span for which the prediction errors can be estimated. Since Brier score is an error measure, a small value is desirable with 0 indicating perfect prediction.

### 2.3.2 Data Description

*2.3.2.1 Benchmark Data*

Six data sets were used the effect of data balancing in survival analysis. Five of the six data sets (except the STEMI dataset) used in this study were obtained from online repositories, each with a different level of imbalance. These 5 data sets consists of survival analysis data for acute diseases such as lung cancer (veteran and lung data sets), a plasma cell immune disorder which may result in malignancy (mgus dataset), acute stroke in patients with atrial fibrillation (COST dataset), a rare and fatal chronic liver disease (pbc dataset). A summary of the class proportions in all the data sets for the survival and the mortality classes is given in Table 2.1.

Table 2.1: Summary of the survival data sets used for model evaluation.

| | **Class proportions** | | |
|---|---|---|---|
| **Dataset** | Total | Censored | Mortality |
| veteran [1] | 137 | 9 | 128 |
| mgus [56] | 241 | 16 | 225 |
| COST [57] | 518 | 114 | 404 |
| STEMI [58] | 267 | 205 | 62 |
| lung [59] | 228 | 63 | 165 |
| pbc [60] | 418 | 257 | 161 |

[*] Minority class size is represented in red.

The next subsection provides the description for the STEMI dataset that was obtained from our collaborators at the Heart, Artery, and Vein center of Fresno.

*2.3.2.2 ST Elevated Myocardial Infarction Data*

The study cohort for the STEMI dataset consisted of 278 consecutive patients. The patients had electrocardiographic criteria for STEMI and a presumed diagnosis of acute coronary syndrome at the time of presentation to the emergency room of a tertiary care hospital in central California, USA. Electrocardiographic, radiographic, and basic laboratory investigations were obtained at the time of presentation and an emergent coronary angiography was performed. Patients underwent coronary artery bypass grafting (CABG) or primary percutaneous coronary intervention. Enrollment into the study began in January 2007 and patient were followed for one year until January 2008. A detailed design of this retrospective study has previously been published [58]. We focused primarily on $N = 267$ patients (187 male and 80 female) who did not have preexisting left bundle branch block or paced rhythm on ECG. Dataset consisted of a large set ($R = 150$) of features. These features included therapy provided, physiological and anatomical variables such as age, gender, ethnicity, BMI, ECG criteria, the occurrence of cardiac arrest during admission, troponin levels at the time of discharge, Brain Natriuretic Peptide (BNP) levels, and clinical risk measures such as TIMI index, Mayo Clinic risk score etc. along with the previously mentioned laboratory measurements. The dataset had ethnically diverse population including Black, Caucasian, and a high percentage of representative minority populations such as American Indian, Asian, and Hispanics. Mortality data were obtained either from the hospital, California Department of Public Health (CDPH) or Social Security Death Index records. To avoid any confounding effects of loss to follow-up and accurate determination of the cause of the death, an all-cause mortality was selected as a primary endpoint. Out of the 267 patients, 62 patients died in one-year duration (representing the minority class for this dataset).

### 2.3.3 Performance Comparison

Most of the data sets contained several missing values which were then imputed using adaptive tree imputation [15] and the data was balanced for equal class representation. To compare CPH and RSF and to determine the effect of balancing on these models, we use the C-index and IBS measures described in subsection 2.3.1. Table 2.2 presents the average C-index and IBS scores for CPH, balanced CPH (BCPH), RSF, and BRSF obtained via a 10 fold cv scheme. The best model obtained for both C-index and IBS are shown in blue. As evident from this Table, BCPH and BRSF consistently performed better than their unbalanced counterparts.

Table 2.2: Performance evaluation results for the benchmark data sets.

| | | **Model** | | | |
|---|---|---|---|---|---|
| **Dataset** | **Error measure** | CPH | BCPH | RSF | BRSF |
| **veteran** | C-index | 59 (0.24) | 58 (0.21) | 61 (0.11) | 77 (0.04) |
| | IBS | 0.15 (0.04) | 0.14 (0.03) | 0.15 (0.05) | 0.09 (0.02) |
| **mgus** | C-index | 71 (0.05) | 89 (0.021) | 69 (0.07) | 88 (0.02) |
| | IBS | 0.13 (0.02) | 0.06 (0.01) | 0.14 (0.02) | 0.04 (0.01) |
| **COST** | C-index | 69 (0.03) | 76 (0.03) | 64 (0.04) | 85 (0.01) |
| | IBS | 0.17 (0.01) | 0.15 (0.01) | 0.18 (0.02) | 0.06 (0.01) |
| **STEMI** | C-index | 80 (0.12) | 79 (0.05) | 82 (0.08) | 82 (0.06) |
| | IBS | 0.18 (0.07) | 0.12 (0.04) | 0.17 (0.06) | 0.08 (0.01) |
| **lung** | C-index | 61 (0.09) | 70 (0.04) | 59 (0.09) | 76 (0.03) |
| | IBS | 0.18 (0.01) | 0.13 (0.01) | 0.18 (0.02) | 0.08 (0.01) |
| **pbc** | C-index | 77 (0.09) | 79 (0.02) | 78 (0.08) | 83 (0.02) |
| | IBS | 0.14 (0.02) | 0.12 (0.01) | 0.13 (0.02) | 0.07 (0.01) |

[*] Numbers inside the bracket represents standard deviation across the 10 fold cv.

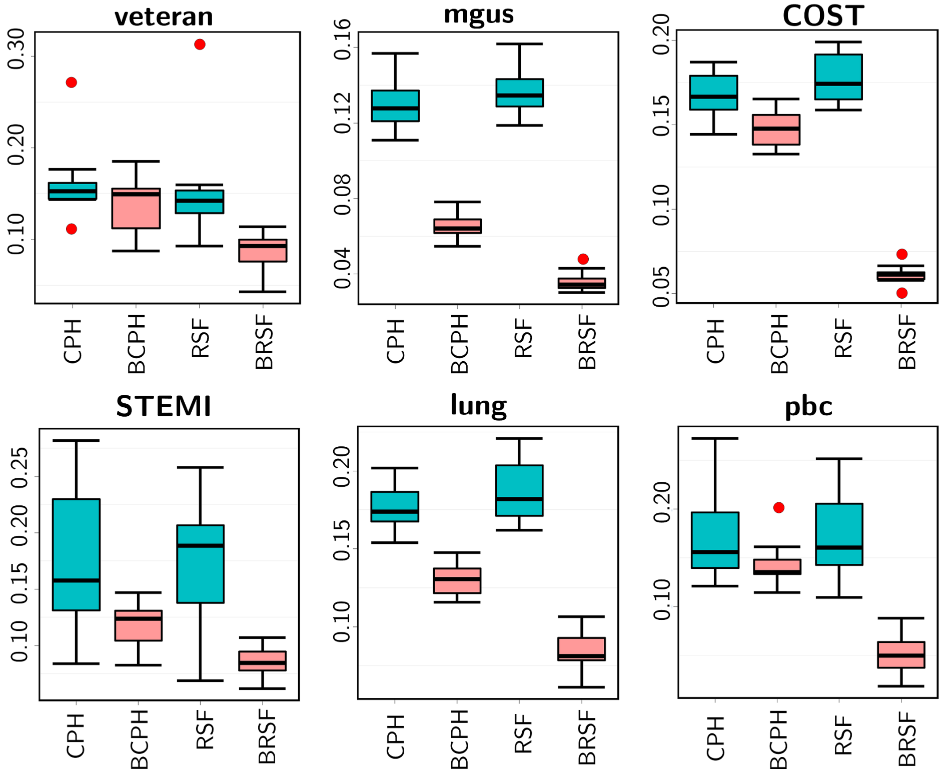

Figure 2.4: Boxplots of estimated IBS calculated for the test data in 10-fold cv scheme for 6 data sets arranged in decreasing order of class imbalance. The horizontal line inside the box represents the median and the box is bounded by the $25^{th}$ and $75^{th}$ percentile (IQR). Whiskers extend to $1.5 \times$ IQR and the outliers are represented by the red dot.

Additionally, the performance of BRSF supersedes all other models. There was an overall improvement of 25% in the C-index and 55% in the IBS score from RSF to BRSF. This is summarized in Figure 2.4.

The performance improvement in C-index can be better represented in terms of the survival probability curves. Figure 2.5, presents the survival probability plots for the veteran, mgus, COST, and lung data sets for which survival class is the minority and STEMI and pbc data sets which has mortality class as minority (refer to Table 2.1). In this figure, the red curve represents survival probability of the mortality samples, blue curve represents the survival of the censored samples at different event times, and the dashed lines are used to represent the synthetic samples. Ideally, the survival probability of the mortality should be low and that for the censored samples should be high. However, due to imbalance, the hazard/survival probability estimates are underestimated when $m_2 << m_1$ and overestimated when $m_1 << m_2$. After the classes are balanced not only their separability (i.e. higher survival probability for the censored samples and lower for the mortality samples) increases but also the survival/hazard probability estimates for the minority samples improves.

Since the data size increases after balancing, to ensure that the performance improvement after balancing was not due to the difference in the number of samples in the leaf node (terminal nodesize) of BRSF and RSF, we compared the results for different nodesize sequence. Figure 2.6 shows the average prediction performance in terms of C-index (Figure 2.6 (a)) and IBS scores (Figure 2.6 (b)) for the 5 benchmark data sets across 7 different nodesizes. This figure shows a consistently better performance of BRSF at all nodesizes with the minimum average C-index of BRSF (0.81) being higher than the maximum average C-index accuracy of RSF (0.69) and the maximum IBS averaged error of the BRSF (0.09) being lower than the minimum average IBS of RSF (0.15).

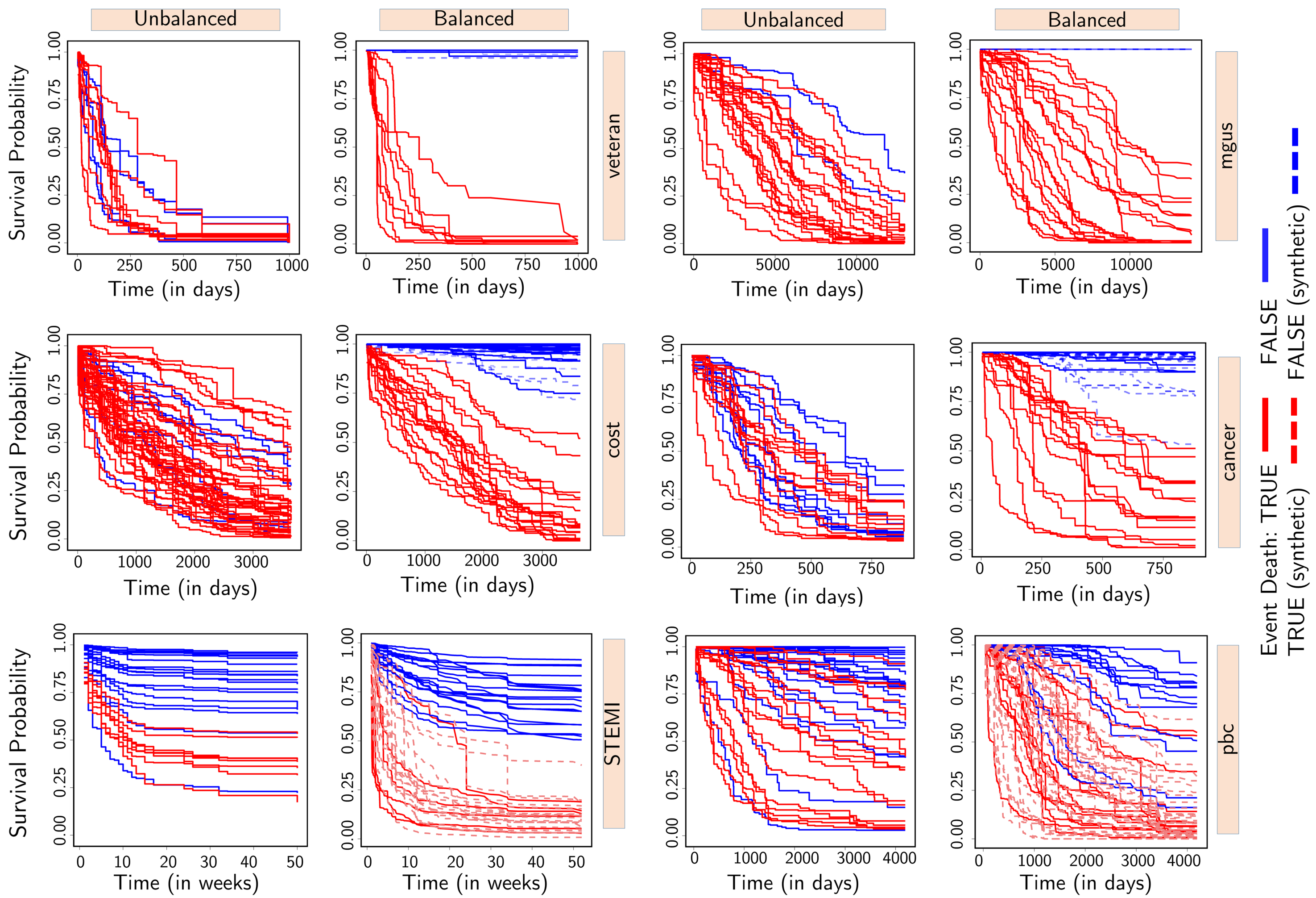

Figure 2.5: Survival probability plots for veteran, mgus, COST, lung, STEMI, and pbc data sets. For the pair of plots for each dataset, the left plot represents the survival probability for original imbalanced data and the right one represent the survival probability after balancing. Survival probability plots for the synthetic samples are shown in dashed lines.

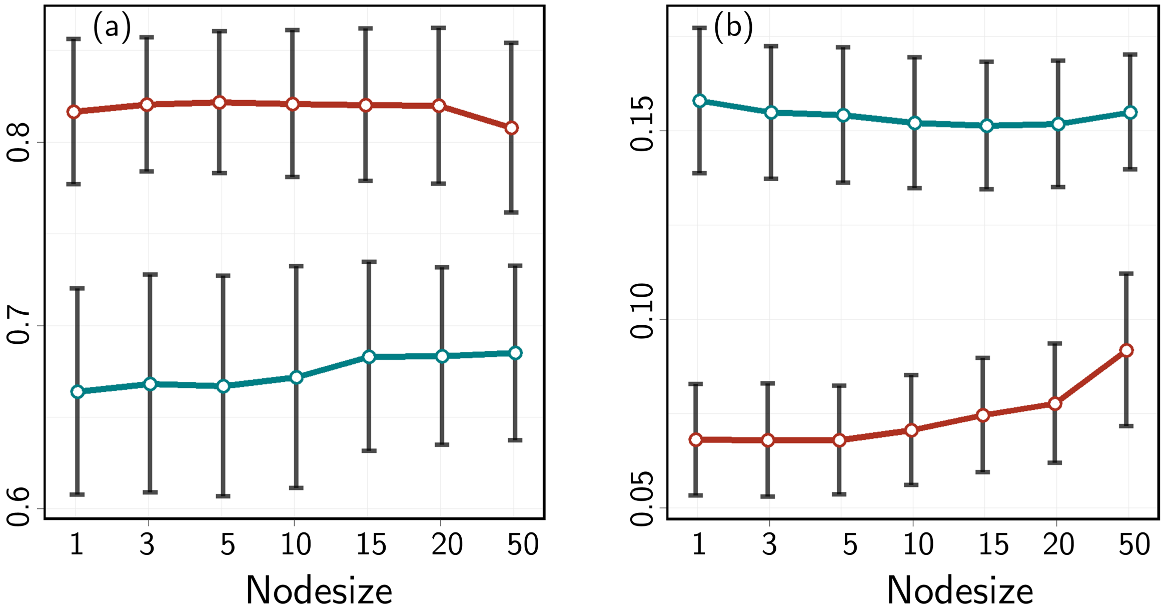

Figure 2.6: Average prediction performance of RSF (green) and BRSF (red) for the 5 benchmark data sets at 7 different node sizes in terms of (a) C-index and (b) IBS score.

Further, we obtained 7 best features for STEMI data based on the backward selection. Their OOB error was 15.8% compared to 18% with all 150 features. It turns out these features have high importance as per both Breiman's variable importance (VIMP) [61] and Ishwaran et. al.'s minimal depth (MD) scores [62] (see Table 2.3). From a physiological standpoint, these features are among the most significant indicators of survival during acute cardiac diseases, as elaborated in the following paragraphs.

Table 2.3: The top 7 features for the STEMI dataset.

| | | | | **features** | | | |
|---|---|---|---|---|---|---|---|
| **Ranks** | Disch | MCRS | Cron | GRACE | MCRS | CHF | ACS |
| **(statistics)** | Trop | MS | DC | Prob | MACE | in1yr | in1yr |
| MD | 1(8.16) | 2(8.39) | 3(8.53) | 4(8.75) | 7(9.28) | 8(9.30) | 12(9.69) |
| VIMP | 9(0.01) | 3(0.02) | 1(0.02) | 8(0.01) | 7(0.01) | 4(0.02) | 5(0.02) |

We graphically explored the relation of thus selected most important features (top 7) with the survival probability using partial dependence plots. In a survival setting, a partial dependence plot represents the response corresponding to the feature of interest at a particular time by averaging out the joint effect of the remaining features [63, 64]. In Figure 2.7, the two curves corresponding to each of the features shows the trend of survival probability with the changing value of the feature at $16^{th}$ and $32^{nd}$ week for 100 randomly chosen subjects. It shows nonlinearly-decreasing survival probability with increasing value of “DischTrop”, “MCRS-Mortality Score(MS)”, “CronDC”, “GRACEProb”, “MCRS-MACE”, “CHFin1yr”, “ACSin1yr”. For all the features, we can see the decreasing survival probability with increasing time (blue line for $32^{nd}$ week is below the green line for $16^{th}$ week). Most importantly, the non-proportional hazard trend is evident from the “WBC” or white blood cell count feature. This trend will be disregarded by the CPH model which has a proportional hazard assumption. The variables selected were evaluated by a cardiologist to have a significant physical meaning and correlation with the prediction of mortality.

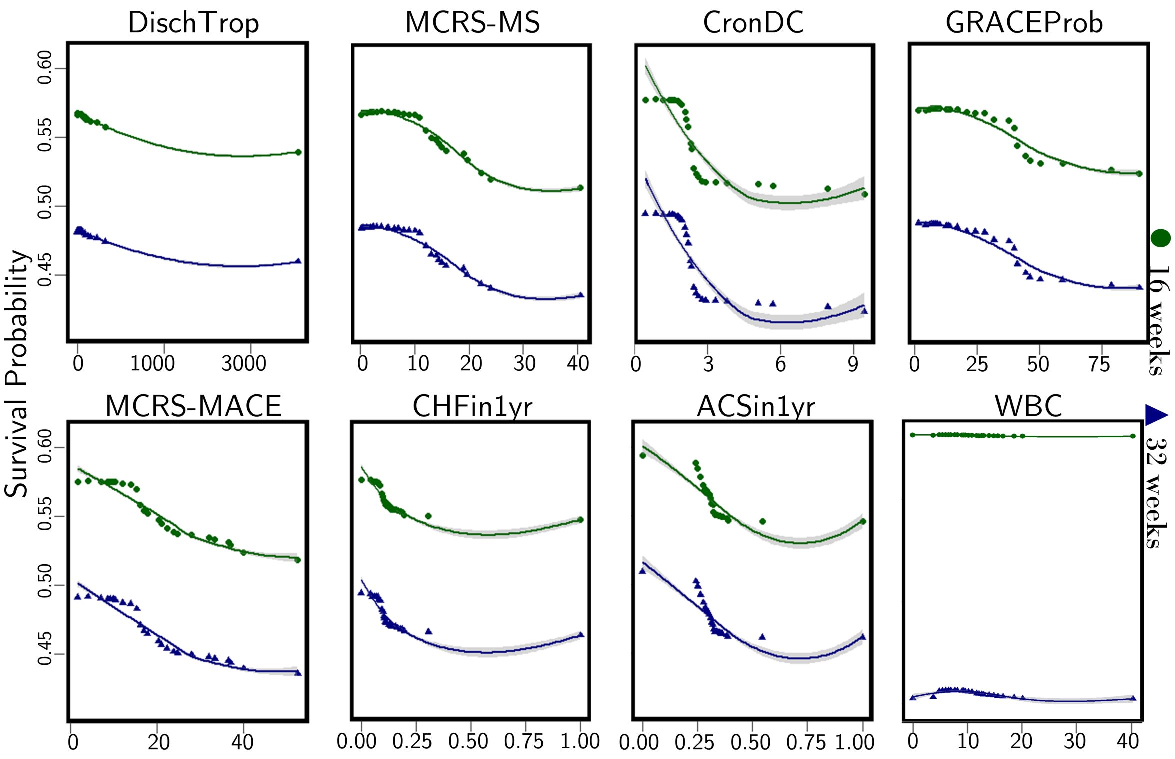

Figure 2.7: Partial dependence plot of predicted survival probability plotted as a function of top 6 features for randomly chosen 100 subjects (market as "circle" and "triangle"). The green and blue lines show the trend at $16^{th}$ and $32^{nd}$ weeks respectively.

The variables selected were evaluated by a cardiologist to have a significant physical correlation with the prediction of mortality. In particular, the DischTrop (discharge troponin) feature, recording the troponin levels during the patient discharge is studied as a primary diagnostic component [65, 66]. Troponin is a protein released during myocardial infarction. A higher level of troponin indicates more damage to the cardiac muscle. Figure 2.8 is a visualization of survival probability trend with varying DischTrop level across the mortality and survival samples. Patients with a higher level of troponin content during the discharge are shown to have lower survival probabilities. Though application of machine learning algorithms, in particular, ensemble based approaches are often criticized for their lack of interpretability in real-world data, the variable and partial dependence plots for all the features can provide insightful information on their relationship with mortality. Consequently, the proposed technique can be used by healthcare practitioners as an analytical tool to achieve improved throughput and accuracy.

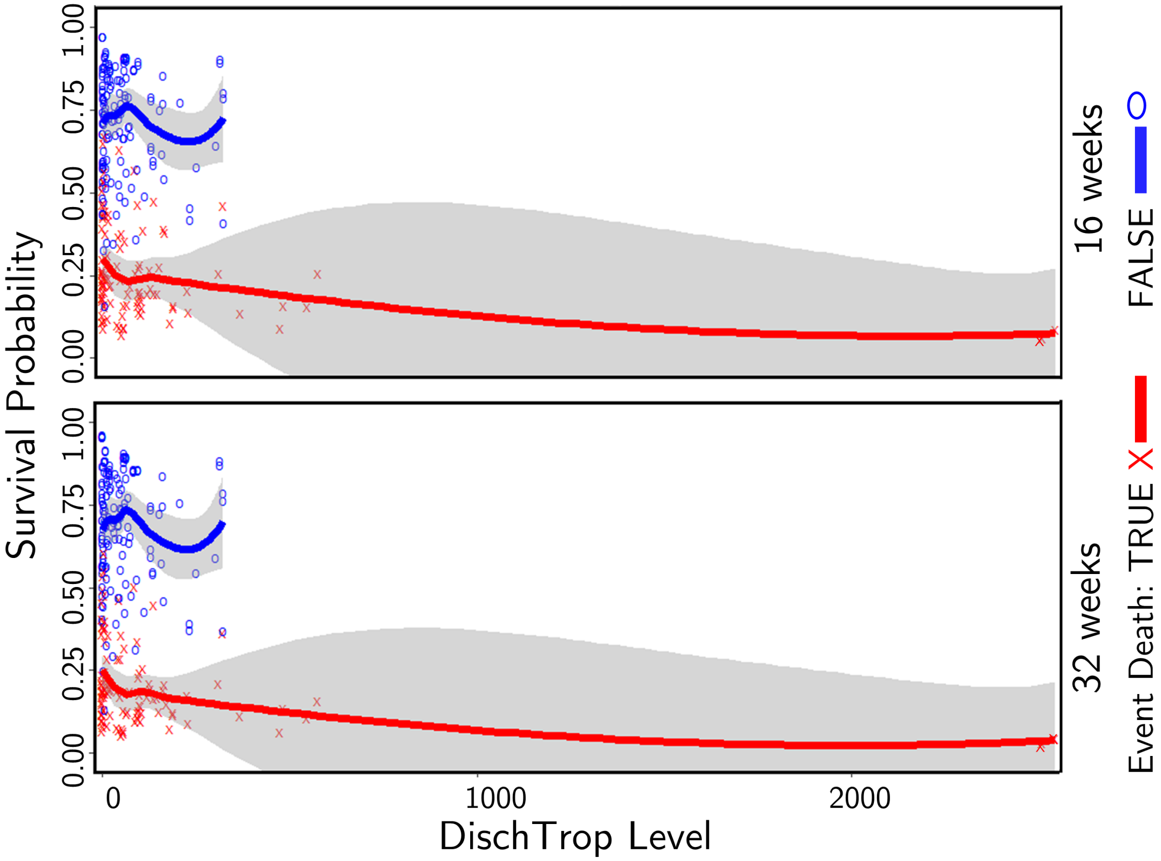

Figure 2.8: Variable dependence plot of survival probability plotted as a function of DischTrop level at $16^{th}$ and $32^{nd}$ weeks.

## 2.4 Summary

In this chapter, we introduced a BRSF model for survival analysis to address the limitations in handling extremely imbalanced data sets in survival analysis. The theoretical results, as well as extensive experimental analysis presented in this chapter, demonstrates the benefits of data balancing in survival learning. Empirical studies with

6 data sets suggest a 55% improvement in IBS score of BRSF as compared to RSF. Although class imbalance has been extensively studied in the machine learning literature, its theoretical analysis and application in the domain of survival analysis still remain largely unexplored. Pertinently, this is among the first extensive investigations into the effect of class imbalance on the performance of survival models. Specifically, the theoretical results on performance improvement accrued from balancing the RSF models as well as the detailed empirical studies can lead to further improvements to the algorithms for RSF as well as more optimized balancing strategies. Future work on extensive comparison of different data balancing techniques could be essential in shedding more light on data imbalance in the survival analysis context and develop more customized balancing techniques.

# 3. NON-PARAMETRIC SURVIVAL LEARNING FOR HIGH DIMENSIONAL MULTI-VIEW DATA

## 3.1 Introduction

Simultaneous advancements in a variety of sensing and data acquisition techniques have created a data explosion that has not only impacted the engineering or IT domain through internet traffic and social network data but also created healthcare big data. Within a few decades, the challenge of learning with limited features has now metamorphosed into the challenge of learning from high dimensional big data with hundreds and thousands of features. More recently, the innovations in high-throughput technologies and the creation and collaborations of big consortium such as The Genome Technology Program at the National Human Genome Research Institute [67], International Cancer Genome Consortium [68], and TCGA program [29] have facilitated low-cost and high-quality genomic data collection from multiple sources [33]. The data (often heterogeneous) collected from multiple sources provide complementary information or different views of the same process and are referred to as multi-view data [28]. For example, in multimedia content understanding, a video and audio signal can provide complementary information for a multimedia segment. It can be simultaneously used to describe it [34], and in molecular biology, different omics data, i.e., genomics and transcriptomics, can provide complementary information for phenotype determination.

Since the integrated analysis of the multi-view data has been shown to provide complementary information and elucidate the underlying complex disease mechanisms [69], multi-view data analysis has emerged as a promising area of research. However, most of the key integrated multi-view analysis methods are exploratory and unsupervised. Studies in survival learning using multi-view data have been very limited [70] despite one of the key applications of multi-view dataset in healthcare being cancer prognosis and progression studies using survival analysis.

In line with our goal of improving survival prognosis accuracy, in this chapter, we focus on addressing the challenges imposed by the multiple multivariate or multi-view dataset for survival learning. Specifically, we will be focusing on vertical multi-view data or multi-view data with the same set of samples/patients (see Figure 1.2) for survival prognosis. This chapter has two specific aims: (1) to turn the challenge of using high dimensional multi-view data into an opportunity of learning from various significant biological processes by utilizing an efficient data integration method and (2) to implement a non-parametric survival learning model on the integrated data for prognosis performance improvement.

This chapter is organized as follows. In the first section, we review the multi-view data integration literature and give a brief description of the Joint and Individual Variation Explained (JIVE) method that is used in this dissertation for multi-view data integration. In the second section, we propose a non-parametric survival learning for the integrated multi-view data. Finally, the last section provides the assessment and performance comparison results for data integration and survival learning for breast invasive carcinoma data.

## 3.2 Multi-view Data Integration

The integration of different data types for understanding the same set of subjects has several advantages. Some of them are as follows: (1) novel biological insights: different data types in the multi-view data can provide insights that are often not available from a single data type. For example, in genetics, expression of biological activity such as a tumor is suspected to depend on the interaction among the gene, protein, and transcriptional regulators, thus their integration provides a different vantage point to observe the same biological phenomena and gather more details than ever possible with single data type [71, 72, 69], (2) increased accuracy in noisy data: in many applications including high-throughput technologies in genomics, the signal-to-noise ratio is often low. In such conditions, obtaining complementary information from multiple-data types can significantly increase the accuracy of the analysis, and (3) robustness in rare disease prognosis: the current diagnostic accuracies for a variety of rare diseases is just 25-50%. The integration of different data types can not only increase this accuracy but can provide additional evidence for a molecular event, establishing a causality chain that can not be established using only a single data type [69, 73].

Hence several efforts have been made for integrated multi-view analysis, and it is a new and growing area of research [70]. However, multi-view data integration is challenging due to the redundancy, HDLSS issue, and complex associations between the constituent data types [28]. Additional challenges arise when multi-view data integration is followed by downstream analysis, such as classification and survival prognosis. Hence a robust and effective data integration is a crucial first step for accurate survival prognosis.

Existing multi-view data integration methods in the literature can primarily be classified into two extremes—early or late integration. In the early integration approach, all data types are concatenated and then used for subsequent analysis. In the late integration, each data type is analyzed separately, followed by an ad-hoc combination of their results [74, 75, 28]. One advantage of the early and late integration method is that they allow for a straightforward application of single multivariate data analysis methods, for example, the application of dimensionality reduction methods such as the principal component analysis (PCA) and singular value decomposition (SVD) on the concatenated data in early integration and the individual data type in late integration [75, 76]

However, both these approaches have significant limitations [77]. The early integration of high dimensional data can have scaling and interpretability issues. It may fail to capture information specific to a single data type and their relation with the response. The late integration, on the other hand, neglects the association and interactions between the data types by analyzing them individually and leads to decreased statistical power.

Hence, effective multi-view data integration methods are required to address these challenges. Recently, several methods have been proposed to address these issues for optimal integration and joint analysis of multi-view data. These include data transformation methods such as matrix factorization, network-based learning, etc. that transform the data types in a common space. Several methods focused on exploring the associations between the data types have also been proposed. Canonical correlation analysis (CCA) [78] and co-training [79] are one of the earliest methods for investigating the association between two sets of variables. Extensions of CCA for high-dimensional [80] and multiple data types (more than two data types) have been proposed [32] for use in multi-view data analysis. Similarly, methods such as O2-PLS, an extension of partial least square (PLS), is used to explore the association between a pair of data types [81] and multi-level functional PCA or MF-PCA is used to explore the between and within variation of grouped samples on the same functional data [77, 82].

Since our emphasis is on exploring the impact of each data type and their association on the response, a method that can extract the common and disparate structures in each data type is of interest. One such method is joint and individual variation explained (JIVE) [31]. JIVE is an extension of PCA that decomposes the multi-view data into a low-rank approximation of joint and individual components capturing the joint (or shared) variation and individual variations in each data type. JIVE was selected as it provides both a robust and insightful understanding of the integrated data structure, including the contribution of each data type [83, 84]. In the next section, we provide a brief summary of the JIVE integration method (details of the JIVE method are presented in [31]).

### 3.2.1 Joint and Individual Variation Explained (JIVE)

It is well established that the data types obtained from multiple sources for the same set of subjects are associated with each other [72]. For example, in molecular biology, the central dogma explains the flow of information from DNA to messenger RNA (mRNA) and mRNA to protein through the translation and transcription process, making these molecular level associated with each other when measured for the same subject [33, 85]. Thus, a multi-view data integration method that can exploit and explore these associations can provide additional information for improved prognosis and elucidate the effect of each data type on the response. JIVE is one such exploratory data integration method that separates and analyzes the joint and individual (information uniquely present in a data type) effects in multi-view data [31].

As discussed in the foregoing, JIVE is based on decomposing each data type into a sum of three terms, namely, a low-rank approximation of joint structure capturing the shared variation across the data types, a low-rank approximation of the individual structure capturing the variations specific to each data type, and the residual noise [31]. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k$ with $k \geq 2$ be vertical data types where each data type has the same number of $N$ columns regarded as subjects and $p_i$ features or rows that may or may not be the same. Hence, their vertical integration, $\boldsymbol{X}$ has $N$ columns and $p = p_1 + p_2 + \cdots + p_k$ rows (see Fig. 1.2). Now, let $\boldsymbol{J}_i$ be the sub-matrix of the joint structure matrix and $\boldsymbol{I}_i$ be the individual structure matrix associated with $X_i$, then the unified JIVE model is given as:

$$\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \\ \vdots \\ \boldsymbol{J}_k \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_1 \\ \boldsymbol{I}_2 \\ \vdots \\ \boldsymbol{I}_k \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_k \end{bmatrix} \tag{3.1}$$

where $\boldsymbol{\epsilon}_i$ are the error matrix of independent entries with zero expectation and has

$N$ columns and $p_i$ rows, corresponding to $\boldsymbol{X}_i$ and the joint structure matrix, $\boldsymbol{J}$ can be given as the vertical concatenation of the joint sub-matrices as:

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \\ \vdots \\ \boldsymbol{J}_k \end{bmatrix} \tag{3.2}$$

The JIVE model imposes rank constraints on the joint and individual matrices where $rank(J) = r$ and $\text{rank}(I_i) = r_i$ such that $r < rank(X)$ and $r_i < rank(X_i)$ for $i = 1, \ldots, k$. The rank selection is done using a permutation testing approach and is a critical first step in the subsequent estimation of the joint and individual structures. The choice of ranks is crucial to avoid over and underestimation of joint and individual variations. Now, for fixed ranks, $r, r_1, \ldots, r_k$, $J$ and $I_i$ are estimated by minimizing the sum of squared error. Let $R$ be the $p \times N$ residual matrix after accounting for the joint and individual structures, given as:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \\ \vdots \\ \boldsymbol{R}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 - \boldsymbol{J}_1 - \boldsymbol{I}_1 \\ \boldsymbol{X}_2 - \boldsymbol{J}_2 - \boldsymbol{I}_2 \\ \vdots \\ \boldsymbol{X}_k - \boldsymbol{J}_k - \boldsymbol{I}_k \end{bmatrix} \tag{3.3}$$

Then the JIVE model iteratively estimated $J$ and $\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_k$ by minimizing $||\boldsymbol{R}||^2$. This iterative estimation is summarized in algorithm 2 (more details on the estimation of joint and individual structures and rank selection can be found in the supplementary section of [31]).

This estimation imposes orthogonality constraint between the rows of the joint and individual matrix, i.e., $JI_i^T = 0_{p \times p_i}$ for $i = 1, \ldots, k$ ensuring that the estimated joint structure between the data types and the individual structures are unrelated and are

**Algorithm 2** Estimation of joint and individual structures using JIVE

```
Initialize: $\boldsymbol{X}^{\text{Joint}} = \boldsymbol{J} = [\boldsymbol{J}_1', \ldots, \boldsymbol{J}_k']'$ and $RF = \Delta$
while $\boldsymbol{R} > RF$ do
    For a given rank $r$, estimate $\boldsymbol{J} = [\boldsymbol{J}_1', \ldots, \boldsymbol{J}_k']'$ by a rank $r$ SVD of $\boldsymbol{X}^{\text{Joint}}$
    while $i \leq k$ do
        Set: $\boldsymbol{X}_i^{\text{Individual}} = \boldsymbol{X}_i - \boldsymbol{J}_i$
        Estimate $\boldsymbol{I}_i$ by a rank $r_i$ SVD approximation of $\boldsymbol{X}_i^{\text{Individual}}$
        Set: $\boldsymbol{X}_i^{\text{Joint}} = \boldsymbol{X}_i - \boldsymbol{I}_i$
        Estimate $||\boldsymbol{R}||$ and
    end while
    Set $\boldsymbol{X}^{\text{Joint}} = [\boldsymbol{X}_i^{\text{Joint}'}, \ldots, \boldsymbol{X}_i^{\text{Joint}'}]'$
end while
```

uniquely determined.

Through the JIVE decomposition, the collinearity and redundancy between the data types are accounted for in the joint component. It also facilitates identifying the data types that contain useful information not present in others [36, 77]. However, JIVE is an exploratory unsupervised method. In the next section, we extend JIVE for survival prediction.

## 3.3 Multi-view Survival Learning

Multi-view data integration has been primarily motivated by real-life applications aimed at improving prognosis and precision medicine for complex and severe diseases such as cancer using multiple data sets that can provide different vantage points to understand the disease prognosis. However, most of the early and prominent methods proposed (as discussed in section 3.2) in the multi-view data analysis literature, including JIVE, are unsupervised methods aimed for exploratory analysis of high-throughput data, dimensionality reduction, and easy visualization. Nonetheless, the importance of incorporating phenotype labels (disease or normal) in biostatistics, i.e., an extension of multi-view method to supervised and semi-supervised learning is paramount as it enables disease diagnosis. Recognizing this significance, several methods are now being proposed for supervised and semi-supervised multi-view data analysis [86, 87, 88].

However, the supervised and semi-supervised approaches are used to classify patients into two groups but are unable to link the features to the time-to-event and fail to provide a continuous risk score for the patients [89]. Nonetheless, the clinical data analysis is now moving from supervised diagnosis to early prediction and long-term prognosis using survival analysis methods with emphasis on prognostic feature determination that can inform decision-making, such as personalized treatment, patient management, and risk stratification. However, a review of the literature until 2015 found no methods for predictive analysis from multi-view data [90, 36]. Currently, only a few methods exist that integrate multi-view data for survival analysis [70]. Further, the adaptation of the unsupervised methods for survival data is not trivial due to the censoring of data making it an open area of study in the multi-view literature.

The recent development in multi-view survival analysis has two fundamental limitations that we are trying to address in this chapter. Firstly, the data types are integrated either using the early or late integration method resulting in several limitations, as explained in the foregoing. Secondly, the survival analysis models used are primarily dominated by the parametric or semi-parametric models like CPH model which inflicts several restrictions (see section 1) that does not hold in real-life, especially when the data is as complicated as the multi-view genomics data for cancer prognosis [36, 37]. To resolve these issues, we propose an integrated non-parametric survival learning method or *iNPS* learning that use JIVE method for effective integration and exploration of multiple data types followed by the implementation of the non-parametric RSF method, for survival learning.

As stated previously, JIVE is an extension of PCA to multi-view data. Similar to the factorization of a single multivariate data matrix using PCA, the JIVE factorization of each data type is given as follows:

$$\boldsymbol{X}_i = \boldsymbol{U}_i\boldsymbol{S} + \boldsymbol{W}_i\boldsymbol{S}_i + \boldsymbol{\epsilon}_i, i = 1, \ldots, k \tag{3.4}$$

where $\boldsymbol{S}$ is $r \times N$ common joint score matrix that summarizes the joint structure between the data types and $\boldsymbol{U}_i$ is $p_i \times r$ are loadings that provide how the joint scores are expressed in features of data type $i$. Similarly, $\boldsymbol{S}_i$ is $r_i \times N$ scores summarizing the individual variation with the data type $i$ and $\boldsymbol{U}_i$ is $p_i \times r_i$ corresponding loadings.

Now, these scores can be used for subsequent analysis of time-to-event. The CPH model to predict the time-to-event for a subject $i$ using two data types ($k = 2$) with joint and individual scores of ranks $r, r_1$, and $r_2$ can be given as:

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t) \exp(\sum_{j=1}^{r} \beta_j J_{ji} + \sum_{l=1}^{r_2} \beta_l I_{1li} + \sum_{m=1}^{r_2} \beta_m I_{2mi}) \tag{3.5}$$

Leveraging the strength of JIVE, using CPH for survival prognosis using the joint and individual scores is promising due to the dimension reduction and removing the redundancy between multiple data types [36]. However, the limitations of the CPH model, such as the proportional hazard assumption (see section 1.1) still undermines the accuracy of the estimated survival prognosis. Hence, we use the RSF learning method with the scores of the joint and individual matrices are treated as composite biomarkers [36]. The subsequent analysis and performance estimation of the JIVE and *iNPS* method is demonstrated using the real-world dataset for cancer prognosis described in the following section.

## 3.4 Results

### 3.4.1 Data Description

#### *3.4.1.1 Breast Invasive Carcinoma Data*

To assess the performance of the parametric and non-parametric survival analysis methods on multi-view data, we used the breast invasive carcinoma (BRCA) dataset from TCGA (https://www.cancer.gov/tcga) [29]. BRCA is one of the benchmark data sets predominately used in the studies for multi-view data analysis [91]. This data contained 620 patients with three omics data types, including gene expression (illuminahiseq_rnaseqv2-RSEM_genes_ normalized), miRNA (illuminahiseq_mirnaseq-miR_gene_expression), and methylation (illumina humanmethylation 450k). The expression data sets were downloaded from Broad institute's Genome Data Analysis Center (GDCA) Firehose (https://gdac.broadinstitute.org) and the clinical and phenotype data with the overall survival status and time to event was downloaded from UCSC XENA (https://xena.ucsc.edu) [92].

The dataset was preprocessed to remove all patients and features with more than 20% missing values, and the remaining missing values were imputed using k-nearest neighbor imputation. For the methylation, only 5000 features with maximal variance were selected [91]. Following the preprocessing, each sample had $p_1 = 20531$ genes, $p_1 = 1046$ miRNAs, and $p_1 = 5000$ methylation. We also chose to include 10 clinical features along with the expression data: gender, age, menopause status, binary response parameter for the occurrence of any new tumor event after the initial treatment, pathologic stage for each patient, neoplasm cancer status, pathologic T, which is a discrete parameter that measuring the total progression of the tumor, progression of the metastases (pathologic M), progression of cancer in the lymph nodes (pathologic N), and the overall progression (pathologic stage) [91] .

The assessment presented in the next section was done using 5-fold cross-validation with each fold containing 80% training data, and 20% testing data and accuracy of time to event prognosis was measured using c-index values (see 2.3.1.1).

### 3.4.2 Performance Comparison

Using the three omics data types, i.e., miRNA($p_1 = 1046$), methylation($p_1 = 5000$), and genes($p_1 = 20531$) in the BRCA dataset, we performed some comparisons and performance evaluation of the proposed *iNPS* method.

Firstly, we applied the unsupervised JIVE on the normalized concatenated data for extracting the joint and individual structures from these three data types and dimensionality reduction for subsequent survival analysis. Using the permutation test, the joint rank of $r = 2$ individual ranks of $r_1 = 4, r_2 = 35$, and $r_3 = 2$ were obtained for miRNA, methylation, and gene data types, respectively. The importance of effective multi-view data integration using JIVE can be seen from Figure 3.1. This figure shows the c-index performance of the time-to-event prediction on the BRCA test data using RSF with 5-fold cross-validation. The performance of each individual data type is shown with the highest median accuracy of the methylation data type. However, a simple concatenation of these data types did not result in a significant accuracy improvement. On the contrary, the median performance of the concatenated data was worse that each of the individual data types. This is most likely due to the HDLSS problem, i.e., $(p >> N)$ and here $p = 26577$ and $N = 620$. For the survival prognosis using JIVE, the joint and individual scores with the estimated ranks were viewed as individual columns of the composite feature vector for each patient [36]. Thus a total of 43 features were considered with $(2, 4, 35, 2)$ joint and individual ranks. This significant elimination of redundancy and dimensionality reduction while preserving the critical variations within and across the 3 data types using JIVE resulted in a 49.6% improvement in accuracy. This result clearly emphasizes the importance of an effective integration method prior to the downstream analysis.

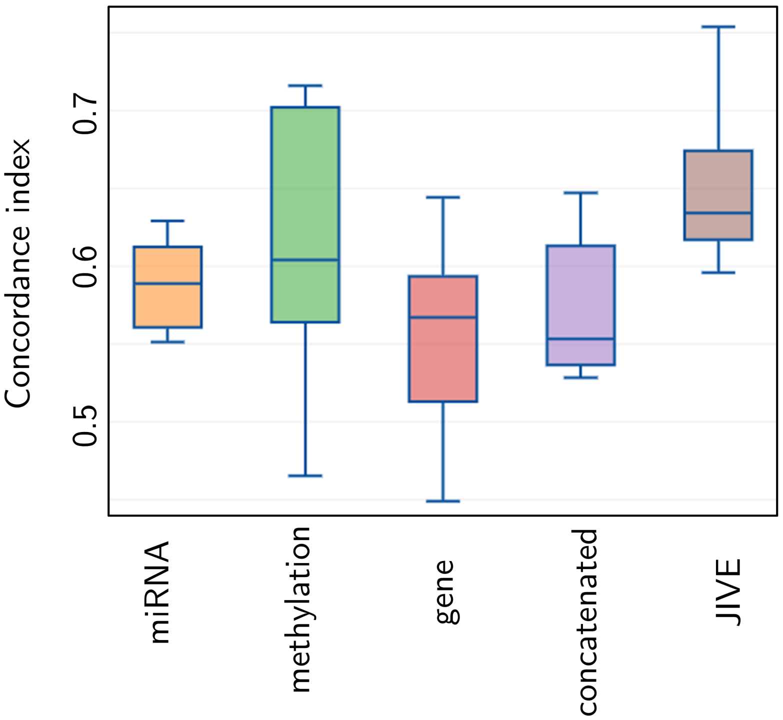

Figure 3.1: Box plot of estimated mean c-index for the test data in 5-fold cv scheme for the BRCA dataset for individual data types, concatenated data, and JIVE using RSF. The horizontal line inside the box represents the median and the box is bounded by the $25^{th}$ and $75^{th}$ percentile (IQR), whiskers extend to 1.5 × IQR.

Nonetheless, the maximum median c-index accuracy achieved by the JIVE even with the non-parametric RSF method was only 63.4%. One of the factors contributing to this low accuracy is the unsupervised nature of JIVE. We know that PCA is an unsupervised approach, hence the score estimation using JIVE (which is an extension of PCA) is also unsupervised, i.e., it does not take the response (survival time or status) into account when determining the scores. However, taking the response into account can help us identify the most important features. Hence we extend the unsupervised JIVE method to supervised JIVE using the supervised PCA method proposed in [93].

As shown in Fig. 3.2, taking the response into selecting the most critical joint and individual components using the supervised JIVE is shown to have improved performance than its unsupervised counterpart.

Nonetheless, the increased accuracy of 6.1% is still insufficient to handle the critical task of mortality prediction. Another reason for this low accuracy is the exclusive use of genomic features and not taking the clinical information into account for survival prognosis. It is known that a more remote genomic data from the physiological trait can have a lesser influence on it [94]. Further, some clinical information has been shown to improve the survival accuracy estimation in several studies. Hence, we selected 10 clinical features (as described in section 3.4.1.1) along with the 43 scores for the genomic data types. As expected, significant improvement can be seen after including the clinical features, as shown in Fig. 3.2. Including 10 clinical features along with the genomic features resulted in the accuracy of *iNPS* to increase to 87.5%. These 10 clinical features are described in section 3.4.1.1.

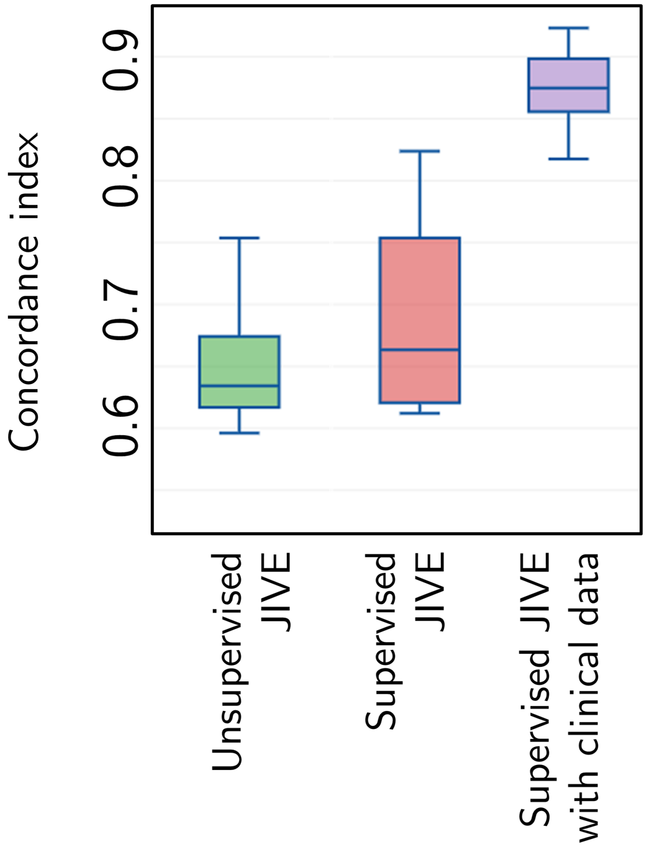

Figure 3.2: Box plot of estimated mean c-index for the test data in 5-fold cv scheme for the BRCA dataset for unsupervised JIVE, supervised JIVE, and supervised JIVE with 10 clinical features. The horizontal line inside the box represents the median and the box is bounded by the $25^{th}$ and $75^{th}$ percentile (IQR), whiskers extend to 1.5 × IQR.

All the time-to-event prognosis result shown so far in this section were done using the non-parametric RSF model, resulting in the best mean c-index of 87.5% for the JIVE dataset, including clinical features. However, as discussed in the foregoing, the survival prognosis in multi-view literature is primarily performed using the CPH model. Hence, in the subsequent assessment, we compared the result of RSF and CPH on the scores estimated by JIVE along with the clinical features. Figure 3.3 presents the box plot for mean error in each of the 5-fold cross-validation. With a 50.9% improvement in the c-index accuracy, significant evidence towards using non-parametric analysis is presented.

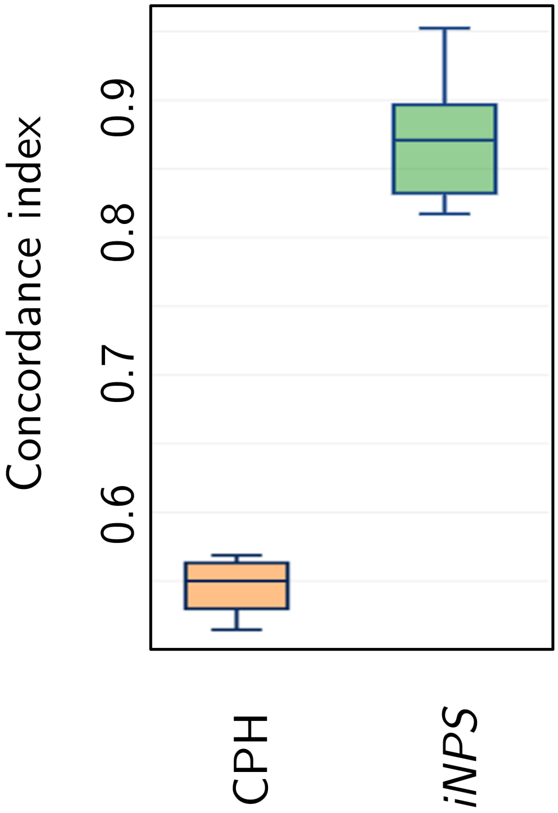

Figure 3.3: Box plot of estimated mean c-index for CPH and *iNPS* using BRCA test dataset in 5-fold cv scheme. The horizontal line inside the box represents the median and the box is bounded by the $25^{th}$ and $75^{th}$ percentile (IQR), whiskers extend to $1.5 \times$ IQR.

### 3.5 Summary

In this chapter, we introduced an *iNPS* model for survival analysis to address the limitations in handling multi-view data in survival analysis. This approach proposed an effective integration multi-view data to extract the joint and individual variations in each data type and reduce the dimension for subsequent downstream analysis using survival learning. The use of a non-parametric survival learning method using RSF is proposed to address the limitations imposed by the traditional proportional hazard models, which are often invalid in complex real-life data. Application for breast cancer prognosis using gene expression, methylation, and miRNA data types suggest a 50.9%

improvement in c-index accuracy of time to event prognosis of *iNPS* as compared to the use of CPH on the joint and individual components.

This is a relatively new area of research where both effective multi-view data integration and survival learning on multi-view data are active areas of research. Future work on extensive comparison of different survival learning techniques, including more interpretable parametric models such as accelerated failure time models and deep learning survival models, could be essential in developing understanding in their utilization for multi-view survival learning. Furthermore, a direct extension of the proposed method can be its application in pan-cancer survival analysis, where the vertical multi-view data is presented for different tumor types. This is a crucial area for discovering the evolution and prognosis of same cancer with varying types of tumors. Such detailed exploration and increased accuracy of the survival learning models can significantly impact cancer risk assessment, treatment decisions, and long-term survival estimations.

# 4. SURVIVAL LEARNING FOR PROGNOSIS USING SMART WEARABLES

## 4.1 Introduction

The key motivation of this study (as presented in chapter 1) is inspired by improving the healthcare data analysis outcomes such as personalized risk prediction and key biomarker identification via statistical development and accuracy improvement of survival models. While the earlier two chapters emphasized one aspect of this goal, i.e., the survival prognosis accuracy improvement by addressing the methodological challenges posed by healthcare big data and semi-parametric survival analysis models. This chapter focuses on increasing the broader impact of our goal by extending the application of the proposed survival learning methods beyond the retrospective analysis of clinical studies with a fixed follow-up/decision horizon time and often time-invariant covariates to continuous time-series healthcare data where the event is not death or mortality. More precisely, the current chapter of this dissertation is focused on applying survival learning for continuous prognosis of recurring epileptic seizure events. This will enable us to use survival learning as a robust prognostic tool to quantitatively estimate the risks of seizure at every time point. Such a seizure warning system will not only enable the patients to be aware of their impending risks over various time horizons and take necessary precautions but also allow the caregivers/medical experts to plan out and provide timely interventions.

This application is motivated by a broader objective of applying survival learning to one of the fastest-growing and most transformative areas of healthcare, i.e., personalized and precision medicine using smart wearables [38]. This application aims at extending survival analysis to wearable prognosis and bringing the strength of survival analysis models such as continuous prognosis to personalized health prediction.

Wearable devices that were initially developed for in-clinic use and health monitoring enthusiasts have significantly developed over time and are already transforming biomedicine through digital health and personalized medicine by facilitating continuous, longitudinal health monitoring outside of the clinic [38]. Such applications of smart wearables is crucial for shifting the status-quo of patient care from hospital to at-home, POC setting and to provide accessible quality care in rural and resource-limited areas. This shift of patient-care can only be possible through the development of machine learning algorithms and their integration with wearable sensors, transforming them from wearable to *smart wearable* technology [39].

Several studies in the wearable sensor literature have been focused on harnesing the innovations in healthcare electronics for the design of wearable devices to collect high-quality, real-time physiological data and its use in providing functionalities ranging from monitoring to rehabilitation [95]. For example, photoplethysmography based wearable watches that collect blood volume change data and provides heart rate and blood oxygenation values to the users [96, 97]. Moreover, the physiological data collection and monitoring has moved from vital measurement of heart rate to critical applications such as continuous glucose monitoring [98] and rehabilitation for disorders such as motion disorder in Parkison's disease [99]. Initial application of advanced analytics with wearable sensing (or smart wearables) have also facilitated the analysis of data for point-of-care diagnosis such as arrhythmia detection [100].

Nonetheless, most of the current wearable technologies are limited to monitoring or diagnosis and prognosis for event prediction using smart wearable is very limited. Furthermore, the limited methods that do exist are based on point-prediction and fail to provide a continuous prognosis which is much needed for smart wearables that are made to be worn continually by the patients. Application of survival analysis instead of the commonly used prediction methods not only enables continuous prediction but also opens other opportunities. For example, survival analysis models can be used to provide continuous and quantitative estimates for the occurrence of an event and thus can be used as a tool in event monitoring using smart wearables for critical and chronic health conditions such as the application proposed in this chapter for epileptic seizure alert systems via continuous prognosis using electroencephalography (EEG) data.

This chapter is organized as follows. In the first section, we provide a detailed description of the methodology for seizure prognosis using survival learning. The second section provides the details for results including the performance measure used, seizure data description, and the performance assessment of using survival learning for seizure prognosis. Finally, the last section summarizes the research work for this chapter.

## 4.2 Epileptic Seizure Prognosis

An epileptic seizure is characterized by recurrent seizures and a major challenge in managing epilepsy comes from the difficulties in predicting impending seizure events and associated adverse outcomes. In the absence of drugs that can halt the development of seizures, patients with epilepsy are always on a lookout for the "next" unforeseen seizure event which impairs their daily activities and are significantly detrimental to the quality of life [101]. Several efforts have been made in the seizure prediction literature to predict a seizure episode and provide early warning to the patients. However, the current seizure prediction studies are based on the conjecture that a "pre-ictal" state exists before a seizure onset and that this state is easily identifiable using EEG signals and can be used to give a warning for imminent seizure events (ictal state) [102]. Thus, the current models are unable to provide continuous risk estimates, i.e., probability of seizure event risk at any given point in time. To address this limitation, we propose a new approach by utilizing time-to-event survival analysis for seizure prediction. In this method, a quantitative risk of a seizure event is predicted continuously at every time point. Consequently, it eliminates the need for pre-ictal stage identification. To our knowledge, the present work is the first study to use survival analysis model in a seizure prediction context.

To apply survival learning on continuous EEG data, we first convert the EEG signal into seizure event data constituting the phase and frequency-based features, survival status, and the seizure event time information based on a decision horizon. A detailed discussion of the seizure event data is provided in the following section.

### 4.2.1 Seizure Event Data

To be able to apply survival learning to the streaming EEG data, we need to transform the streaming EEG data into the classical survival analysis format, i.e., for every subject, there will be an event time (seizure onset time), censoring status, and a fixed decision horizon or study duration, $T_D$ until when the subjects are monitored. However, this transformation is not trivial and has several challenges. These challenges and their corresponding solutions are listed below:

1. Multiple/ recurring event: the first and the main challenge is of multiple/recurring events. Unlike time-to-death analysis, where the event happens only once, seizure episodes can happen multiple times during a patient's lifetime. Hence, our primary task was to process and segment the data to allow only one event during a study. To accommodate this constraint, we selected the time horizon, $T_D$ as the global minimum time duration between two consecutive seizure events for the entire study population. Since a majority of inter-ictal times for the specimens were below 300 sec and following recommendations in the literature [103], a decision horizon $T_D = 300$ sec was chosen. Intuitively, this allows for treating every 300 sec of the data for a subject $i$ as a new subject. Let $T_i^0$ represent the actual time to first seizure event for the subject $i$ then the observed survival time for that $i^{th}$ subject in the decision horizon $T_D$ is given as $T_i = \min(T_i^0, T_D)$.

2. Censoring status: In accordance with the decision horizon, $T_D$ every 300 seconds of a subject's data is treated as a new subject, hence in all decision horizons where the seizure event was not observed, the status was given as "right-censored" ($\delta_i = 0$) and where the event occurred, status was given as "event" ($\delta_i = 1$). Status

is also "right-censored" during a decision horizon when the subject dies or is removed from the study (due to injuries or other possible reasons). Furthermore, since the seizure event can continue for some time duration, the status remained "event" until a decision horizon where the seizure has stopped and no episodes are found for the entire decision horizon.

Let $\boldsymbol{x}_i$ be the feature vector extracted from subject $i$ during time horizon $T_D$, then the survival data for subject $i$ is given as $\Phi_i = (\boldsymbol{x}_i, T_i, \delta_i)_{(1 \leq i \leq N)}$, where N is the total number of subjects for a given decision horizon. In this dissertation, we selected $T_D$ as 300 seconds, nonetheless, extension of this method to a much longer decision horizon is straightforward by selecting a longer $T_D$ value. The feature extraction is detailed in the next subsection.

### 4.2.2 Feature extraction and selection

Numerous features have been explored in the literature for their statistical significance in differentiating between the seizure and non-seizure characteristics of the EEG signal. These features include, but are not limited to, those extracted from frequency, time-scale domain analyses [104, 105] as well as other sophisticated geometric and graphical representations [106] and physiological transfer functions [107]. With the recent interest in studying the ripple patterns of EEG as precursors to seizure episodes, spectral features offer one of the convenient means to capture EEG signal patterns [102]. This formed the motivation for using the frequency domain features of EEG for seizure prognosis. We also propose a novel *snowball* feature which can accumulate the temporal variations in the EEG signal characteristics.

1. Frequency and phase features:

   The frequency-based features were estimated by applying a windowed Fourier transform, also referred to as a short-time Fourier transform (STFT) [108]. Here, the length of the sliding window, i.e., the time window in which the feature is estimated, determines the consistency of the estimated feature values. It is well known that seizure spreads very quickly and lasts just for a few seconds [109, 110]. Hence, the length of the window should be long enough to assure consistency of the feature value estimates but sufficiently short to maintain the purity of the EEG patterns the emerge during a seizure episode [110, 111, 112]. Based on different recommendations in the seizure prediction literature [109, 113], we considered three non-overlapping time windows of lengths $w = 1024, 2048, \text{and } 4096$ data points (i.e., $\approx 0.25 - 1s$) and extracted the spectral features, $x^E$ for different $w$, $x^E_{1024}$, $x^E_{2048}$, and $x^E_{4096}$ as the energy of the frequency content (measured as the sum square of the FFT magnitude). These were calculated over every 4Hz frequency band (e.g. 0-4 Hz) with an overlap of 1Hz over the $0 - 50$ Hz frequency range for all three windows. Due to the lack of consensus in the literature on the most effective time window length, $w$ for seizure prognosis [109, 113], we used features estimated from all three window lengths (i.e., every feature that was calculated had three values corresponding to each of the windows). This was done to ensure the robustness of the feature estimates. The frequency bandwidth and the range were selected to identify sensitive neural pathological activities that give rise to a seizure event [114, 115, 116]. For example, three energy feature values were computed for the delta band (0-4 Hz), one each for 1024, 2048, and 4096 window lengths. Following this procedure, 51 features (17 spectral features with each of the three window lengths) were obtained. Additionally, one overall variance, as a measure of the volatility of the data was calculated to give a total of 52 frequency spectral features. This feature estimation is summarized in Figure 4.1.

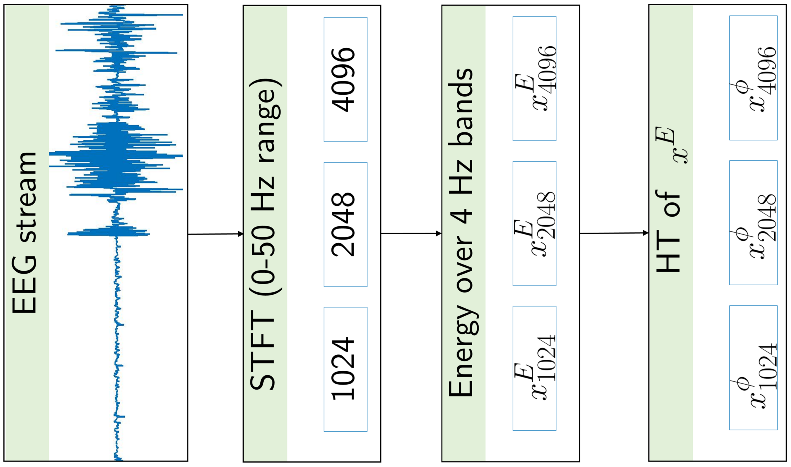

Figure 4.1: Summary of the spectral feature extraction procedure.

Although the frequency domain features are known to have a high significance in characterizing seizure and non-seizure EEG signals, they are bounded by the linearity and stationarity constraints. However, the EEG dynamics, in turn, are inherently nonlinear and non-stationary [117]. To overcome this limitation, we used the Hilbert Transform (HT), which is a prominent method for analyzing nonlinear and non-stationary processes [118]. Our approach employed HT to capture the shape of the frequency spectrum, especially the relative intensities of various frequency bands. These features are particularly relevant for early-stage prediction of seizure for the following reason. A well-developed seizure is marked by a high amplitude response over narrow frequency bands, typically spread over 4-30 Hz range. However, such a pronounced high-intensity frequency response is unlikely to standout a few minutes before seizure onset. Therefore, it becomes hard to detect seizures at early stages using frequency features alone. However, systematic, subtle changes in the shape of the frequency portrait, especially in the distribution of energy across the various frequency bands, can portend the emergence of seizure and associated intensification of certain frequency bands of

EEG signals. The instantaneous phase values $x^{\phi}_{1024}$, $x^{\phi}_{2048}$, $x^{\phi}_{4096}$ estimated from the HT of the spectral features are used to capture this subtle variation.

This procedure holds a similarity to cepstrum analysis, in that we treated the spectral energy features as a waveform [119]. However, unlike cepstrum analysis, we do not conduct a logarithmic transform of the energy values. Consequently, this approach allows the capture of small variations in the dominant frequency bands, and thereby track the variation of the salient frequency bands over time relative to each other [120]. Corresponding to the 52 spectral features $x^E$, 52 phase features $x^{\phi}$ were obtained.

2. Snowball features: In this dissertation, we introduce a new feature called Snowball feature as a way to accumulate the time-related changes [121] in the frequency and phase-based features. In the epilepsy literature, it is well known that impending seizure characteristics often build slowly over time [122]. Therefore, the accuracy of seizure prognosis can be increased by accumulating the subtle, consistent change between the features in successive windows. The proposed snowball features, based on capturing the cumulative changes in the feature values over time, are given by Equation 4.1.

$$\mathcal{S}(W) = x^w(i+1) - x^w(i) \tag{4.1}$$

In Equation 4.1, $x^w(i)$ represents a given feature at time window $i$ and $x^w(i+1) - x^w(i)$ captures the dynamic change in the features measured between two consecutive windows. In our work, these changes were accumulated over a period of $t = 30$ sec ($W = t/w$). Furthermore, by calculating the features over a small window length (that ensures consistency of the estimates while reducing the effects of transients) and then obtaining their corresponding snowball features, we ensured that our feature set encapsulates non-stationarity without compromising statistical accuracy. The snowball features were obtained for every frequency and phase-based features.

Finally, with 52 snowball frequency features and 52 snowball phase features, a 104-dimensional feature vector was extracted from the data. However, the high dimensionality of the feature space (104 features) can lead to added computational complexity in a large dataset, especially with streaming data. To reduce the dimensionality, the features most effective in the prognosis of seizure events were selected based on Minimal Depth (MD) statistics of RSF. MD is a relatively new high-dimensional feature selection method based on the order statistics for a tree [62]. MD measures the feature importance in terms of its splitting performance relative to the root node. Here, the argument employed is that the feature that frequently split the nodes farther down from the root node effects a relatively smaller sample of the original data in the root node. Thus, these features do not significantly affect the leaf node assignments as compared to the features that frequently split the root node and nodes closer to it and hence are less informative. Additionally, MD is more robust as compared to the Variable Importance (VIMP) measure, which is commonly employed for feature selection in tree-based methods [42]. VIMP measures the feature importance by the increase/decrease in prediction error by features when it is randomly "noised up". Reliance on the prediction error makes VIMP largely an ad-hoc method and highly susceptible to bias in the model. Hence, we preferred MD over VIMP. Out of the 104 features, 20 most significant features which were selected where features less than an MD threshold were most important (for more details on threshold selection please refer to [42]). In the selected top 20 features, 6 were spectral features and the rest 14 were phase-based features.

## 4.3 Results

### 4.3.1 Performance Measure

In this chapter, four performance metrics were used to evaluate the performance of the proposed seizure prognosis model. The first two metrics, C-index and Integrated Brier Score (IBS), as described in section 2.3.1 are used to assess the survival learning model. In addition to C-index and IBS, we also used the sensitivity and specificity evaluation metrics that are commonly used in seizure prediction literature. It is commonly noted that sensitivity is more critical than specificity as a missed prediction of a seizure episode can become perilous for the patient. At the same time, a very low specificity can become a nuisance by providing seizure alert even without the presence of an actual seizure episode. Furthermore, the sensitivity and specificity are measured at four different prediction horizons to test the prognosis accuracy as the seizure event becomes imminent. To validate the consistency of the result, all four performances were measured via a 10 fold cross-validation.

### 4.3.2 Data Description

#### *4.3.2.1 Small Mammal EEG Data*

The performance was assessed on intracranial EEG data from 80 small mammals (mice and rats). Experiments, as well as data collection, were conducted at the Department of Neuroscience and Experimental Therapeutics at the Texas A&M College of Medicine. As part of the experiments, an intracranial electrode (PlasticsOne, Roanoke, VA) was inserted into the hippocampus of 80 small mammals. EEG readings (in milli-Volt) at 4 KHz sampling frequency were acquired for each of these specimens continuously over a 3-month period. These specimens were exposed to organophosphate (OP) intoxication according to the protocol previously published [123, 124]. This treatment induces the mammal subjects with progressive chronic epilepsy, which were random in occurrence and seizure intensity, closely mimicking epileptic episodes of human patients [125]. This model is quite different from chemical toxin injections, such as kainic acid or pilocarpine that are excitatory chemicals not normally exposed to humans or not used for poisoning, so they are purely for lab research only, not relevant to field situation or cannot be extrapolated to humans. All animal procedures were approved by the university's institutional animal care and use committee in compliance with the guidelines of NIH Guide for the Care and Use of Laboratory Animals [126]. The EEG signals were recorded at 4 KHz sampling rate in an abf-1.8 (Axon Binary Format) file format. The continuous EEG data was collected 24/7 for 3 months. For each animal recording, we took snapshots of data for 2 days for seizure prognostics study. The spontaneous seizures occurring during this time span is a representation of the chronic epilepsy state with seizure occurring normally [123]. Figure 4.2 shows a 20 second long strip of EEG data (10 sec of normal state juxtaposed with 10 sec of epileptic EEG) acquired from one of the 80 rat specimens employed for this test. As shown in the figure, the EEG patterns (including the signal energy) during a seizure episode differ from those during a normal state. The EEG signals were carefully annotated by trained technicians to identify epileptic episodes and different period of EEG, i.e., non-ictal, pre-ictal, ictal segments including their onset and offset points [127]. For verification, seizure episodes were correlated with continuous video measurements taken during the tests.

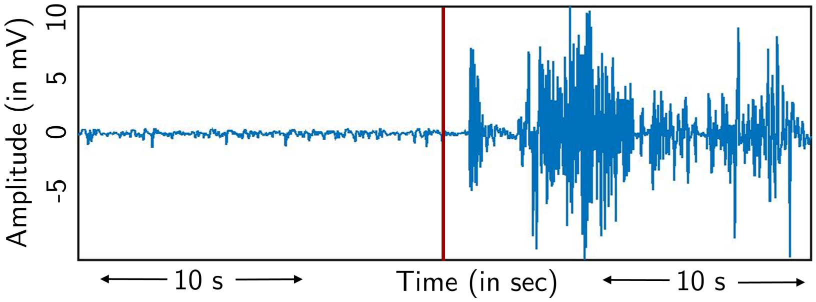

Figure 4.2: Representation of 10-sec strips of normal and epileptic EEG signal recorded by the hippocampal intracranial electrodes of a rat specimen.

### 4.3.3 Performance Comparison

To determine which survival learning method to use, we first compare the effectiveness of RSF and CPH for the prognosis of seizure events by comparing them on the IBS error measure. Figure 4.3 shows the variation of the prediction error over the decision horizon for RSF, CPH, and a reference Kaplan Meir method [12]. On the basis of IBS, RSF yields the best predictions among the three methods tested. To validate the consistency of this result, we performed a 10 fold cross validation test. The results of the 10 fold cross validation is summarized in Figure 4.4. Evidently, RSF performed consistently better in terms of the IBS values as well as their variation. The C-index score for both the model CPH and RSF models averaged over the 300-sec duration are summarized in Table 4.1 with RSF outperforming CPH with an 87.5% reduction in the IBS score and a 17.5% increase in the C-index value. Extension of this approach to other, less controlled scenarios where the mean inter-ictal times can be much higher is rather straightforward in the sense that we need to extend the time of prediction error curve evaluation beyond the current 300-sec decision span [14].

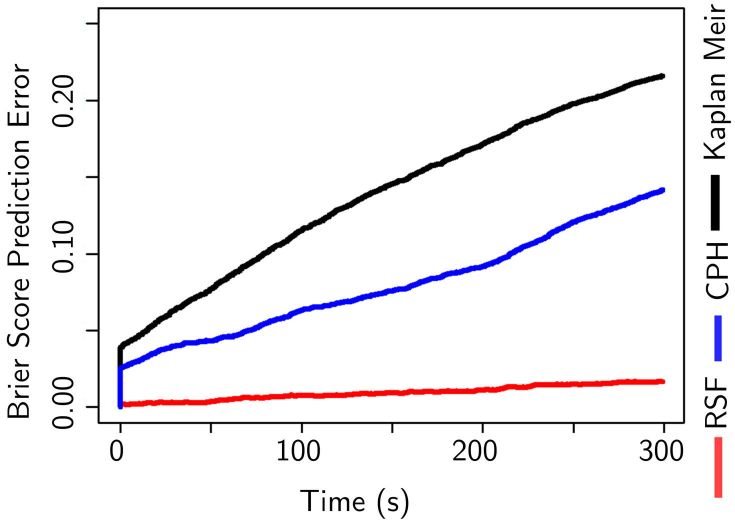

Figure 4.3: Prediction error curves (PEC) for RSF (in red), CPH (in blue), and Kaplan-Mier (in black) for the test dataset.

Table 4.1: Summary of RSF seizure prognosis performance for the test dataset.

| **Error measure** | CPH | RSF |
|---|---|---|
| C-index | 80.6 | 94.7 |
| IBS | 0.08 | 0.01 |

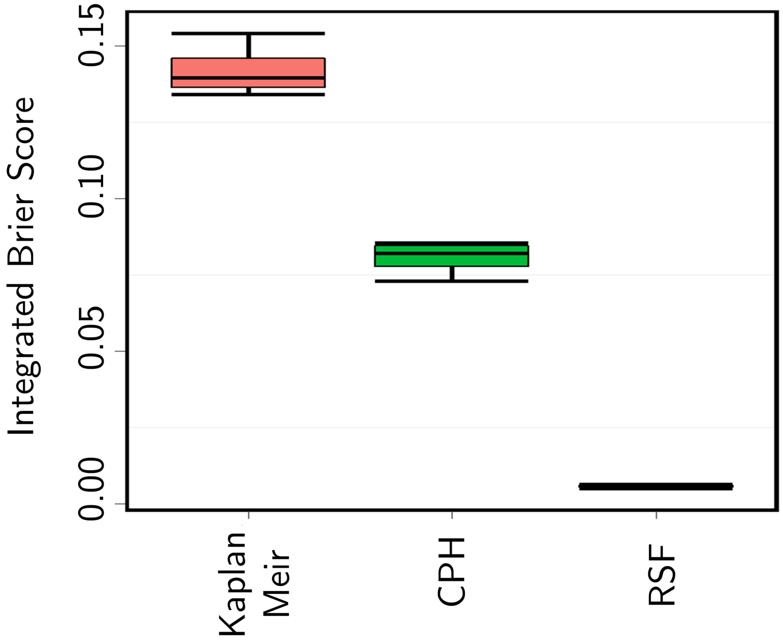

Figure 4.4: Boxplots of IBS error for Kaplan-Meir, CPH, and RSF models, calculated for the test data in 10-fold cross validation scheme. The horizontal line inside the box represents the median and the box is bounded by the 25th and 75th percentiles.

Now with established preference of RSF, we used it for further analysis of the seizure data using the other two error metrics, time varying sensitivity and specificity. Figure 4.5 presents the survival probability plots for non-seizure and a seizure events with each curve representing RSF's estimate of the survival probability (or inverse of cumulative hazard) for the event occurrence at any given time. Clearly, the survival probability is higher when the event does not occur and is lower otherwise.

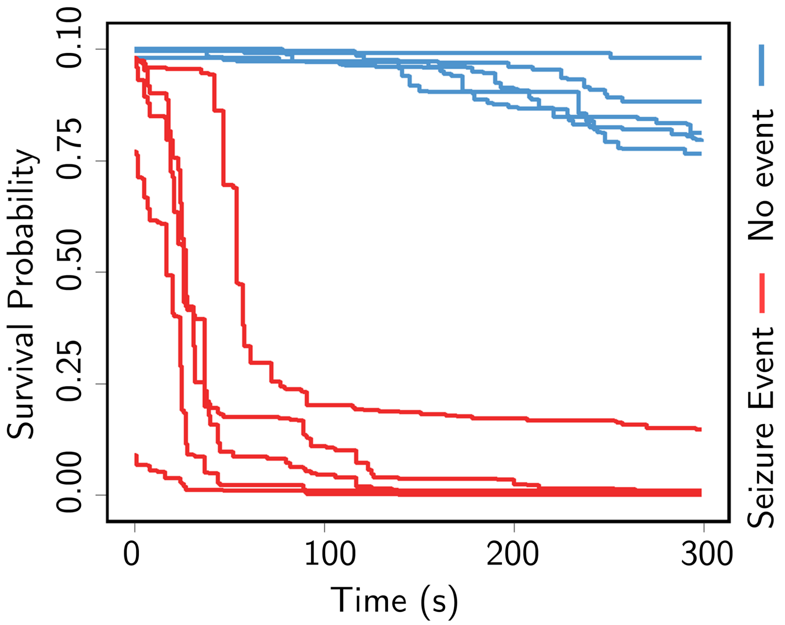

Figure 4.5: Survival probability curves for seizure event (in red) and no seizure event (in blue) obtained using the RSF methodology.

Furthermore, the sensitivity and specificity at different prediction horizons from time periods of $60 - 300$ seconds and for the out-of-sample test data are summarized in Table 4.2. These accuracies were calculated based on a threshold probability value of 0.69 to classify between a seizure vs non-seizure event. This threshold was chosen to be significantly higher than 0.5 for the robustness of the estimated sensitivity and specificity and was also based on the accuracy of the model during the training phase. It is noted that both sensitivity and specificity improve as the time to seizure event decreases. This result is in line with the expected outcome and supports the relevance of the proposed work for seizure warning applications.

Table 4.2: Performance of the prediction model on test dataset at different time points.

| Event ID | Time to event (s) | Sensivity (%) | Specificity (%) |
|---|---|---|---|
| 1 | 300 | 82.40 | 77.40 |
| 2 | 240 | 77.43 | 85.78 |
| 3 | 120 | 82.19 | 85.00 |
| 4 | 60 | 82.99 | 86.79 |

### 4.4 Summary

In this chapter, we introduced a survival learning- based seizure prognosis model to provide a continuous prognosis of seizure events,i.e., the probability distribution of the time remaining until the next seizure event based on harnessing information from the measured EEG data. This new approach allows the estimation of the times till the next event when the event is not "death" but instead a recurring epileptic seizure episode. EEG data used in this study was continuously collected over a period of 3-months from intracranial electrodes placed in the hippocampus of organophosphate intoxicated rodents. Using EEG data from the rat and mice specimens, we demonstrated the robust prognostic ability of the proposed method. The average test IBS error and C-index of using RSF for continuous seizure prognosis was 0.01 and 94.7, respectively. As compared to the current "gold standard" CPH survival model, the IBS decreased by 87.5% and C-index increased by 17.5%. These consistent empirical results open new opportunities for using survival analysis-based seizure prediction and quantitative risk management approaches. The current study in this chapter employs rodent EEG to assess the proof of concept of the novel approach. Our future work includes using the proposed model for seizure prediction in human subjects and the adoption of an adaptive, risk-informed threshold for a seizure warning system using smart wearables.

Application of the proposed method in a seizure warning system can enable the patient and caregivers to monitor the likelihood of seizure at any given point in time instead of waiting for sudden and unexpected warnings or having to train a separate model to predict the likelihood of seizure at different future times. From a clinical perspective, multiple physical symptoms and patterns of EEG exist as a forbearer of an impending seizure, but these symptoms rely on patient and health care provider's "intuition" or other qualitative assessment. Having a tool with continuous, quantitative assessment of seizure will help to relate these symptoms quantitatively to the seizure likelihood. Such studies can ultimately improve the fundamental knowledge of epilepsy and its behavior. Further, it can address the missing data problem by the use of censoring. Signal recording using smart wearables is extremely prone to missing data for multiple reasons including device malfunction, turning the device off at night, etc. While the traditional machine learning method suffers from a significant decrease in accuracy as a result of missing data, the concept of censoring in survival analysis can be used to effectively handle them.

Another future work direction is using the proposed model for seizure prognosis over a much longer prediction horizon with consideration of the influence of environment and other exogenous factors. We anticipate that our work would spur further efforts in this front.

# 5. CONCLUSIONS AND FUTURE DIRECTIONS

## 5.1 Conclusions

The overarching goal of this dissertation was to improve the time-to-event prognosis accuracy for better healthcare outcomes. To achieve this goal, we focused on the methodological developments required to address the big data challenges in the survival data and address the limitation of traditional parametric and proportional hazard models. Accuracies of the proposed methods were demonstrated through application in critical healthcare areas such as cardiac disease and cancer to improve mortality prediction. The application was also extended to beyond the time-to-death prediction by applying survival learning for epileptic seizure prediction where events are multiple/recurring seizure episodes. The big data challenges in survival learning were addressed in Chapter 2 and Chapter 3, and the application for seizure prediction and smart wearables were addressed in Chapter 4.

In Chapter 2, we proposed a balanced random survival forest or *BRSF* method to address the challenge of data imbalance in healthcare big data. The proposed *BRSF* method integrated a synthetic minority over-sampling technique with RSF learning for mortality prediction. Theoretical results were used to establish the negative impact of imbalance in survival data for hazard estimation. Through intensive empirical studies, we demonstrated that the prognosis accuracy significantly increased after survival models were trained with balanced data sets. In terms of the integrated Brier score and concordance index, the balanced RSF performed 25% and 55% better that the RSF, respectively. The balanced CPH performed 8.2% and 24%, respectively, as compared to CPH.

In Chapter 3, we proposed an integrated non-parametric survival learning or *iNPS* learning method to address the challenge of accurate survival prognosis using multi-view data. In the first part of this chapter, we used a joint and individual variation explained (JIVE) method to effectively integrate the multi-view data to separate and analyze the joint (or shared) and individual variations in different data types. Next, a non-parametric survival learning method was used on the integrated data for increased prognosis accuracy compared to the most gold-standard CPH model. Empirical studies on simulated data and the breast invasive carcinoma data showed that the proposed *iNPS* method had a 50.9% better performance than the CPH.

Finally, in the last chapter, we extended the application of survival learning beyond time to "death" prediction. We used an RSF model for time to epileptic seizure prognosis using continuous EEG signals and recurrent/multiple seizure events. We assessed the proposed application using intracranial EEG data from small mammals. The model performance was 0.01 and 94.7 based on the integrated Brier score and concordance index. Further, the sensitivity and specificity were 82.4% and 77.4%, respectively, 300s before a seizure episode was imminent.

### 5.2 Future Directions

Survival analysis is one of the oldest methods that merged medical or clinical data and data analysis. With the increasing data collection and use of machine learning and data analytics for clinical decision making, it's relevance is greater than ever. Several opportunities exist, both for its statistical advancement and more innovative applications. Along these lines, we plan to extend the future work in the following directions:

1. Directionally dependent multi-view survival learning: In chapter 3, we introduced survival learning using multi-view data. In recent healthcare studies, it is a fundamental problem to integrate different data types for the same subject/patients. Exploring the association between multiple data types has been shown to reveal significant insights. Hence, we used a joint and individual variation explained

method for the integrated analysis of the multi-view data and to analyze the joint and individual structure in multiple data types. However, accurately estimating the associations between real-world data types is rarely feasible. This is especially true in genomic data with complex underlying biology. For example, the central dogma of molecular biology that governs the flow of information between different omics level makes this association directional or causal. Different data types are now directionally dependent with a pre-specified direction of dependence. Incorporating this directional dependency in survival learning can provide several insights, including the most significant features and the influence structure/network of the features on the mortality prediction. One of our recent work incorporates this directionality using a biology-inspired Bayesian integrated multi-view clustering model [33]. We used an asymmetric copula to accommodate the directional dependencies between the data types before clustering. Significant research needs to be done in this nascent area to capture such biological insights in the statistical model effectively. Furthermore, extensions in survival analysis, i.e., a complete analysis of dependence/directional dependence from genotype to phenotype, can provide unprecedented insights.

2. Application of multi-view survival learning for smart wearable seizure prognosis: One primary issue that limits the broader adoption of smart wearables for diagnosis and prognosis is their limited accuracy, especially when concerning a critical application such as seizure prognosis. This limited accuracy is often a result of low signal to noise ratio of the data collected using smart wearables due to movement-related artifacts. Nonetheless, due to the miniaturization of electronic components, smart wearables often have more than one sensor, for example, the most common wearable for physiological monitoring, Fitbit, within its small watch configuration includes a photoplethysmography sensor, 3-axis accelerometer, an altimeter, and a gyroscope. Similarly, a device for seizure prediction can

have an electroencephalography sensor, accelerometer, and electrooculography to detect the seizure status of a person at any given time. These data types can be treated as multi-view data, and vertical integration techniques (see Chapter 3) along with survival learning, can be used for effective seizure prognosis. Using multi-view survival analysis will increase the prognosis accuracy by jointly improving the signal to noise ratio.

# REFERENCES

[1] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, vol. 360. Hoboken, NJ: John Wiley & Sons, 2011.

[2] J. Graunt, *Natural and political observations made upon the bills of mortality*. London: The Royal Society, 1939.

[3] X. Liu, *Survival analysis: models and applications*. Chichester, West Sussex: John Wiley & Sons, 2012.

[4] S. T. S. Bukkapatnam, K. Afrin, D. Dave, and S. R. T. Kumara, "Machine learning and AI for long-term fault prognosis in complex manufacturing systems," *CIRP Annals*, vol. 68, no. 1, pp. 459–462, 2019.

[5] B. Liang and H. Park, "Predicting hedge fund failure: A comparison of risk measures," *Journal of Financial and Quantitative Analysis*, vol. 45, no. 1, pp. 199–222, 2010.

[6] P. Giot and A. Schwienbacher, "IPOs, trade sales and liquidations: Modelling venture capital exits using survival analysis," *Journal of Banking & Finance*, vol. 31, no. 3, pp. 679–702, 2007.

[7] A. D. Galbreath, R. A. Krasuski, B. Smith, K. C. Stajduhar, M. D. Kwan, R. Ellis, and G. L. Freeman, "Long-term healthcare and cost outcomes of disease management in a large, randomized, community-based population with heart failure," *Circulation*, vol. 110, no. 23, pp. 3518–3526, 2004.

[8] E. A. Amsterdam, N. K. Wenger, R. G. Brindis, D. E. Casey, T. G. Ganiats, D. R. Holmes, A. S. Jaffe, H. Jneid, R. F. Kelly, M. C. Kontos, *et al.*, "2014 aha/acc guideline for the management of patients with non–st-elevation acute coronary syndromes: executive summary: a report of the american college of

cardiology/american heart association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 64, no. 24, pp. 2645–2687, 2014.

[9] P. T. O'Gara, F. G. Kushner, D. D. Ascheim, D. E. Casey, M. K. Chung, J. A. De Lemos, S. M. Ettinger, J. C. Fang, F. M. Fesmire, B. A. Franklin, *et al.*, "2013 accf/aha guideline for the management of st-elevation myocardial infarction," *Journal of the American College of Cardiology*, vol. 61, no. 4, pp. e78–e140, 2013.

[10] L. Ohno-Machado, "A comparison of cox proportional hazards and artificial neural network models for medical prognosis," *Computers in Biology and Medicine*, vol. 27, no. 1, pp. 55–65, 1997.

[11] S. Walker and B. K. Mallick, "A bayesian semiparametric accelerated failure time model," *Biometrics*, vol. 55, no. 2, pp. 477–483, 1999.

[12] B. Efron, "Logistic regression, survival analysis, and the kaplan-meier curve," *Journal of the American Statistical Association*, vol. 83, no. 402, pp. 414–425, 1988.

[13] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[14] B. Mogensen, H. Ishwaran, and T. A. Gerds, "Evaluating random forests for survival analysis using prediction error curves," *Journal of Statistical Software*, vol. 50, no. 11, pp. 1–23, 2012.

[15] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, *et al.*, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[16] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, 2018.

[17] K. Afrin, G. Illangovan, S. S. Srivatsa, and S. T. S. Bukkapatnam, "Balanced random survival forests for extremely unbalanced, right censored data," *arXiv preprint arXiv:1803.09177*, 2018.

[18] A. P. Furnary, J.-C. Chachques, L. F. Moreira, G. L. Grunkemeier, J. S. Swanson, N. Stolf, S. Haydar, C. Acar, A. Starr, A. D. Jatene, *et al.*, "Long-term outcome, survival analysis, and risk stratification of dynamic cardiomyoplasty," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 112, no. 6, pp. 1640–1650, 1996.

[19] R. Dankner, U. Goldbourt, V. Boyko, H. Reicher-Reiss, B. S. Group, *et al.*, "Predictors of cardiac and noncardiac mortality among 14,697 patients with coronary heart disease," *The American Journal of Cardiology*, vol. 91, no. 2, pp. 121–127, 2003.

[20] D. Delen, A. Oztekin, and Z. J. Kong, "A machine learning-based approach to prognostic analysis of thoracic transplantations," *Artificial Intelligence in Medicine*, vol. 49, no. 1, pp. 33–42, 2010.

[21] Z.-H. Zhou, N. V. Chawla, Y. Jin, and G. J. Williams, "Big data opportunities and challenges: Discussions from data analytics perspectives," *IEEE Computational Intelligence Magazine*, vol. 9, no. 4, pp. 62–74, 2014.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[23] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[24] E. DeRouin, J. Brown, H. Beck, L. Fausett, and M. Schneider, "Neural network training on unequally represented classes," *Intelligent Engineering Systems through Artificial Neural Networks*, pp. 135–145, 1991.

[25] N. V. Chawla, N. Japkowicz, and A. Kotcz, “Special issue on learning from imbalanced data sets,” *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[26] N. Japkowicz *et al.*, “Learning from imbalanced data sets: a comparison of various strategies,” in *AAAI workshop on learning from imbalanced data sets*, vol. 68, pp. 10–15, Menlo Park, CA, 2000.

[27] S. V. Parikh and J. A. De Lemos, “Biomarkers in cardiovascular disease: integrating pathophysiology into clinical practice,” *The American Journal of the Medical Sciences*, vol. 332, no. 4, pp. 186–197, 2006.

[28] Y. Li, F.-X. Wu, and A. Ngom, “A review on machine learning principles for multi-view biological data integration,” *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 325–340, 2018.

[29] Cancer Genome Atlas Network, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[30] G. Martinelli and N. Foreman, “Advancing precision medicine through multi-omics: An integrated approach to tumor profiling,” *Science*, vol. 349, no. 6253, pp. 1246–1246, 2015.

[31] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, “Joint and individual variation explained (JIVE) for integrated analysis of multiple data types,” *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 523–542, 2013.

[32] D. M. Witten and R. J. Tibshirani, “Extensions of sparse canonical correlation analysis with applications to genomic data,” *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 1–27, 2009.

[33] K. Afrin, A. S. Iquebal, M. Karimi, A. Souris, S. Y. Lee, and B. K. Mallick, “Directionally dependent multi-view clustering using copula model,” *arXiv preprint arXiv:2003.07494*, 2020.

[34] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[35] L.-J. Wei, "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis," *Statistics in Medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992.

[36] A. Kaplan and E. F. Lock, "Prediction with dimension reduction of multiple molecular data sources for patient survival," *Cancer Informatics*, vol. 16, pp. 1–11, 2017.

[37] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, *et al.*, "Salmon: Survival analysis learning with multi-omics neural networks on breast cancer," *Frontiers in Genetics*, vol. 10, p. 166, 2019.

[38] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution," *Personalized medicine*, vol. 15, no. 5, pp. 429–448, 2018.

[39] X. Lopez, K. Afrin, and B. Nepal, "Examining the design, manufacturing, and analytics of smart wearables," *Medical Devices & Sensors*, vol. 3, no. 3, p. e10087, 2020.

[40] C.-C. Chia, I. Rubinfeld, B. M. Scirica, S. McMillan, H. S. Gurm, and Z. Syed, "Looking beyond historical patient outcomes to improve clinical models," *Science Translational Medicine*, vol. 4, no. 131, pp. 131ra49–131ra49, 2012.

[41] V. V. Belle, K. Pelckmans, J. A. Suykens, and S. V. Huffel, "Learning transformation models for ranking and survival analysis," *Journal of Machine Learning Research*, vol. 12, no. 23, pp. 819–862, 2011.

[42] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn, "Random survival forests for high-dimensional data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 1, pp. 115–132, 2011.

[43] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. New York, NY: Springer, 2013.

[44] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] M. R. Segal, “Regression trees for censored data,” *Biometrics*, vol. 44, no. 1, pp. 35–47, 1988.

[46] E. Hsich, E. Z. Gorodeski, E. H. Blackstone, H. Ishwaran, and M. S. Lauer, “Identifying important risk factors for survival in patient with systolic heart failure using random survival forests,” *Circulation: Cardiovascular Quality and Outcomes*, vol. 4, no. 1, pp. 39–45, 2011.

[47] C. Chen, A. Liaw, and L. Breiman, “Using random forest to learn imbalanced data,” *University of California, Berkeley*, vol. 110, pp. 1–12, 2004.

[48] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[49] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, “Learning from imbalanced data in surveillance of nosocomial infection,” *Artificial Intelligence in Medicine*, vol. 37, no. 1, pp. 7–18, 2006.

[50] E. Byon, A. K. Shrivastava, and Y. Ding, “A classification procedure for highly imbalanced class sizes,” *IIE Transactions*, vol. 42, no. 4, pp. 288–303, 2010.

[51] D. D. Lewis and J. Catlett, “Heterogeneous uncertainty sampling for supervised learning,” in *Machine learning proceedings 1994*, pp. 148–156, Elsevier, 1994.

[52] N. Japkowicz, “The class imbalance problem: Significance and strategies,” in *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, pp. 111–117, 2000.

[53] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, “Evaluating the yield of medical tests,” *JAMA*, vol. 247, no. 18, pp. 2543–2546, 1982.

[54] T. A. Gerds and M. Schumacher, "Consistent estimation of the expected brier score in general survival models with right-censored event times," *Biometrical Journal*, vol. 48, no. 6, pp. 1029–1040, 2006.

[55] T. A. Gerds and M. Schumacher, "Efron-type measures of prediction error for survival analysis," *Biometrics*, vol. 63, no. 4, pp. 1283–1287, 2007.

[56] R. A. Kyle, ""Benign" monoclonal gammopathy—after 20 to 35 years of follow-up," in *Mayo Clinic Proceedings*, vol. 68, pp. 26–36, Elsevier, 1993.

[57] H. S. Jørgensen, H. Nakayama, J. Reith, H. O. Raaschou, and T. S. Olsen, "Acute stroke with atrial fibrillation," *Stroke*, vol. 27, no. 10, pp. 1765–1769, 1996.

[58] A. C. Sawant, S. R. Narra, P. K. Mills, and S. Srivatsa, "Prognostic value of frontal qrs-t angle in predicting survival after primary percutaneous coronary revascularisation/coronary artery bypass grafting for stemi," *Journal of the American College of Cardiology*, vol. 61, no. 10, p. E97, 2013.

[59] C. L. Loprinzi, J. A. Laurie, H. S. Wieand, J. E. Krook, P. J. Novotny, J. W. Kugler, J. Bartel, M. Law, M. Bateman, and N. E. Klatt, "Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group.," *Journal of Clinical Oncology*, vol. 12, no. 3, pp. 601–607, 1994.

[60] H. Ishwaran and U. B. Kogalur, "Random survival forests for r," *R news*, vol. 7, no. 2, pp. 25–31, 2007.

[61] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees.* Boca, Raton, FL: CRC press, 1984.

[62] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 205–217, 2010.

[63] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. New York, NY: Springer series in statistics, 2001.

[64] J. Ehrlinger, "ggrandomforests: Exploring random forest survival," *arXiv preprint arXiv:1612.08974*, 2016.

[65] F. Ottani, M. Galvani, F. A. Nicolini, D. Ferrini, A. Pozzati, G. Di Pasquale, and A. S. Jaffe, "Elevated cardiac troponin levels predict the risk of adverse outcome in patients with acute coronary syndromes," *American Heart Journal*, vol. 140, no. 6, pp. 917–927, 2000.

[66] B. L. Croal, G. S. Hillis, P. H. Gibson, M. T. Fazal, H. El-Shafei, G. Gibson, R. R. Jeffrey, K. G. Buchan, D. West, and B. H. Cuthbertson, "Relationship between postoperative cardiac troponin i levels and outcome of cardiac surgery," *Circulation*, vol. 114, no. 14, pp. 1468–1475, 2006.

[67] National Human Genome Research Institute, "Genome technology program." https://www.genome.gov/Funded-Programs-Projects/Genome-Technology-Program. Accessed: June 10,2020.

[68] International Cancer Genome Consortium, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–966, 2010.

[69] K. J. Karczewski and M. P. Snyder, "Integrative omics for health and disease," *Nature Reviews Genetics*, vol. 19, no. 5, p. 299, 2018.

[70] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: recent progress in multi-omics data integration methods," *Frontiers in Genetics*, vol. 8, p. 84, 2017.

[71] Z. Yang and G. Michailidis, "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data," *Bioinformatics*, vol. 32, no. 1, pp. 1–8, 2015.

[72] D. M. Reif, B. C. White, and J. H. Moore, "Integrated analysis of genetic, genomic and proteomic data," *Expert Review of Proteomics*, vol. 1, no. 1, pp. 67–75, 2004.

[73] B. B. Cummings, J. L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort, A. R. Foley, V. Bolduc, L. B. Waddell, S. A. Sandaradura, G. L. O'Grady, *et al.*, "Improving genetic diagnosis in mendelian disease with transcriptome sequencing," *Science Translational Medicine*, vol. 9, no. 386, p. eaal5209, 2017.

[74] A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, and D. Greco, "MVDA: a multi-view genomic data integration methodology," *BMC Bioinformatics*, vol. 16, no. 1, p. 261, 2015.

[75] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[76] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, "Analysis of multiblock and hierarchical pca and pls models," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 12, no. 5, pp. 301–321, 1998.

[77] E. F. Lock, *Vertical Integration of Multiple High-DimensionalDatasets*. PhD thesis, The University of North Carolina at Chapel Hill, 2012.

[78] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190, Springer, 1992.

[79] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 92–100, 1998.

[80] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 1–34, 2009.

[81] J. Trygg and S. Wold, “O2-pls, a two-block (x–y) latent variable regression (lvr) method with an integral osc filter,” *Journal of Chemometrics*, vol. 17, no. 1, pp. 53–64, 2003.

[82] C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi, “Multilevel functional principal component analysis,” *The Annals of Applied Statistics*, vol. 3, no. 1, pp. 458–488, 2009.

[83] I. Måge, A. K. Smilde, and F. M. van der Kloet, “Performance of methods that separate common and distinct variation in multiple data blocks,” *Journal of Chemometrics*, vol. 33, no. 1, p. e3085, 2019.

[84] M. Jiang, *Statistical Learning of Integrative Analysis*. PhD thesis, The University of North Carolina at Chapel Hill, 2018.

[85] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[86] W. Wang, V. Baladandayuthapani, J. S. Morris, B. M. Broom, G. Manyam, and K.-A. Do, “iBAG: integrative bayesian analysis of high-dimensional multiplatform genomics data,” *Bioinformatics*, vol. 29, no. 2, pp. 149–159, 2012.

[87] M. Ruffalo, M. Koyutürk, and R. Sharan, “Network-based integration of disparate omic data to identify" silent players" in cancer,” *PLoS Computational Biology*, vol. 11, no. 12, p. e1004595, 2015.

[88] D. Kim, H. Shin, Y. S. Song, and J. H. Kim, “Synergistic effect of different levels of genomic data for cancer clinical outcome prediction,” *Journal of Biomedical Informatics*, vol. 45, no. 6, pp. 1191–1198, 2012.

[89] H. Chai, X. Zhou, Z. Cui, J. Rao, Z. Hu, Y. Lu, H. Zhao, and Y. Yang, “Integrating multi-omics data with deep learning for predicting cancer prognosis,” *bioRxiv*, p. 807214, 2019.

[90] Y. Wei, “Integrative analyses of cancer data: a review from a statistical perspective,” *Cancer Informatics*, vol. 14, no. Suppl. 2, pp. 173–181, 2015.

[91] N. Rappoport and R. Shamir, “Multi-omic and multi-view clustering algorithms: review and cancer benchmark,” *Nucleic Acids Research*, vol. 46, no. 20, pp. 10546–10562, 2018.

[92] M. Goldman, B. Craft, A. Brooks, J. Zhu, and D. Haussler, “The UCSC Xena platform for cancer genomics data visualization and interpretation,” *bioRxiv*, p. 326470, 2018.

[93] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, “Prediction by supervised principal components,” *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 119–137, 2006.

[94] H. Qin, T. Niu, and J. Zhao, “Identifying multi-omics causers and causal pathways for complex traits,” *Frontiers in genetics*, vol. 10, p. 110, 2019.

[95] G. Appelboom, E. Camacho, M. E. Abraham, S. S. Bruce, E. L. Dumont, B. E. Zacharia, R. D’Amico, J. Slomian, J. Y. Reginster, O. Bruyère, *et al.*, “Smart wearable body sensors for patient self-assessment and monitoring,” *Archives of Public Health*, vol. 72, no. 1, pp. 1–9, 2014.

[96] H. Fukushima, H. Kawanaka, M. S. Bhuiyan, and K. Oguri, “Estimating heart rate using wrist-type photoplethysmography and acceleration sensor while running,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2901–2904, IEEE, 2012.

[97] Y. Khan, A. E. Ostfeld, C. M. Lochner, A. Pierre, and A. C. Arias, “Monitoring of vital signs with flexible and wearable medical devices,” *Advanced Materials*, vol. 28, no. 22, pp. 4373–4395, 2016.

[98] J. Kim, A. S. Campbell, and J. Wang, “Wearable non-invasive epidermal glucose sensors: A review,” *Talanta*, vol. 177, pp. 163–170, 2018.

[99] N. L. Keijsers, M. W. Horstink, and S. C. Gielen, “Online monitoring of dyskinesia in patients with parkinson’s disease,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 22, no. 3, pp. 96–103, 2003.

[100] R. Gopalakrishnan, L. Korzinov, F. Wang, E. Thomson, N. Srivastava, O. Dawood, I. Abuzeid, and D. E. Albert, “Methods and systems for arrhythmia tracking and scoring,” Aug. 23 2016. US Patent 9,420,956.

[101] I. Younus and D. S. Reddy, “A resurging boom in new drugs for epilepsy and brain disorders,” *Expert Review of Clinical Pharmacology*, vol. 11, no. 1, pp. 27–45, 2018.

[102] H.-T. Shiao, V. Cherkassky, J. Lee, B. Veber, E. E. Patterson, B. H. Brinkmann, and G. A. Worrell, “Svm-based system for prediction of epileptic seizures from ieeg signal,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1011–1022, 2016.

[103] Y. Varatharajah, R. K. Iyer, B. M. Berry, G. A. Worrell, and B. H. Brinkmann, “Seizure forecasting and the preictal state in canine epilepsy,” *International Journal of Neural Systems*, vol. 27, no. 01, p. 1650046, 2017.

[104] V. Srinivasan, C. Eswaran, Sriraam, and N, “Artificial neural network based epileptic detection using time-domain and frequency-domain features,” *Journal of Medical Systems*, vol. 29, no. 6, pp. 647–660, 2005.

[105] Y. Li, X.-D. Wang, M.-L. Luo, K. Li, X.-F. Yang, and Q. Guo, “Epileptic seizure classification of eegs using time–frequency analysis based multiscale radial basis functions,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 386–397, 2017.

[106] T. Q. Le, S. T. S. Bukkapatnam, B. A. Benjamin, B. A. Wilkins, and R. Komanduri, “Topology and random-walk network representation of cardiac dynamics for

localization of myocardial infarction," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2325–2331, 2013.

[107] T. Q. Le, S. T. S. Bukkapatnam, and R. Komanduri, "Real-time lumped parameter modeling of cardiovascular dynamics using electrocardiogram signals: toward virtual cardiovascular instruments," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2350–2360, 2013.

[108] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 55–69, 1980.

[109] M. D'Alessandro, R. Esteller, G. Vachtsevanos, A. Hinson, J. Echauz, and B. Litt, "Epileptic seizure prediction using hybrid feature selection over multiple intracranial eeg electrode contacts: a report of four patients," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 5, pp. 603–615, 2003.

[110] S. Blanco, H. Garcia, R. Q. Quiroga, L. Romanelli, and O. Rosso, "Stationarity of the eeg series," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 4, pp. 395–399, 1995.

[111] Z. Wang and S. T. S. Bukkapatnam, "A dirichlet process gaussian state machine model for change detection in transient processes," *Technometrics*, vol. 60, no. 3, pp. 373–385, 2018.

[112] C. Cheng, A. Sa-Ngasoongsong, O. Beyca, T. Le, H. Yang, Z. Kong, and S. T. S. Bukkapatnam, "Time series forecasting for nonlinear and non-stationary processes: A review and comparative study," *IIE Transactions*, vol. 47, no. 10, pp. 1053–1071, 2015.

[113] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," *Clinical Neurophysiology*, vol. 116, no. 3, pp. 569–587, 2005.

[114] A. T. Tzallas, M. G. Tsipouras, and D. I. Fotiadis, "Epileptic seizure detection in eegs using time–frequency analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 703–710, 2009.

[115] A. J. Shackman, B. W. McMenamin, J. S. Maxwell, L. L. Greischar, and R. J. Davidson, "Identifying robust and sensitive frequency bands for interrogating neural oscillations," *Neuroimage*, vol. 51, no. 4, pp. 1319–1333, 2010.

[116] J. J. Howbert, E. E. Patterson, S. M. Stead, B. Brinkmann, V. Vasoli, D. Crepeau, C. H. Vite, B. Sturges, V. Ruedebusch, J. Mavoori, *et al.*, "Forecasting seizures in dogs with naturally occurring epilepsy," *PloS one*, vol. 9, no. 1, p. e81920, 2014.

[117] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[118] N. E. Huang and Z. Wu, "A review on hilbert-huang transform: Method and its applications to geophysical studies," *Reviews of Geophysics*, vol. 46, no. 2, 2008.

[119] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[120] W. J. Freeman and G. Vitiello, "Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics," *Physics of Life Reviews*, vol. 3, no. 2, pp. 93–118, 2006.

[121] S. Reddy, K. Afrin, M. Aguirre, S. Kancharla, T. Nakkina, B. Arnold, S. Manoharan, R. Doodipala, and S. T. S. Bukkapatnam, "Smart prognostic wearable for epileptic seizure alert," 2020.

[122] W. A. Chaovalitwongse, W. Suharitdamrong, C.-C. Liu, and M. L. Anderson, "Brain network analysis of seizure evolution," in *Annales Zoologici Fennici*, vol. 45, pp. 402–414, BioOne, 2008.

[123] R. Kuruba, X. Wu, and D. S. Reddy, "Benzodiazepine-refractory status epilepticus, neuroinflammation, and interneuron neurodegeneration after acute organophosphate intoxication," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1864, no. 9, pp. 2845–2858, 2018.

[124] National Institutes of Health, *Guide for the care and use of laboratory animals.* Washington, D.C.: National Academies, 1985.

[125] D. S. Reddy and R. Kuruba, "Experimental models of status epilepticus and neuronal injury for evaluation of therapeutic interventions," *International Journal of Molecular Sciences*, vol. 14, no. 9, pp. 18284–18318, 2013.

[126] X. Wu, R. Kuruba, and D. S. Reddy, "Midazolam-resistant seizures and brain injury after acute intoxication of diisopropylfluorophosphate, an organophosphate pesticide and surrogate for nerve agents," *Journal of Pharmacology and Experimental Therapeutics*, vol. 367, no. 2, pp. 302–321, 2018.

[127] B. Litt and J. Echauz, "Prediction of epileptic seizures," *The Lancet Neurology*, vol. 1, no. 1, pp. 22–30, 2002.

[128] J. Lagarias, "Euler's constant: Euler's work and modern developments," *Bulletin of the American Mathematical Society*, vol. 50, no. 4, pp. 527–628, 2013.

[129] M. Abramowitz, I. A. Stegun, *et al.*, *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* 1972.

[130] H. Ishwaran and U. B. Kogalur, "Consistency of random survival forests," *Statistics & Probability Letters*, vol. 80, no. 13, pp. 1056–1064, 2010.

# APPENDIX A

# INVESTIGATION OF THE EFFECT OF CLASS BALANCING ON SURVIVAL ANALYSIS

**Proposition 1** *For $m_2 << m_1$, let $\{m_1, m_2, \rho(t)\}$ and $\{m_1, m_2', \rho'(t)\}$ be the surviving and the mortality class sizes, and the Brier score (BS) before and after balancing, respectively, and let $d_0$ be the minimum number of unique death (mortality class) samples present in the censored leaf nodes of a survival tree to obtain non-zero hazard. Assuming an almost perfect split with $m_2 - d_0$ samples in the mortality node and $m_1 + d_0$ samples in the censoring node, $\rho'(t)$ can be approximated as:*

$$\rho'(t) = \rho(t)\left(\frac{m_1 + m_2}{m_1 + m_2'}\right)\left\{\frac{(m_2' - d_0)e^{-2\hat{H}'_M(t)} + d_0 e^{-2\hat{H}_C(t)} + m_1(1 - e^{-\hat{H}_C(t)})^2}{(m_2 - d_0)e^{-2\hat{H}_M(t)} + d_0 e^{-2\hat{H}_C(t)} + m_1(1 - e^{-\hat{H}_C(t)})^2}\right\}$$

**Proof**. Growing a survival tree proceeds with recursively splitting of the parent nodes into daughter nodes such that the survival difference between the daughter nodes is maximized. Further, the forest ensemble hazard for the individual is an average across all such leaf nodes in the forest. Consequently, determining the ensemble hazard and proving related result reduces to demonstrating them for a single node split. Therefore, for simplicity, we demonstrate our result for a single split and results thus derived can be adapted to a generalized survival tree construction and hence to the RSF. Nonetheless, we consider all possible case splits, i.e., (a) case 1: impure censored node, (b) case 2: impure mortality node, and (c) case 3: impure mortality and censored node.

Let $M$ and $C$ denote the mortality and censored/survival nodes respectively and the parent node has $m_1$ censored and $m_2$ mortality samples. For case 1, the leaf nodes then have the following conditions: (i) the censored nodes has $d_0$ mortality samples and (ii)

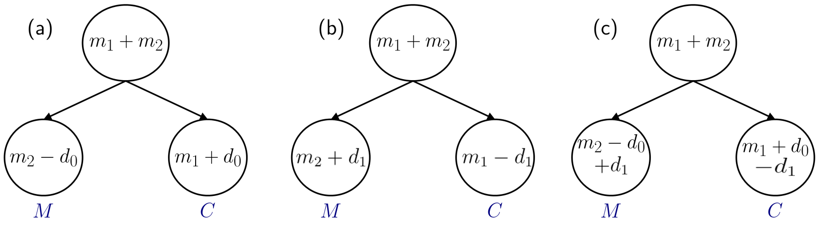

Figure A.1: Survival tree node split for three different cases (a) impure censored leaf node, (b) impure mortality leaf node, and (c) impure mortality and censored leaf nodes.

except for the $d_0$ misclassified mortality samples, both the nodes have a homogeneous population. Thus, the censored leaf node has $m_1 + d_0$ samples and the mortality leaf node has $m_2 - d_0$ samples. Note that, with this construction, a reasonable assumption is that $m_1 > d_0$, i.e., censored node has more censored samples than the mortality samples. Further, both leaf nodes have distinct event times. For case 2, the leaf nodes have the following conditions: (i) the mortality nodes has $d_1$ misclassified censored and $m_2$ mortality samples and (ii) except for the $d_1$ misclassified censored samples, both the nodes have a homogeneous population. Finally, the leaf nodes for case has the following conditions: (i) the censored node has $d_0$ misclassified mortality and a total of $m_1 - d_1 + d_0$ samples (ii) the mortality nodes has $d_1$ misclassified censored samples and a total $m_2 - d_0 + d_0$ samples.

The prediction error of an individual $i$ at any time $t$ can be defined in terms of the expected Brier score (refer to section 2.3.1.2) which is given as follows:

$$\rho_i(t) = E(\tilde{\mathcal{Y}}_i(t) - \hat{S}_i(t))^2$$

where, $\tilde{\mathcal{Y}}_i(t) = \mathbb{1}_{T_i > t}$ is the actual survival status of individual $i$ at time $t$ and $\hat{S}_i(t)$ is the predicted survival, which we know is equal to the survival estimates of the node they belong to (refer to section 2.2 of the main text). Further, the survival function estimated at a time $t$ for the mortality node, $\hat{S}_M(t)$ and the censored node, $\hat{S}_C(t)$ are given as follows:

$$\begin{aligned}\hat{S}_M(t) &= e^{-\hat{H}_M(t)}\\ \hat{S}_C(t) &= e^{-\hat{H}_C(t)}\end{aligned}$$

Now, given the case 1 node split, as defined above, BS score calculated for the unbalanced data, $m_1$ and $m_2$ can be represented as:

$$\begin{aligned}\rho(t) &= \frac{(m_2 - d_0)(0 - \hat{S}_M(t))^2 + d_0(0 - \hat{S}_C(t))^2 + m_1(1 - \hat{S}_C(t))^2}{m_1 + m_2}\\ &= \frac{(m_2 - d_0)e^{-2\hat{H}_M(t)} + d_0 e^{-2\hat{H}_C(t)} + m_1(1 - e^{-\hat{H}_C(t)})^2}{m_1 + m_2}\end{aligned}$$

For an imbalance with $m_2 << m_1$, let the mortality class size after balancing be $m_2'(m_2' \geq m_2)$ and $m_2' \approx m_1$ (with fixed $m_1$ and $d_0$), the balanced Brier Score, $\rho'(t)$ can then be given as:

$$\rho'(t) = \frac{(m_2' - d_0)e^{-\hat{H}'_M(t)} + d_0 e^{-2\hat{H}_C(t)} + m_1(1 - e^{-\hat{H}_C(t)})^2}{m_1 + m_2'}$$

Hence the ratio of $\rho'(t)$ and $\rho(t)$ can be represented as:

$$\frac{\rho'(t)}{\rho(t)} = \left(\frac{m_1 + m_2}{m_1 + m_2'}\right)\left\{\frac{(m_2' - d_0)e^{-2\hat{H}'_M(t)} + d_0 e^{-2\hat{H}_C(t)} + m_1(1 - e^{-\hat{H}_C(t)})^2}{(m_2 - d_0)e^{-2\hat{H}_M(t)} + d_0 e^{-2\hat{H}_C(t)} + m_1(1 - e^{-\hat{H}_C(t)})^2}\right\} \quad \text{(A.1)}$$

■

For $m_1 << m_2$ before balancing, let the censored class size after balancing be $m_1'(m_1' \geq m_1)$ and $m_1' \approx m_2$ (with fixed $m_2$ and $d_0$), the proposition 1 can be represented as:

$$\frac{\rho'(t)}{\rho(t)} = \left(\frac{m_1+m_2}{m_1'+m_2}\right)\left\{\frac{(m_2-d_0)e^{-2\hat{H}_M(t)}+d_0e^{-2\hat{H}_C'(t)}+m_1(1-e^{-\hat{H}_C'(t)})^2}{(m_2-d_0)e^{-2\hat{H}_M(t)}+d_0e^{-2\hat{H}_C(t)}+m_1(1-e^{-\hat{H}_C(t)})^2}\right\} \quad \text{(A.2)}$$

The expression for ratio of $\rho'(t)$ and $\rho(t)$ for case 2 and case 3 can be derived similarly. We skip these derivations for the brevity of the paper and proceed to an interesting result in Corollary 1.

**Corollary 1** Let $\{\rho(t), \rho'(t)\}$ be the Brier scores before and after balancing the class sizes, then $\rho'(t) < \rho(t)$.

**Proof**. Let $\{\rho(t), \rho'(t)\}$ be the Brier scores before and after balancing the class sizes, as expressed in Proposition 1. Since $m_2' > m_2$, we know that $\left(\frac{m_1+m_2}{m_1+m_2'}\right) < 1$. Now, let $f(m_2) = (m_2 - d_0)e^{-2\hat{H}_M(t)}$, showing that $f(m_2)$ is a decreasing function of $m_2$ would suffice to prove Corollary 1. In order to do so, we first derive the expression for $\hat{H}_M(t)$ for Case 1. Let $M$ and $C$ denote the mortality and censored/survival nodes, respectively. At the termination point, there are $\mathcal{L}(\mathcal{T}_b)$ terminal/leaf nodes in the tree, $\mathcal{T}_b$. Let, $t_{1,h} < t_{2,h} < ... < t_{N(h),h}$ be $N(h)$ ordered, unique event (death) times in the leaf node, $h \in \mathcal{L}(\mathcal{T}_b)$, then the CHF for individuals in this node is given using the Nelson-Aalen estimator as:

$$\hat{H}(t|\boldsymbol{x}_i) = \hat{H}_h(t) = \sum_{t_{l,h}\leq t} \frac{d_{l,h}}{Y_{l,h}}, \quad \text{if } \boldsymbol{x}_i \in h \quad \text{(A.3)}$$

In (A.3), $d_{l,h}$ and $Y_{l,h}$ represent, respectively, the number of deaths and the number of patients at risk in node $h$ at times $\{t_{l,h}\}_{1\leq l\leq N(h)}$. Since the construction survival tree is based on binary splits, $\boldsymbol{x}_i$ corresponding to each individual $i$ ends up in a unique leaf node of $\mathcal{L}(\mathcal{T}_b)$. Forest ensemble hazard for the individual is an average across all such leaf nodes in the forest. Further, in practice the trees are grown using bootstrap data which needs to be considered while estimating the ensemble hazard (for details of growing RSF and estimating ensemble hazard please see section 2.2 of the main text). For ease of calculation, we estimate the cumulative hazard of the nodes at their respective maximum event times. Let $t^*$ be the maximum event time at node $M$ then $\hat{H}_M(t^*)$ is given as:

$$\begin{aligned}\hat{H}_M(t^*) &= \sum_{t_{l,M} \leq t^*} \frac{d_{l,M}}{Y_{l,M}} \\ &= \frac{1}{(m_2 - d_0) - 1} + \frac{2}{(m_2 - d_0) - 2} + ... + \frac{(m_2 - d_0 - 1)}{1}\end{aligned}$$

Let $m_2 - d_0 = y$, then $\hat{H}_M(t^*)$ can be represented as:

$$\hat{H}_M(t^*) = \left( \frac{1}{y-1} + \frac{2}{y-2} + ... + \frac{y-1}{1} \right) \tag{A.4}$$

Alternately, (A.4) can be written in the form of a harmonic series as follows:

$$\begin{aligned}
z_1^{y-1} &= \frac{1}{y-1} + \frac{1}{y-2} + \frac{1}{y-3} \cdots + \frac{1}{2} + \frac{1}{1} \\
z_2^{y-2} &= \frac{1}{y-2} + \frac{1}{y-3} + \cdots + \frac{1}{2} + \frac{1}{1} \\
z_3^{y-3} &= \frac{1}{y-3} + \frac{1}{y-4} + \cdots + \frac{1}{2} + \frac{1}{1} \\
&\vdots \\
z_{y-2}^{2} &= \frac{1}{2} + \frac{1}{1} \\
z_{y-1}^{1} &= \frac{1}{1}
\end{aligned}$$

The sum of the $1^{st}$ series, $z_1^{y-1}$ with $(y-1)$ terms can be approximated using the following:

$$z_1^{y-1} = \sum_{n=1}^{y-1} \frac{1}{n} = \gamma + \psi_0((y-1)+1) = \gamma + \psi_0(y)$$

Where, $\gamma \approx 0.577$ is the Euler-Mascheroni constant [128] and $\psi_0(\cdot)$ is the diagmma function [129]. Similarly, $z_2^{y-2} = \gamma + \psi_0(y-1)$ and so forth. Hence, the hazard estimate for the mortality node can be given as:

$$\begin{aligned} \hat{H}_M(t^*) &= (y-1)\gamma + \sum_{n=1}^{y-1} \psi_0(n+1) \\ &= (m_2 - d_0 - 1)\gamma + \sum_{n=1}^{m_2-d_0-1} \psi_0(n+1) \end{aligned} \tag{A.5}$$

We perform first order differentiation by parts of $f(m_2)$ with respect to $m_2$ which results in:

$$\frac{df(m_2)}{dm_2} = e^{-2\hat{H}_M(t)}\big(1 - 2(m_2 - d_0)\frac{d\hat{H}_M(t)}{dm_2}\big) \tag{A.6}$$

Using $\hat{H}_M(t^*)$ from (A.5) in (A.6), the differentiation is given as follows:

$$\begin{aligned} \frac{df(m_2)}{dm_2} &= e^{-2(y-1)\gamma + \sum_{i=2}^{y}\psi_0(i)}\big(1 - 2y\frac{d(2(y-1)\gamma + \psi_0(2) + ...\psi_0(m_2 + d_0))}{dm_2}\big) \\ &= e^{-1.154(y-1) + \sum_{i=2}^{y}\psi_0(i)}(1 - 2y(0.577 + \psi_1(2) + ... + \psi_1(y))) \end{aligned}$$

Here, $y = (m_2 - d_0)$ and $\gamma = 0.577$. Clearly, with exponential and Trigamma function being positive [129], $df(m_2)/dm_2 < 0$. Now that we have established $f(m_2)$ is a decreasing function, for $m_2' > m_2$, the right hand side of (A.1) becomes less than 1 and hence $\rho'(t) < \rho(t)$. ■

This implies that the prediction of RSF improves after balancing. For $m_1 << m_2$, similar result holds. This is empirically corroborated from the results present in Table 2.2 and Figure 2.4 in the main text.

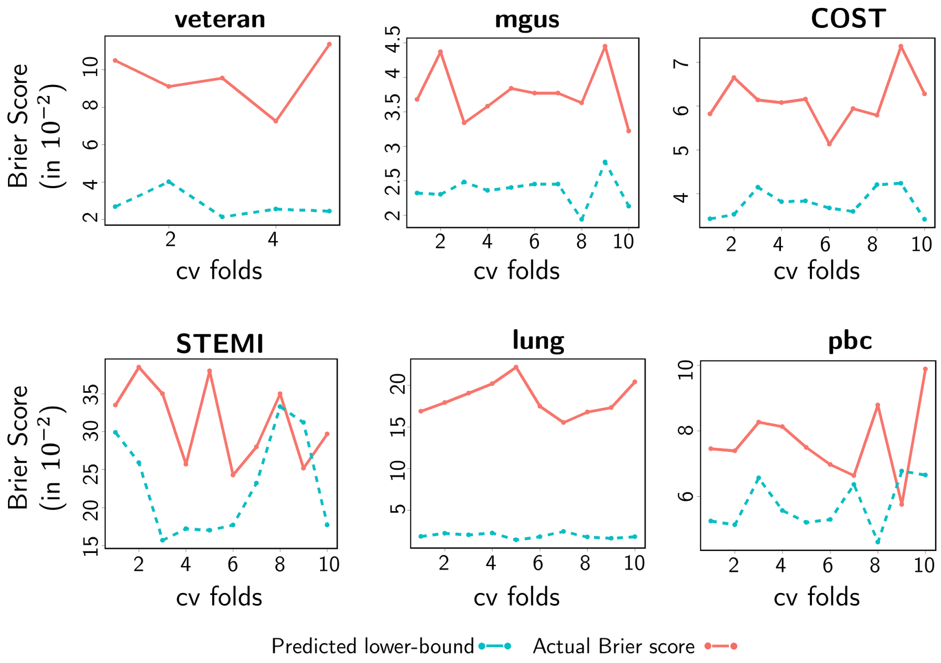

Figure A.2: Representation of the lower-bound on $\tilde{\rho}(t)$ for all the 6 datasets used in this paper.

Finally, from the propositions and corollary, we establish a lower bound on the BS after balancing. Figure A.2, we show the actual BS, $\tilde{\rho}(t)$ obtained after balancing and the predicted lower-bound for all the six datasets across 10 fold cv iterations. As evident from Figure A.2, the proposed lower-bound is a reasonable estimate.

**Remark 2** *Let $m_1$ and $m_2$ be the number of censored and mortality samples in $\mathbf{\Phi}$ and $\hat{H}_C(t)$, $\hat{H}_M(t)$ be the estimated cumulative hazard function for the censored and mortality nodes, respectively. Then for $m_2 << m_1$, $\hat{H}_M(t) < \hat{H}'_M(t)$ and for $m_1 << m_2$, $\hat{H}_C(t) > \hat{H}'_C(t)$, where $\hat{H}'_M(t)$ and $\hat{H}'_C(t)$ are the estimated cumulative hazard function for the censored and mortality nodes, respectively, after balancing.*

In this remark, we establish the relationship between the hazard functions before and after balancing and more importantly the result on improvement in the hazard estimates for both censored and mortality class after balancing. Let us first start with case 1. We already know the expression for hazard function, $\hat{H}_M(t^*)$ for the mortality node,

$M$ from (A.5). Now, the hazard function for the censoring node, $C$ estimated at or after the maximum event time, $t^{**}$, $\hat{H}_C(t^{**})$ can be represented as:

$$\hat{H}_C(t^{**}) = \frac{1}{(m_1+d_0)-1} + \frac{2}{(m_1+d_0)-2} + ... + \frac{d_0}{m_1}$$

Let $m_1 + d_0 = u$, then $\hat{H}_C(t^{**})$ can be represented as:

$$\hat{H}_C(t^{**}) \quad = \quad \left(\frac{1}{u-1} + \frac{2}{u-2} + ... + \frac{d_0}{u-d_0}\right) \tag{A.7}$$

Alternately, (A.7) can be written in the form of a harmonic series as follows:

$$\begin{aligned}
v_1^{d_0} &= \left\{\frac{1}{u-1} + \cdots + \frac{1}{u-d_0} + \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\right\} - \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\\
v_2^{d_0-1} &= \left\{\frac{1}{u-2} + \cdots + \frac{1}{u-d_0} + \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\right\} - \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\\
v_3^{d_0-2} &= \left\{\frac{1}{u-3} + \cdots + \frac{1}{u-d_0} + \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\right\} - \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\\
&\vdots\\
v_{d_0-1}^{2} &= \left\{\frac{1}{u-d_0+1} + \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\right\} - \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\\
v_{d_0}^{1} &= \left\{\frac{1}{u-d_0} + \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)\right\} - \left(\frac{1}{u-d_0-1} + \cdots + \frac{1}{1}\right)
\end{aligned}$$

The sum of the $1^{st}$ series, $v_1^{d_0}$ with $d_0$ terms can be approximated using the following:

$$\begin{aligned}
v_1^{d_0} &= \sum_{n=1}^{u-1}\frac{1}{n} - \sum_{n=1}^{u-d_0-1}\frac{1}{n}\\
&= \big(\gamma + \psi_0((u-1)+1)\big) - \big(\gamma + \psi_0((u-d_0-1)+1)\big) = \psi_0(u) - \psi_0(u-d_0)
\end{aligned}$$

Similarly, the sum of the last series, $v_{d_0}^1 = \psi_0(u-d_0+1) - \psi_0(u-d_0)$. Hence, the hazard estimate for the censored node can be given as:

$$
\begin{aligned}
\hat{H}_C(t^{**}) &= \sum_{n=u-d_0}^{u-1} \psi_0(n+1) - d_0\psi_0(u-d_0) \\
&= \sum_{n=m_1}^{m_1+d_0-1} \psi_0(n+1) - d_0\psi_0(m_1) \qquad \text{(A.8)}
\end{aligned}
$$

It can be shown that the hazard estimate or survival function of the leaf nodes and thus the RSF is consistent [130]. Let us consider a class imbalance with $m_2 << m_1$. Then, for case 1, $m_2 - d_0$ mortality samples have hazard $\hat{H}_M(t^*)$ and the remaining $d_0$ samples have a small hazard of the censored node, $\hat{H}_C(t^{**})$. With $m_2 << m_1$ and $\hat{H}_C(t^{**}) << \hat{H}_M(t^*)$, the overall estimate of the hazard for $m_2$ mortality samples is $(m_2 - d_0)\hat{H}_M(t^*) + d_0\hat{H}_C(t^{**}) < m_2\hat{H}_M(t^*)$. Now, with additional synthetic mortality samples and the new mortality class size $m_2'(m_2' > m_2)$, the hazard of the mortality class at $t^*$ becomes:

$$
\hat{H}_M'(t^*) = (m_2' - d_0 - 1)\gamma + \sum_{n=1}^{m_2'-d_0-1} \psi_0(n+1)
$$

Also, $\hat{H}_M'(t^*) - \hat{H}_M(t^*) = (m_2' - m_2)\gamma + (\sum_{n=1}^{m_2'-d_0-1} \psi_0(n+1) - \sum_{n=1}^{m_2-d_0-1} \psi_0(n+1))$. Since $m_2' > m_2$ and $\psi_0$ is an increasing function in $\mathbb{R}_+$ and the hazard estimate of the individuals in the mortality node has now improved. Further, the $d_0$ mortality samples present in censored node still have hazard $\hat{H}_C(t^{**})$. Nonetheless, the proportion, $d_0/m_2 > d_0/m_2'$, thus overall hazard of the individuals in the unbalanced mortality class is underestimated which is further worsened when the size $m_2$ itself is small. Conversely, when $m_1 << m_2$ with an additional $d_0$ mortality samples in the censored node, the unbalanced hazard of the censored node $\hat{H}_C(t^{**})$ is more than the hazard of the censored node with balanced samples, $\hat{H}_C'(t^{**})$. Without any mortality cases in the censored node, it has a flat survival curve or zero hazard. Since digamma is an increasing function and the first part of (A.8) also has $d_0$ terms $\sum_{n=m_1}^{m_1+d_0-1} \psi_0(n+1) > d_0\psi_0(m_1)$, i.e., $\hat{H}_C(t^{**}) > 0$, which implies an inflated hazard for the censored node. Further, after balancing, the hazard estimate of the censored node with $m_1'(m_1' > m_1)$ can now be represented as:

$$\hat{H}_C'(t^{**}) \quad = \quad \sum_{n=m_1'}^{m_1'+d_0-1} \psi_0(n+1) - d_0\psi_0(m_1') \tag{A.9}$$

$$\hat{H}_C'(t^{**}) - \hat{H}_C(t^{**}) = \big( \sum_{n=m_1'}^{m_1'+d_0-1} \psi_0(n+1) - \sum_{n=m_1}^{m_1+d_0-1} \psi_0(n+1)\big) \\ -d_0\big(\psi_0(m_1') - \psi_0(m_1)\big) \tag{A.10}$$

In (A.11), both parts have the same number of terms and since $m_1' > m_1$, $\psi_0(m_1') > \psi_0(m_1)$ and $\big(\sum_{n=m_1'}^{m_1'+d_0-1}, \psi_0(n+1) > \sum_{n=m_1}^{m_1+d_0-1} \psi_0(n+1)\big)$. Further, the derevative of digamma is a decreasing function (Trigamma [129]), therefore $\big(\psi_0(m_1') - \psi_0(m_1)\big) > \big(\psi_0(m_1'+1) - \psi_0(m_1+1)\big)$. Hence, $\hat{H}_C'(t^{**}) < \hat{H}_C(t^{**})$, i.e., the hazard for the censored node improves/decreases after balancing.

For case 2, $\hat{H}_M(t^*)$ is given as:

$$\hat{H}_M(t^*) \quad = \quad \frac{1}{(m_2+d_1)-1} + \frac{2}{(m_2+d_1)+2} + ... + \frac{m_2}{d_1} \tag{A.11}$$

Now, (A.11) can be simplified as:

Let $m_2 + d_1 = \mu$, then $\hat{H}_M(t^*)$ can be represented as:

$$\hat{H}_M(t^*) \quad = \quad \left(\frac{1}{\mu-1} + \frac{2}{\mu-2} + ... + \frac{m_2}{\mu-m_2}\right) \tag{A.12}$$

Alternately, (A.12) can be written in the form of a harmonic series as follows:

$$\begin{aligned}
\theta_1^{m_2} &= \left\{\frac{1}{\mu-1}+\cdots+\frac{1}{\mu-m_2}+\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\right\}-\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\\
\theta_2^{m_2-1} &= \left\{\frac{1}{\mu-2}+\cdots+\frac{1}{\mu-m_2}+\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\right\}-\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\\
\theta_3^{m_2-2} &= \left\{\frac{1}{\mu-3}+\cdots+\frac{1}{\mu-m_2}+\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\right\}-\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\\
&\vdots\\
\theta_{m_2-1}^{2} &= \left\{\frac{1}{\mu-m_2+1}+\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\right\}-\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\\
\theta_{m_2}^{1} &= \left\{\frac{1}{\mu-m_2}+\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)\right\}-\left(\frac{1}{\mu-m_2-1}+\cdots+\frac{1}{1}\right)
\end{aligned}$$

The sum of the $1^{st}$ series, $\theta_1^{m_2}$ with $m_2$ terms can be approximated using the following:

$$\begin{aligned}
\theta_1^{m_2} &= \sum_{n=1}^{\mu-1}\frac{1}{n}-\sum_{n=1}^{\mu-m_2-1}\frac{1}{n}\\
&= \big(\gamma+\psi_0((\mu-1)+1)\big)-\big(\gamma+\psi_0((\mu-m_2-1)+1)\big)=\psi_0(\mu)-\psi_0(\mu-m_2)
\end{aligned}$$

Similarly, the sum of the last series, $\theta_1^{m_2}=\psi_0(\mu-m_2+1)-\psi_0(\mu-m_2)$ and so forth. Hence, the hazard estimate for the mortality node can be given as:

$$\begin{aligned}
\hat{H}_M(t^*) &= \sum_{n=\mu-m_2}^{\mu-1}\psi_0(n+1)-d_0\psi_0(\mu-m_2)\\
&= \sum_{n=d_1}^{m_2+d_1-1}\psi_0(n+1)-m_2\psi_0(d_1) \qquad \text{(A.13)}
\end{aligned}$$

However, the censored leaf node has $m_1-d_1$ censored samples and no mortality samples. Without any mortality cases, the censored node will have a flat survival curve or zero hazard. After balancing, the hazard estimate of the mortality node, $\hat{H}'_M(t^*)$ with $m'_2$ samples and

$\hat{H}'_M(t^*) - \hat{H}_M(t^*)$ can now be represented as:

$$\hat{H}'_M(t^*) = \sum_{n=d_1}^{m'_2+d_1-1} \psi_0(n+1) - m'_2\psi_0(d_1) \tag{A.14}$$

$$\hat{H}'_M(t^*) - \hat{H}_M(t^*) = \sum_{n=m_2+d_1-1}^{m'_2+d_1-1} \psi_0(n+1) - d_1\big(\psi_0(m'_2) - \psi_0(m_2)\big) \tag{A.15}$$

Since, $m'_2 > m_2$ and digamma is an increasing function, we know that $\psi_0(m'_2) > \psi_0(m_2)$ and $\sum_{n=m_2+d_1-1}^{m'_2+d_1-1} \psi_0(n+1) > d_1(\psi_0(m'_2) - \psi_0(m_2))$ and thus $\hat{H}'_M(t^*) > \hat{H}_M(t^*)$.

Finally, for case 3, the estimated cumulative hazard of the mortality node is given as:

$$\hat{H}_M(t^*) = \frac{1}{(m_2 - d_0 - 1) + d_1} + \frac{2}{(m_2 - d_0 - 2) + d_1} + ... + \frac{(m_2 - d_0)}{d_1}$$

Again, after simplifying, $\hat{H}_M(t^*)$ can be written as:

$$\hat{H}_M(t^*) = \sum_{n=d_1}^{m_2-d_0+d_1-1} \psi_0(n+1) - (m_2 - d_0)\psi_0(d_1)$$

On the other hand, the estimated cumulative hazard of the censored node is given as:

$$\begin{aligned}\hat{H}_C(t^{**}) &= \frac{1}{m_1 - d_1 + (d_0 - 1)} + \frac{2}{m_1 - d_1 + (d_0 - 2)} + ... + \frac{d_0}{m_1 - d_1} \\ &= \sum_{n=m_1-d_1}^{m_1-d_1+d_0-1} \psi_0(n+1) - d_0(\psi_0(m_1 - d_1))\end{aligned}$$

Now when $m_2 << m_1$, after balancing $\hat{H}'_M(t^*) = \sum_{n=d_1}^{m'_2-d_0+d_1-1} \psi_0(n+1) - (m'_2 - d_0)\psi_0(d_1)$ and $\hat{H}'_M(t^*) - \hat{H}_M(t^*)$ is given as:

$$\hat{H}'_M(t^*) - \hat{H}_M(t^*) = \sum_{n=m_2-d_0+d_1-1}^{m'_2-d_0+d_1-1} \psi_0(n+1) - \big((m'_2 - m_2)\psi_0(d_1)\big) \tag{A.16}$$

Both terms in (A.16) have the same number of terms. Further, it is reasonable to assume that $m_2 - d_0 > d_0$, i.e., the number of mortality samples in mortality node is greater than the number of mortality samples in the censored node, then it follows that $(m_2 - d_0 + d_1) > d_1$.

Since $\psi_0$ is an increasing function, $\hat{H}'_M(t^*) > \hat{H}_M(t^*)$. Conversely, for $m_1 << m_2$ after balancing $\hat{H}'_C(t^{**}) = \sum_{n=m_1-d_1}^{m_1-d_1+d_0-1} \psi_0(n+1) - d_0(\psi_0(m_1-d_1))$ and $\hat{H}'_C(t^{**}) - \hat{H}_C(t^{**})$ is given as:

$$\hat{H}'_C(t^*) - \hat{H}_C(t^*) = \left( \sum_{n=m'_1-d_1}^{m'_1-d_1+d_0-1} \psi_0(n+1) - \sum_{n=m_1-d_1}^{m_1-d_1+d_0-1} \psi_0(n+1) \right) - d_0 \left( \psi_0(m'_1-d_1) - \psi_0(m_1-d_1) \right) \quad (A.17)$$

In (A.17), $\big( \sum_{n=m'_1-d_1}^{m'_1-d_1+d_0-1} \psi_0(n+1) - \sum_{n=m_1-d_1}^{m_1-d_1+d_0-1} \psi_0(n+1) \big)$ and $d_0 \big( \psi_0(m'_1-d_1) - \psi_0(m_1-d_1) \big)$ have the same number of terms, i.e., $d_0$. Again, using the same reasoning as in case 1, $\big( \psi_0(m'_1-d_1) - \psi_0(m_1-d_1) \big) > \big( \psi_0(m_1-d_1+1) - \psi_0(m_1-d_1+1) \big)$. Thus, $\hat{H}'_C(t^*) < \hat{H}_C(t^*)$.

These improvements in the hazard of the mortality and the censored node is shown empirically in Figure 2.5 of the main text.