Novel Deep Learning Approaches for Semi-Competing Risk Prediction

by

Stephen Salerno Jr.

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Yi Li, Chair
Professor Moulinath Banerjee
Associate Professor Peisong Han
Professor Richard A. Hirth
Professor Jian Kang

Stephen Salerno Jr.
salernos@umich.edu
ORCID iD: 0000-0003-2763-0494

DEDICATION

This thesis is dedicated to my *mentors*, my *friends*, my *family*, and my *partner* who have created a community of support throughout my academic journey. Your guidance, encouragement, and love have been instrumental in my growth and helping me on my way to attaining my PhD.

To my mentors, thank you for your wisdom, expertise, and patience. Your guidance has been invaluable in expanding my knowledge and honing my skills. Thank you for believing in me, pushing me to reach new heights, and encouraging me to pursue excellence in my research. I am grateful for the countless hours you have dedicated to shaping my intellect and fostering my personal and professional development.

To my friends, thank you for the joy, laughter, and comfort during the highs and the lows. Thank you for the late nights, the early mornings, and the endless reasons to celebrate. I am grateful to be surrounded by people who I respect and admire deeply. Your friendship has made this journey unforgettable.

To my family, thank you for your unconditional love, unwavering encouragement, and unrelenting faith in me. Your hard work and sacrifices, both big and small, have afforded me these opportunities for education that you never had for yourselves. Your constant presence, even from afar, has provided the motivation I needed to persevere. I am forever grateful for the work ethic and the values you have instilled in me.

To my partner, thank you for being my rock, my confidant, and my biggest supporter. Your love and understanding have sustained me during the most challenging times. Your belief in me, even during moments of self-doubt, has been a source of strength, and your belief in my dreams gave me the courage to pursue them wholeheartedly. You have been my sounding board, my cheerleader, and my biggest source of inspiration. This thesis closes one important chapter of our lives and starts many more to come.

ACKNOWLEDGEMENTS

I would like to thank Dr. Li's long-term collaborators, Dr. David C. Christiani and Dr. Xinan Wang for partnering with us on these projects, for providing the Boston Lung Cancer Study data, and for many helpful discussions on our results. I would also like to thank Jui Kothari for her support of our data needs. I thank Dr. Ingrid van Keilegom and a reviewer for many helpful suggestions that significantly improved the quality of the first paper contained in this thesis as we prepared it for publication and to Dr. Jeffrey Morris and two reviewers as we prepare the second paper.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

In the era of precision medicine, time-to-event outcomes such as time to death or disease progression are routinely collected, along with high-throughput covariates which defy classical survival regression models. Given challenges with high-dimensional survival data, recent emphasis has been placed on developing novel deep learning approaches for survival estimation and prognostication. However, many survival processes in real applications involve multiple competing events. Semi-competing risks, a variant of competing risk problems, have commonly been encountered in clinical studies. In this dissertation, we propose a series of deep learning approaches in this setting of semi-competing risks. Our motivation comes from the Boston Lung Cancer Survival Cohort study, a large cancer epidemiology cohort investigating the complex mechanisms of lung cancer.

In Chapter II, we first propose a novel, multi-task deep neural network for semi-competing risks based on the illness-death model, a compartment-type model for the rates at which individuals transition between disease states. We develop our objective function based on the hazards of experiencing a disease progression or death from being event-free (e.g., from time of diagnosis) and the hazards of death following progression. Our deep learning model consists of three risk-specific sub-networks, respectively corresponding to the three possible state transitions, and a finite set of trainable parameters for specifying the baseline hazards and the degree of dependence among the three transition processes. We further introduce a novel framework for evaluating predictive performance in this setting by extending the widely used Brier score for censored univariate time-to-event data to the bivariate survival function.

In Chapter III we further extend this method to allow the nonparametric estimation of our transition-specific baseline hazard functions. We propose a hybrid approach to deep learning via our so-called neural expectation-maximization (NEM) algorithm. By viewing the subject-specific frailty as a missing variable, the algorithm iterates between three steps. In the E-step, we update the conditional expectation of the frailties, given the data and current values for the model parameters. In the M-step, we estimate the jump sizes for the piecewise-constant baseline hazards, then fixing these quantities, update our estimates of the log risk functions and frailty variance as outputs of our neural network architectures in the N-step.

While mortality is often the primary endpoint for studying the effect of a particular treatment or exposure, non-fatal events impact illness trajectories and treatment decisions related to disease

management. The integration of causal inference into machine learning approaches has shown great promise for estimating the causal effects of treatments on survival outcomes, however, little work has been done in settings where a non-fatal event is potentially 'truncated by death.' In Chapter IV, we propose a deep learning approach for estimating the causal effect of a given treatment on a non-fatal outcome. We estimate the marginal survival function for the non-fatal event based on an Archimedean copula model and use a jackknife pseudo-value approach to circumvent the need for a complex loss function, whereby we estimate pseudo-survival probabilities at fixed time points as target values. We relate our pseudo-outcomes to our causal variable of interest and additional confounders in a deep neural network S-learner. Throughout, we provide a series of numerical studies to evaluate our proposed approach and apply our method to the Boston Lung Cancer Study. We conclude with some discussion on our current work and areas of future research.

CHAPTER 1

Introduction

1.1 Background and Motivation

Survival analysis is an area of statistics where the random variate is *survival time* or the time until the occurrence of a specific event, which represents a qualitative change or the transition from one discrete state to another (e.g., alive to deceased). The most often studied event in biomedicine is death, though events of interest in fields ranging from sociology to industry, to engineering, to finance, to astronomy are widely encountered. The goals of survival analysis are to describe the probability of an event occurring by some time, to detect associations between risk factors and events, or to predict survival times based on informative characteristics. What distinguishes survival outcomes from other outcomes is the presence of *censoring*, meaning that the event of interest may not be observed for all subjects; subjects whose event times are not observed are said to be *censored*. In practice, the fraction of event times that are censored in a study population can be substantial, prohibiting the direct use of standard regression methods. Estimation methods in survival analysis are built around extracting information from all subjects, censored or not.

In the era of precision medicine, survival outcomes with high-throughput covariates or predictors are routinely collected. These *high-dimensional data* (i.e., with the number of predictors exceeding the number of observations) challenge classical survival regression models, which are either infeasible to fit or likely to incur low predictability due to over-fitting. Recent emphasis has been placed on developing novel machine learning approaches for survival prognostication.

However, many survival processes in real applications involve multiple competing events. Semi-competing risk problems, a variant of competing risk problems, have commonly been encountered in clinical studies. By semi-competing, we mean that the occurrence of one event, i.e., a non-terminal event, is subject to the occurrence of another, terminal event, but not vice versa. As the non-terminal event (e.g., cancer progression) is often a strong precursor to the terminal event (e.g., death), semi-competing events are often related and, hence, the terminal event may informatively censor the non-terminal event. To overcome such informative censoring, researchers either consider

only univariate outcomes such as overall survival, or composite outcomes such as progression-free survival, that is, time to progression or death, whichever comes first. What is lacking in these approaches is how to model a predictor's potentially different roles in disease progress and death, and they ignore crucial information about the *sojourn time* between progression and death. Even in settings where the non-terminal and terminal event times are only modestly correlated, failing to acknowledge this *sojourn time* may lead to biased predictions. In this thesis, we propose a series of deep learning approaches for survival prediction in this setting of semi-competing risks. Our motivation comes from the Boston Lung Cancer Survival Cohort study, one of the largest cancer epidemiology cohorts investigating the complex mechanisms of lung cancer.

The rest of this chapter is organized as follows. In Section 1.2, we provide a brief overview of some key concepts and notation in survival analysis and introduce the necessary perquisites on which much of the subsequent literature is built. In Section 1.3, we turn to machine learning for survival prediction. We first discuss the application of common machine learning concepts in these settings, such as a support vector machines, recursive partitioning and survival trees, and ensemble learners such as random survival forests. We then focus on a review artificial neural networks and extend this notion to survival prediction. In Section 1.4, we conclude with a review of existing deep learning procedures for competing risk analysis and motivate the body of this work with the introduction of *semi-competing* risks.

1.2 Notation

Consider a study of n subjects. The outcome variable is the time to the event of interest, such as death or cancer progression. Events in other contexts can be bankruptcy, COVID-19 infection, graduation, missing a mortgage payment, etc. A *time zero* also needs to be set carefully, to have a practical interpretation when helping to address specific scientific questions. For instance, some common choices of time zero in medical studies include date of birth, time of diagnosis, date of randomization in a clinical trial, or first date receiving a treatment. A unique aspect of survival analysis is that the event may go unobserved for some individuals. In particular, *right censoring* occurs when a subject's follow-up ends before the event can be observed (**Figure 1.1**). Though other types of censoring exist, we focus on right censoring, which happens most often in practice.

We denote the ith subject's survival and censoring times respectively by T_i and C_i ($i = 1, \ldots, n$), which are non-negative random variates. For the ith subject, we observe $\boldsymbol{X}_i$, a p-vector of covariates, $Y_i = \min(T_i, C_i)$, and the event indicator $\delta_i = \mathbb{I}(T_i \leq C_i)$, where $\mathbb{I}(\cdot)$ is an indicator function. We assume that subjects are independent from each other, and that $T_i \perp C_i$, given $\boldsymbol{X}_i$. Often, the goal of survival analysis is to associate $\boldsymbol{X}_i$ with the distribution of T_i, and, in particular, model the conditional hazard function given $\boldsymbol{X}_i$, i.e.,

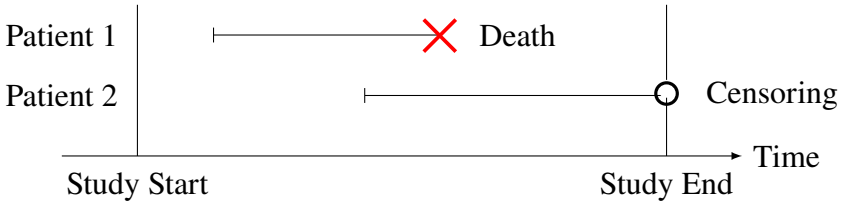


Figure 1.1: Schematic of observations for two example patients, with different entry times, over the course of a study. The event of interest, death, is observed for Patient 1, whereas Patient 2 is censored, as the patient is still alive at the end of the study.

$$\lambda(t|\boldsymbol{X}_i) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T_i < t + \Delta | T_i \geq t, \boldsymbol{X}_i), \tag{1.1}$$

which measures the instantaneous failure rate at a given time among those who are alive and whose risk factors are characterized by $\boldsymbol{X}_i$. Throughout this review, for simplicity, we assume that $\boldsymbol{X}_i$ is time-invariant, though in many circumstances extensions to time-dependent covariates are possible.

1.3 Machine Learning Techniques for Survival Prediction

Significant work has gone into the development of machine learning algorithms that can accommodate survival data. These non-parametric learning approaches can handle non-linear relationships or higher-order interaction that would otherwise be costly in classical methods, and can improve accuracy in prediction for survival outcomes.

1.3.1 Support Vector Machines

Support vector machines (SVMs) fall under the *supervised learning* family [139, 100] and seek to find a hyperplane that provides maximal separation between groups. Specifically, consider a binary outcome $Y_i \in \{-1, 1\}$ for each individual i with a corresponding p-dimensional covariate vector, $\boldsymbol{X}_i$. The goal of SVM is to identify a hyperplane, $H(\psi, a) = \{\mathbf{v} \in \mathbb{R}^p | \langle \psi, \mathbf{v} \rangle + a = 0\}$, separating these two groups so that the margin, $2/||\psi||$, can be maximized, where $\psi \in \mathbb{R}^p$ is the slope vector, and $\langle \cdot, \cdot \rangle$ denotes the inner product. Often, the two classes may not be separable in the original feature space within $\mathbb{R}^p$, and we use $\mathbf{F}(\cdot)$ to map the original predictors to a higher dimensional space where the outcomes can be distinguished, in which case, the hyperplane is $H(\psi, a) = \{\mathbf{v} \in \mathbb{R}^p | \langle \psi, \mathbf{F}(\mathbf{v}) \rangle + a = 0\}$ and, with slight overuse of notation, the dimension of ψ is the same as that of $\mathbf{F}(\mathbf{v})$. In practice, $\mathbf{F}(\cdot)$ does not have to be obtained explicitly and $\langle \psi, \mathbf{F}(\mathbf{v}) \rangle$ can be calculated by using a reproducing kernel [141]. We further introduce a slack variable,

$\xi_i = [1 - Y_i\{\langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle + a\}]_+$, to dictate the degree to which the ith data point is misclassified.

SVMs have been extended to model continuous time-to-event data, which are prone to censoring, by predicting the survival time to be $\langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle + a$. Van Belle et al. (2007) formulated the *survival SVM* based on the rank concordance between the prediction and observed survival time, Y_i, among comparable individuals in the presence of censoring. Specifically, they introduced a *comparability indicator*, $v_{ij} = \delta_i \mathbb{I}(Y_i < Y_j)$, such that the ordering of the observed survival times for subjects i, j can only be determined when $v_{ij} = 1$ [137]. For a comparable pair with $v_{ij} = 1$, a concordance in rank is reached if and only if $\langle \psi, \mathbf{F}(\boldsymbol{X}_j)\rangle - \langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle > 0$. Allowing varying degrees of pairwise slacks, i.e., when $\langle \psi, \mathbf{F}(\boldsymbol{X}_j)\rangle - \langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle \leq 0$ with $v_{ij} = 1$, across comparable pairs, Van Belle et al. proposed to solve

$$\begin{aligned} &\min_{\psi,\xi} \frac{1}{2}\|\psi\|^2 + \gamma \sum_{(i,j):Y_i<Y_j} v_{ij}\xi_{ij} \\ &\text{subject to } \left\langle \psi, \mathbf{F}\left(\boldsymbol{X}_j\right)\right\rangle - \langle \psi, \mathbf{F}\left(\boldsymbol{X}_i\right)\rangle \geq -\xi_{ij}, \\ &\text{and } \xi_{ij} \geq 0, i, j = 1, \ldots, n, \end{aligned}$$

where ξ_{ij}'s are pair-specific slacks, whose summation is to be minimized, and $\gamma > 0$ is a regularization parameter controlling the maximal margin and misclassification penalties. This formulation can be shown to maximize the Harrell rank-based concordance index (C-index) [61]. Hence, it is termed the *rank-based* SVM approach for survival data and does not estimate the "intercept," a. An alternative *regression* approach [121, 123] aimed to find a prediction, $\langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle + a$, for continuous survival times, by identifying a hyperplane that best fit the data that are subject to censoring, i.e.,

$$\begin{aligned} &\min_{\psi,a,\xi,,\xi^*} \frac{1}{2}\|\psi\|^2 + \gamma \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right) \\ &\text{subject to } Y_i - \langle \psi, \mathbf{F}\left(\boldsymbol{X}_i\right)\rangle - a \leq \xi_i, \\ &\qquad \delta_i \left(\langle \psi, \mathbf{F}\left(\boldsymbol{X}_i\right)\rangle + a - Y_i\right) \leq \xi_i^*, \\ &\text{and } \xi_i, \xi_i^* \geq 0. \end{aligned}$$

With censoring indicators incorporated into the constraints, the formulation utilizes available information from both censored and non-censored observations. To make full use of the strengths of both approaches, [138] and [109] further proposed *hybrid* approaches, combining the penalties imposed by both methods.

1.3.2 Tree-Based Methods

While SVMs are adept at estimating non-linear relationships, they do not scale well for large datasets and often under-perform when the outcomes are noisy. Also there may be no clear in-

terpretations for classifying data points above or below the estimated hyperplane [124]. Decision trees are an alternative for classifying patients that provide an intuitive interpretation of the hierarchical relationships between predictors. Broadly, classification and regression trees (CART) is an umbrella term for a set of *recursive partitioning* algorithms, which predict the group membership (classification) or target value (regression) for an observation based on a set of binary decision rules. Gordon and Olshen (1985) first presented survival trees, and Ciampi et al. (1985, 1986) solidified the notion and established splitting criteria based on the log-rank and likelihood ratio test statistics, respectively, gaining predictive accuracy and interpretability [55, 30, 29]. A recursive partitioning algorithm for generating a survival tree is given as follows.

1. Discretize each covariate to be a binary variable (categorical variables with m levels are expressed as $m - 1$ dummy variables).

2. For every binary covariate, X_j, $j = 1, \ldots, p$, compute the log-rank statistic to test the difference between the survival curves for the two groups defined by X_j.

3. Choose the covariate, X_{j^*}, with the largest significant test statistic and partition the full sample (i.e., the *root node*) into two groups (*child nodes*) based on X_{j^*}.

4. Repeat steps 2-3 for each subset (*child node*) until reaching the *terminal nodes*, that is, no covariates produce a significant test statistic and there are enough events (exceeding a prespecified number) in each *terminal node*.

The resulting terminal nodes split the original sample into distinct groups, who are deemed more homogeneous within each group, and will output survival estimates via Kaplan-Meier estimation in each group. Further variations in splitting are based on metrics that accommodate censored data and by either minimizing within-node homogeneity or maximizing between-node heterogeneity. For example, these metrics can be Martingale residuals [133] or deviance residuals [87]. With an established splitting criterion, to select a final tree, either a full survival tree is 'grown' and 'pruned' or a stopping rule is applied in backward or forward selection [17].

1.3.3 Ensemble Learners

While survival trees provide a fast and intuitive means of studying hierarchical relationships of predictors with outcomes, they are prone to over-fitting and high variability [67, 126]. Ensemble learners overcome instability issues with techniques such as *bagging*, *boosting*, and *random forests*.

1.3.3.1 Bagging

Bootstrap aggregation or *bagging* refers to a means of training an ensemble learner by resampling the data with replacement, training weak learners (e.g., individual survival trees) in parallel, and combining these results over the multiple *bootstrapped* samples [18]. It has three steps.

1. **Bootstrapping**: Resample from the original data of size *n with replacement* to form a new sample also of size n, and obtain 'B' such samples.

2. **Parallel Training**: With each bootstrap sample, $b = 1, \ldots, B$, independently train the weak learners in parallel.

3. **Aggregation**: Combine the B individual predictions by averaging over them or by taking a majority vote.

Bagging for survival trees was first proposed by [64]; in contrast to bagging for classification trees, aggregation is done by averaging survival predictions, rather than a 'majority vote.' Each survival tree is grown so that every terminal node has enough events, which are used to predict the survival function node-wise at each terminal node. Then, for any newcomer, the predictions are averaged over the individual trees to yield an ensemble prediction of their survival function.

1.3.3.2 Boosting

In a similar vein, *boosting* trains a series of weak learners with the goal of aggregating them into a better ensemble learner [23]. [63] proposed a gradient boosting algorithm for survival settings. Consider a mortality risk prediction based on covariates, $\boldsymbol{X}_i$. For an M-step gradient boosting algorithm, a prediction, $\mathcal{F}_m(\boldsymbol{X}_i)$, is made at each step, say $m = 1, \ldots, M$, based on a previous prediction, $\mathcal{F}_{m-1}(\boldsymbol{X}_i)$, and an additional weak learner $f_m(\boldsymbol{X}_i)$, which is the projection of the "residual error" of $\mathcal{F}_{m-1}(\boldsymbol{X}_i)$ to the space spanned by $\boldsymbol{X}_i$,

$$\mathcal{F}_m(\boldsymbol{X}_i) = \mathcal{F}_{m-1}(\boldsymbol{X}_i) + w_m f_m(\boldsymbol{X}_i),$$

where $0 < w_m \leq 1$ (e.g., $w_m = 0.1$) is the step size, the residual error refers to the gradient of the loss function, e.g., the negative log partial likelihood function in a survival setting, evaluated at $\mathcal{F}_{m-1}(\boldsymbol{X}_i)$, and the number of steps, M, can be viewed as a tuning parameter.

Boosting has two notable differences from bagging. First, boosting trains weak learners sequentially, updating the weights placed on learners iteratively, whereas in bagging individual weak learners such as survival trees are trained independently and in parallel, which are aggregated via majority voting or averaging. Second, boosting is applicable to settings where learners have low

variability and high bias, as the performance is improved by redistributing the weights. In contrast, bagging is often applied when individual learners exhibit high variability, but low bias, as it reduces variations arising from individual trees.

1.3.3.3 Random Forests

Yet another class of ensemble learners are random forests [19], which, like bagging, aggregate predictions from individual trees generated over bootstrap resampled datasets. However, differing from bagging, random forests randomly select a subset of features, say $p' < p$ features, when generating each tree and use them for the individual tree's growth. By doing so, random forests reduce correlations among individual trees, leading to gains in accuracy [19]. The choice of p' is problem-specific, which can also be viewed as a tuning parameter. In survival settings, [71] aggregated the survival predictions arising from each tree by averaging the predicted cumulative hazard functions into an ensemble prediction. Further notable developments include [72], which extended random survival forests to high dimensions by incorporating regularization, [73], which provided standard errors and confidence intervals for variable importance, and [127], which proposed censoring unbiased regression trees and ensembles.

1.3.4 Deep Learning and Artificial Neural Networks

Deep learning has emerged as a powerful tool for risk prediction. This work stems from artificial neural networks that tried to mirror how the human brain functions [115], wherein nodes (or *neurons*) are connected in a network as a weighted sum of inputs through a series of affine transformations and non-linear activations.

A fully-connected, feed-forward artificial neural network is made up of L layers, with k_l neurons in the lth layer ($l = 1, \ldots, L$) (**Figure 1.2**). With an input, network predictions are made based on an L-fold composite function, $f_L \circ f_{L-1} \circ \cdots \circ f_1(\cdot)$ with $(g \circ f)(\cdot) = g(f(\cdot))$. At the lh layer, $f_l(\cdot)$, is defined as

$$f_l(\mathbf{v}) = \sigma_l\left(\mathbf{W}_l\mathbf{v} + \mathbf{b}_l\right) \in \mathbb{R}^{k_l},$$

where $\mathbf{v}$ is a $k_{l-1} \times 1$ input vector fed from the $(l-1)$th layer, $\sigma_l(\cdot) : \mathbb{R}^{k_l} \to \mathbb{R}^{k_l}$ is an activation function, $\mathbf{W}_l$ is a $k_l \times k_{l-1}$ weight matrix, $\mathbf{b}_l$ is a $k_l \times 1$ bias vector, and the 0th layer is the input layer. Typical choices of $\sigma_l(\cdot)$ include the sigmoid function or the rectified linear unit activation function (ReLU), that is, $\sigma_l(\mathbf{b}) = \max(0, \mathbf{b})$, where $\mathbf{b} \in \mathbb{R}^{k_l}$ and $\max(0, \cdot)$ operates component-wise.

For survival prediction, several deep learning approaches have emerged, beginning with the seminal work of [42], which adopted a fully-connected, feed-forward neural network to extend the Cox model to nonlinear predictions. Other feed-forward neural networks [93, 21, 16, 41]

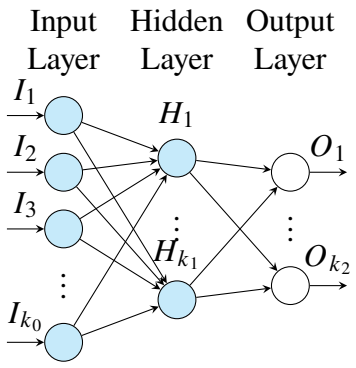


Figure 1.2: Diagram of a feed-forward, fully-connected two-layer artificial neural network, including the hidden (1st) and output (2nd) layer. The input (0-th) layer is not counted as a real neural network layer.

used the survival status as a training label, and output predicted survival probabilities. Further developments have been made in Bayesian networks [15, 94, 43], convolutional neural networks [148, 79, 80, 111], and recurrent neural networks [147]. From a different perspective, Zhao and Feng (2020) proposed transforming each subject's survival time into a series of jackknife pseudo conditional survival probabilities, thus circumventing the need for complex cost functions for censored survival data [154]. Given that the Kaplan-Meier estimator is approximately unbiased under independent censoring, for the ith subject, the pseudo-survival probability is computed by

$$\hat{S}_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{-i}(t),$$

where $\hat{S}(t)$ and $\hat{S}^{-i}(t)$ are the Kaplan-Meier (KM) estimates of $S(t)$ using all n subjects and excluding the ith subject, respectively. For subjects $i = 1, \cdots, n$, the $\hat{S}_i(t)$ are then used as the numeric response, similar to model fit to $\mathbb{I}(T_i > t)$, and the ANN minimizes the so-called *binary cross-entropy loss*, or simply, the mean of the squared differences between the pseudo- survival probabilities and the predicted survival probabilities from the neural network output.

1.4 Prediction for Competing and Semi-Competing Risks

Many survival processes in real applications involve multiple competing events. Risk prediction in these settings is an up-and-coming field with many potential developments. We focus on two common competing event settings, i.e., competing and semi-competing risks.

1.4.1 Competing Risks

In a competing risk setting, observing an event type, labeled by $c \in \{1, \ldots, K\}$, effectively eliminates the chance of observing other event types happening to the same individual [149]. For example, when studying the survival of patients with cancer, competing events can be cancer-related death ($c = 1$) or death by cardiac disease ($c = 2$) (**Figure 1.3**); an individual cannot die of cardiac disease once they have died of cancer, and vice versa. For characterizing the risk of competing events, there are two commonly used statistical metrics, namely, the *cause-specific* and the *subdistribution* hazard, which target different counterfactual scenarios. The former describes the risk under hypothetical elimination of competing events, while the latter is about the observable risk without elimination of any competing events [117].

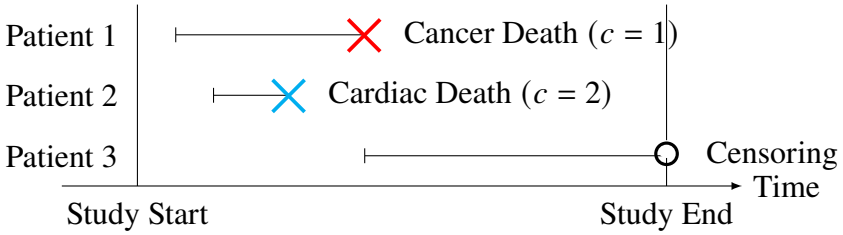


Figure 1.3: Schematic of observation times for three example patients with competing risks: cancer death (red cross) versus cardiac death (blue cross), with censoring denoted by an open circle.

Several authors [85, 84] have stated that the subdistribution hazard is useful for predicting the probability of having an event of a type of interest by a given time, termed the cumulative incidence function (CIF), which reflects an individual's actual risks and prognosis. In the following, we focus on the subdistribution hazard, which is derived from CIF, i.e., $F_c(t) = \Pr(T_i < t, C_i = c)$, where C_i marks the event type for subject i. Specifically, for each event type $c = 1, \ldots, K$, it is defined as

$$\lambda_c(t) = \lim_{\Delta \to 0} \frac{\Pr\left(t \leq T_i < t + \Delta, C_i = c \mid T_i \geq t \cup \{T_i < t \wedge C_i \neq c\}\right)}{\Delta} = \frac{dF_c(t)/dt}{1 - F_c(t)},$$

which denotes the instantaneous risk of failure from event type c among those who have not experienced this type of event. That is, the risk set at t includes those who are event free as well as those who have experienced a competing event (other than type c) by t. The subdistribution hazard model [45] links a subdistribution hazard function to covariates via

$$\lambda_c(t|\boldsymbol{X}_i) = \lambda_{0c}(t) \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}), \tag{1.2}$$

where $\lambda_{0c}(t)$ is the baseline subdistribution hazard function for event type c, and $\boldsymbol{\beta}$ specifies the effect of $\boldsymbol{X}_i$ on the probability of event c occurring over time. In fact, model (1.2) implies that

$1 - F_c(t|\boldsymbol{X}_i) = \{1 - F_{0c}(t)\}^{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})}$, where $F_c(t|\boldsymbol{X}_i)$ and $F_{0c}(t)$ are the CIF given $\boldsymbol{X}_i$ and the baseline CIF, respectively.

With high-dimensional predictors, several authors [81, 58, 5] proposed regularized subdistribution hazard models for variable selection, and [65] further performed inference using a one-step debiased LASSO estimator. For prediction, several deep learning works for competing risks have been proposed based on CIFs. DeepHit [89] developed a multi-task network to nonparametrically estimate $F_c(t|\boldsymbol{X}_i)$ for $c = 1, \ldots, K$. The network is trained to minimize a loss function, which is constructed based on the joint distribution of the first hitting time for competing events of each subject, while ensuring the concordance of estimates across subjects [61], that is, a patient who died at a given time should have a higher risk at that time than a patient who survived longer. Dynamic DeepHit [88] further incorporated longitudinal information for dynamic predictions. Other approaches have included DeepCompete [1], as well as a hierarchical, multi-state models [134].

1.4.2 Semi-Competing Risks

Semi-competing risk problems, a variant of competing risk problems, have commonly been encountered in clinical studies. By *semi-competing*, we mean that the occurrence of one event, i.e., a *non-terminal* event, is subject to the occurrence of another *terminal* event, but not vice versa (**Figure 1.4**). As the non-terminal event (e.g., cancer progression) is often a strong precursor to the terminal event (death), semi-competing events are often related and, hence, the terminal event may informatively censor the non-terminal event [74]. To overcome such informative censoring, researchers either consider only the terminal event (i.e., mortality) or a composite outcome such as *progression-free survival*, that is, time to progression or death, whichever comes first.

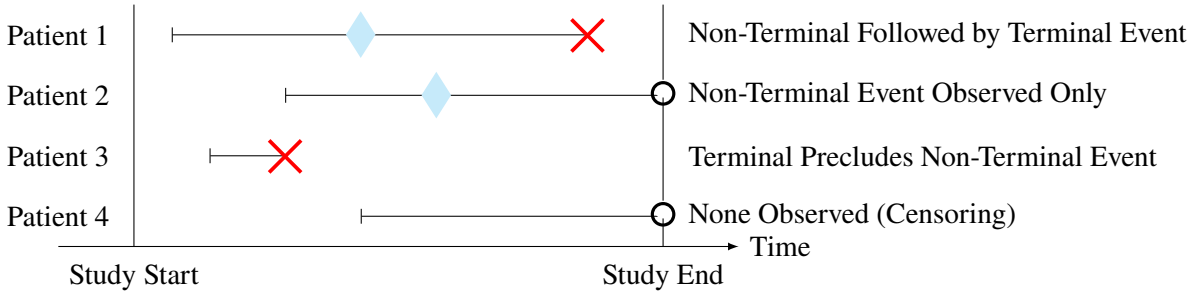


Figure 1.4: Schematic of four example patients with semi-competing risks: non-terminal event (blue diamond); terminal event (red cross); censoring (black circle).

What is lacking here is how to model a predictor's potentially different roles in disease progress and death, while utilizing the crucial information about the *sojourn* time between progression

and death. Even in settings where the non-terminal and terminal event times are only modestly correlated, failing to acknowledge this sojourn time may lead to incorrect inference or biased predictions [35]. In Chapter 2, we propose a novel, multi-task deep neural network for semi-competing risks based on the illness-death model, a compartment-type model for the rates at which individuals transition between disease states. We develop our objective function based on the hazards of experiencing a disease progression or death from being event-free (e.g., from time of diagnosis) and the hazards of death following progression. We further introduce a novel framework for evaluating predictive performance in this setting by extending the widely-used Brier score for censored univariate time-to-event data to the bivariate survival function. We then apply our method to the Boston Lung Cancer Study, where we investigate the impact of clinical and genetic predictors on disease progression and mortality.

In Chapter 3 we extend this method to allow for the flexible estimation of our transition-specific baseline hazard functions as well. We propose a hybrid approach to deep learning via our so-called neural expectation-maximization (NEM) algorithm. Under our neural EM approach, we update the conditional expectation of a subject-specific *frailty* in the E-step, estimate the jump sizes for piecewise-constant baseline hazards in the M-step, and update our estimates of the log risk functions and frailty variance as outputs of our neural network architectures in the N-step. As deep learning can recover non-linear risk scores, we test our method by simulating risk surfaces of varying complexity and revisit the Boston Lung Cancer Study.

While mortality is often the main focus of cancer studies, non-fatal events, such as disease progression, can vitally impact patient outcomes. For example, recurrence after curative treatment is a crucial endpoint in lung cancer, affecting available second-line treatments and personalized care. Estimating the true effect of interventions on disease recurrence is a key aspect of assessing cancer treatments. However, semi-competing risks complicate causal inference when death prevents disease recurrence. Existing approaches for estimating causal quantities in semi-competing survival functions rely on complex objective functions with strong assumptions and are challenging to estimate accurately. To address these challenges, in Chapter 4 we propose a deep learning approach for estimating the causal effect of treatment on non-fatal outcomes in the presence of dependent censoring and complex covariate relationships. Our three-stage approach involves estimating the marginal survival function using an Archimedean copula representation, and a jackknife pseudo-value approach that estimates pseudo-survival probabilities at fixed time points. These pseudo-survival probabilities serve as target values for developing causal estimators that are consistent and do not rely on assumptions like proportional hazards across all time points. In the final stage, we employ a deep neural network to link pseudo-outcomes, the causal variable, and additional confounders. This enables us to estimate survival average causal effects through direct standardization. We evaluate our approach through numerical studies and apply it to the Boston

Lung Cancer Study, specifically examining the effect of surgical tumor resection in early-stage non-small cell lung cancer patients. We conclude with remarks on future work and open areas of study in this exciting field.

CHAPTER 2

Deep Learning for Semi-Competing Risk Prediction

2.1 Background

Lung cancer remains one of the leading cause of cancer mortality worldwide, with a 5-year survival rate less than 20% [13, 114]. Prognostication for individuals with lung cancer is a complex task, often relying on the use of risk factors and health events spanning their entire life course [53, 86]. One challenge is that an individual's disease course involves non-terminal (e.g., disease progression) and terminal (e.g., death) events, which form *semi-competing* relationships. Further, prognosis varies greatly for patients with lung cancer, and accurate prediction of long-term events such as progression or mortality depends on several individualized risk factors including smoking status, genetic variants, and other comorbid conditions [22, 12, 49].

To facilitate prediction in clinically complex settings, machine learning techniques are becoming increasingly popular for studying the potential non-linear and higher-order interactions between large numbers of risk factors. Following developments in the prediction of time-to-event outcomes with neural networks [42, 94, 79, 111, 75, 60], deep learning has become a key area of focus for the development of risk prediction methods in survival analysis. Many deep learning approaches for time-to-event outcomes extend the Cox proportional hazards model [34] to nonlinear predictions or use a patient's survival status directly as a binary training label, predicting a patient's survival probability rather than their survival time. More recently, competing risk and multi-state models extend these methods to settings where multiple event types mutually censor one another [89, 88, 1, 134]. Such methods characterize the risk of one or more competing events by estimating either the cause-specific or subdistribution hazards of each event type, where the survival times are viewed as the first hitting times of the underlying stochastic processes.

While methods for competing risks improve greatly upon univariate approaches, they cannot accommodate prediction of the joint risk of two events or the study of outcome trajectories in the so-called *sojourn time* between two events. In fact, there is currently a lack of literature dealing with risk prediction for semi-competing outcomes. Most recently [113] proposed a penalized estimation

approach under semi-competing risks. In this chapter, we propose a novel, multi-task deep neural network for semi-competing risks based on the illness-death model, a compartment-type model for the rates at which individuals transition between disease states. We develop our objective function based on the hazards of experiencing a disease progression or death from being event-free (e.g., from time of diagnosis) and the hazards of death following progression. Our deep learning model consists of three risk-specific sub-networks, respectively corresponding to the three possible state transitions, and a finite set of trainable parameters for specifying the baseline hazards and the degree of dependence among the three transition processes.

This chapter is laid out as follows. In Section 2.2, we go over our notation and review the illness-death model, a compartment-type model for studying the hazards, or transition rates, between semi-competing events. In Section 2.3, we propose our deep learning approach for semi-competing risk prediction. In Section 2.4 we introduce a novel framework for evaluating predictive performance in this setting by extending the widely-used Brier score for censored univariate time-to-event data to the bivariate survival function. We then assess the predictive accuracy of our method in Section 2.5. In Section 2.6, we apply our method to analyze the BLCS cohort. We conclude with a discussion and directions for future work.

2.2 The Illness-Death Model

Consider two events, a non-terminal event and a terminal event. Let T_{i1} denote the time to the non-terminal event and T_{i2} denote the time to the terminal event for the ith observation in an analytic sample of n subjects. Central to the formulation of the semi-competing problem is the *illness-death model*, a compartment-type model for the rates at which individuals transition between event states [47, 131, 78, 10]. Within the framework of illness-death models, we stipulate a three compartment model for the rates at which individuals transition between an initial, event-free state (e.g., diagnosis), a non-terminal event state (e.g., progression), and a terminal event state (e.g., death). The hazard rates corresponding to the transitions from diagnosis to progression, $\lambda_1(t_1)$, diagnosis to death, $\lambda_2(t_2)$, and from progression to death, $\lambda_3(t_2 \mid t_1)$, are defined as

$$\lambda_1(t_1) = \lim_{\Delta \to 0} \Pr\left[T_1 \in [t_1, t_1 + \Delta) \mid T_1 \geq t_1, T_2 \geq t_1\right] / \Delta; \quad t_1 > 0 \tag{2.1}$$

$$\lambda_2(t_2) = \lim_{\Delta \to 0} \Pr\left[T_2 \in [t_2, t_2 + \Delta) \mid T_1 \geq t_2, T_2 \geq t_2\right] / \Delta; \quad t_2 > 0 \tag{2.2}$$

$$\lambda_3(t_2 \mid t_1) = \lim_{\Delta \to 0} \Pr\left[T_2 \in [t_2, t_2 + \Delta) \mid T_1 = t_1, T_2 \geq t_2\right] / \Delta; \quad t_2 > t_1 > 0, \tag{2.3}$$

where $\lambda_3(t_2 \mid t_1)$ depends on t_1 and t_2 via their difference, $t_2 - t_1$, i.e., the *sojourn time* since the

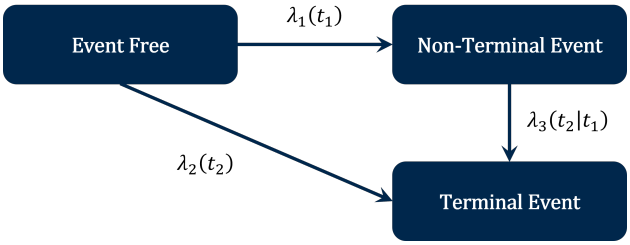


Figure 2.1: Graphical representation of the illness-death model with three states: the event free, or initial state, the non-terminal event state, and the terminal event state. Transition rates between states are characterized by $\lambda_1(t_1)$, $\lambda_2(t_2)$, and $\lambda_3(t_2 \mid t_1)$, respectively.

non-terminal event (Figure 2.1). This is a semi-Markov model with respect to $\lambda_3(t_2 \mid t_1)$ [145]. Alternatively, this would be Markov in time if $\lambda_3(t_2)$ only depended on t_2, the terminal event time.

In practice, both non-terminal and terminal events are subject to independent censoring. We focus only on the case of right censoring, whereby a subject may be lost to follow-up or the study ends before the event has occurred. We denote the censoring time for individual i by C_i. Our observed data are defined as

$$\mathcal{D} = \{(Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2});\ i = 1, \ldots, n\},$$

where $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = I(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = I(T_{i1} \leq Y_{i2})$, and $I(\cdot)$ denotes the indicator function. Our observable data take on probability only in the so-called *upper wedge* on which $Y_{i1} \leq Y_{i2}$ and arise from four potential cases: (1) the subject experiences both the non-terminal and the terminal event, (2) the subject experiences only the terminal event, (3) the subject experience only the non-terminal event, or (4) the subject experiences neither event prior to the end of follow up (Figure 2.2). We can model (2.1) - (2.3) in the context of our observed data as follows. We extend the Cox model [34] to our semi-competing risks setting [145, 59] by formulating each hazard function in terms of the baseline hazard for the transition of states, a shared frailty term, and a patient's covariates as

$$\lambda_1\left(t_1 \mid \gamma_i, \boldsymbol{x}_i\right) = \gamma_i \lambda_{01}\left(t_1\right) \exp\{h_1(\boldsymbol{X}_i)\}; \quad t_1 > 0 \tag{2.4}$$

$$\lambda_2\left(t_2 \mid \gamma_i, \boldsymbol{x}_i\right) = \gamma_i \lambda_{02}\left(t_2\right) \exp\{h_2(\boldsymbol{X}_i)\}; \quad t_2 > 0 \tag{2.5}$$

$$\lambda_3\left(t_2 \mid t_1, \gamma_i, \boldsymbol{x}_i\right) = \gamma_i \lambda_{03}\left(t_2 - t_1\right) \exp\{h_3(\boldsymbol{X}_i)\}; \quad t_2 > t_1 > 0, \tag{2.6}$$

where γ_i is a patient-specific random effect, or *frailty*, $\lambda_{01}\left(t_1\right)$, $\lambda_{02}\left(t_2\right)$, and $\lambda_{03}\left(t_2 - t_1\right)$ are the baseline hazard functions for the three state transitions, respectively, $\boldsymbol{X}_i$ is a p-vector of clinically relevant predictors such as patient socio-demographic status, medical history data, and comorbid

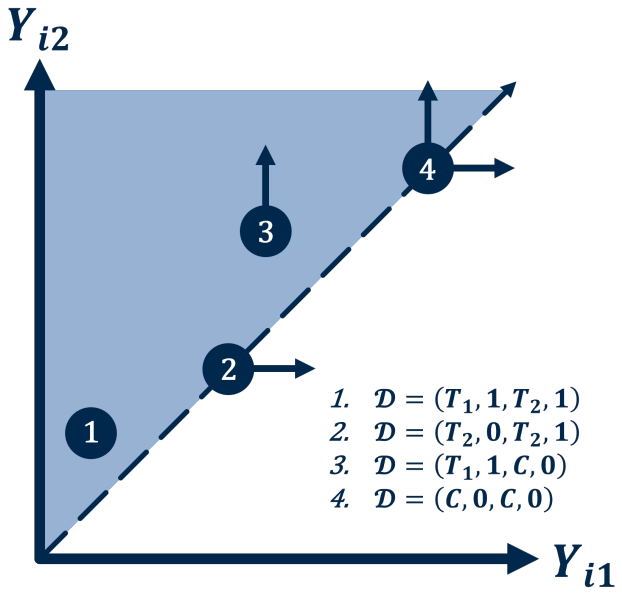


Figure 2.2: Graphical representation of the observable space for (T_1, T_2) with example observations: (1) both events are observed, (2) only the terminal event is observed, (3) only the non-terminal event is observed, and (4) neither event is observed. The arrows represent the direction of censoring, and $\mathcal{D} = (Y_1, \delta_1, Y_2, \delta_2)$ represents the data under each example observation.

conditions, and $h_g(\boldsymbol{X}_i); g \in \{1, 2, 3\}$ are log-risk functions which relate a patient's covariates to the hazard rates for each potential transition. As opposed to existing works, we do not parameterize the $h_g(\boldsymbol{X}_i)$. Instead, we estimate these functions non-parametrically as outputs from our proposed neural network architecture. The λ_{0g} functions can be taken to be Weibull functions or piecewise constant with jumps at the distinct observed event times. Including a shared frailty term in (2.4) - (2.6) induces a dependence structure between the multiple event times taken on a given subject. We assume $\gamma_i \overset{i.i.d}{\sim} \text{Gamma}(1/\theta, 1/\theta)$ (i.e., both shape and rate are $1/\theta$ so that the mean and variance are respectively 1 and θ). A larger value of θ reflects a stronger dependence. Further, model (2.6) stipulates that the hazard is a function of the *sojourn time*, a reasonable and common assumption [59, 90]. Given (2.4)-(2.6), and by integrating out the frailty term, we can derive the marginal likelihood based on n independent subjects as

$$\begin{aligned}\mathcal{L} = \prod_{i=1}^{n} &\{\lambda_{1i}(Y_{i1})\}^{\delta_{i1}} \{\lambda_{2i}(Y_{i1})\}^{(1-\delta_{i1})\delta_{i2}} \{\lambda_{3i}(Y_{i2} - Y_{i1})\}^{\delta_{i1}\delta_{i2}} \left(1 + \theta^{-1}\right)^{\delta_{i1}\delta_{i2}} \\ &\times \left[1 + \theta^{-1} \left\{\Lambda_{1i}\left(Y_{i1}\right) + \Lambda_{2i}\left(Y_{i1}\right) + \Lambda_{3i}\left(Y_{i2} - Y_{i1}\right)\right\}\right]^{-\theta - \delta_{i1} - \delta_{i2}},\end{aligned} \tag{2.7}$$

where $\lambda_{gi}(s) = \lambda_{0g}(s) \exp\left\{h_g(\boldsymbol{X}_i)\right\}$ and $\Lambda_{gi}(t) = \int_0^t \lambda_{gi}(s)ds$ for $g = 1, 2, 3$. See Appendix A for additional details. This serves as the objective function for our proposed algorithm.

2.3 Deep Learning for Semi-Competing Risks

We propose a multi-task deep neural network for semi-competing risks by using Equation (2.7) as the objective function with potentially high-dimensional covariates. Our neural network consists of three risk-specific sub-architectures, respectively corresponding to the three possible state transitions, and a finite set of trainable parameters for specifying the baseline hazards (i.e., the ϕ parameters in **Figure 2.3**) and the dependence among the three transition processes (i.e., θ in **Figure 2.3**). For example, if we specify Weibull baseline hazards in (2.4)-(2.7), then $\lambda_{0g}(s) = \phi_{g1}\phi_{g2}s^{\phi_{g2}-1}$ for $g = 1, 2, 3$. As opposed to the classical models, we opt for flexible, non-parametric estimation of $h_g(\cdot), g = 1, 2, 3$, to better capture potential non-linear dependencies of covariates on semi-competing events and to maximize the predictive accuracy.

In particular, we design three sub-architectures to estimate the h functions non-parametrically as outputs. Each sub-network is made up of L layers, with k_l neurons in the lth layer ($l = 1, \ldots, L$). Sub-network predictions are based on an L-fold composite function, $F_L(\cdot) = f_L \circ f_{L-1} \circ \cdots \circ f_1(\cdot)$, where $(g \circ f)(\cdot) = g(f(\cdot))$. Each layer-specific function, $f_l(\cdot)$, is defined as

$$f_l(x) = \sigma_l\left(\mathbf{W}_l x + \mathrm{b}_l\right) \in \mathbb{R}^{k_{l+1}},$$

where $\sigma_l : \mathbb{R}^{k_{l+1}} \to \mathbb{R}^{k_{l+1}}$ is an activation function, $\mathbf{W}_l$ is a $k_{l+1} \times k_l$ weight matrix, and b_L is a $k_{l+1} \times 1$ bias vector. For identifiability, we require $h_g(\mathbf{0}) = 0, g = 1, 2, 3$, where $\mathbf{0}$ is a $p \times 1$ vector of 0's. Each sub-network is a fully-connected feed-forward neural network with rectified linear unit activations (ReLU; $\sigma_l(x) = \max(0, x)$) and a linear activation in the final layer (**Figure 2.3**). The number of hidden layers, nodes per layer, dropout fraction, regularization rate, and learning rate can be optimized as hyperparameters over a grid of values based on predictive performance. We implement our approach using the deep learning library `TensorFlow` [8], with model building and fitting done using `Keras` [7]. Finite dimensional parameter training is done via the GradientTape API [4] for automatic differentiation.

2.4 Bivariate Brier Score

To assess the predictive performance of methods in a semi-competing risk setting, we propose a bivariate extension to the inverse probability of censoring weighting (IPCW)-approximated Brier Score [20]. Let $S_i(t) = \Pr(T_{i1} > t, T_{i2} > t)$ denote the disease-free survival function for individual i at a given, fixed time point t. Further, denote an estimate of $S_i(t)$ by $\pi_i(t)$, e.g., based on (2.4)-(2.6). If $S_i(t)$ were known, a Bivariate Brier Score would simply be the mean squared error

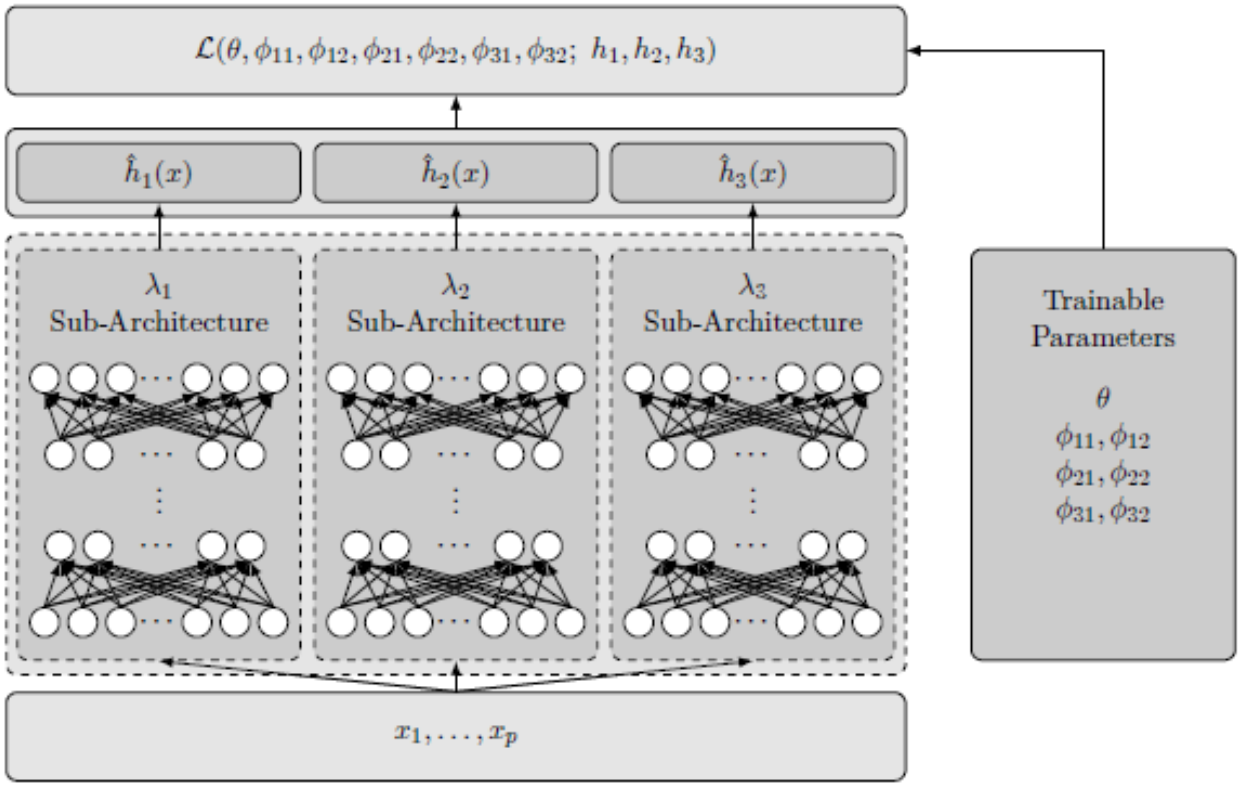


Figure 2.3: Architecture for the proposed semi-competing risk deep neural network.

$$MSE(t) = \frac{1}{n} \sum_{i=1}^{n} \left[S_i(t,t) - \pi_i(t,t) \right]^2 .$$

However, with unknown $S_i(t)$, we need to estimate it with the observed data, $\mathcal{D}$, and in the presence of censoring. In particular, we approximate the bivariate survival function by the indicator function $I(T_{i1} > t,\ T_{i2} > t)$, which is equal to one if both conditions are true and zero, otherwise. This provides an approximation to the true, unknown survival functions through step functions with jumps at the observed event times. Further, in practice, patients may be lost to follow up, or the observation period may end before the events are observed. In these situations, it is known that the event times, T_{i1} and T_{i2}, occur after some censoring time, C_i. In this setting, inverse probability of censoring weights (IPCW) are necessary to incorporate information loss due to censoring, as we must reweight the contributions of the individuals who do contribute information to the Bivariate Brier Score [56, 50]. Let $G_i(t) = \Pr(C_i > t) > 0$ be the survival function of the censoring distribution for the ith individual. We propose a Bivariate Brier Score for assessing $\pi_i(t)$ as

$$
\begin{aligned}
BBS_c(t) = & \frac{\pi_i(t,t)^2 \cdot I\left(Y_{i1} \leq t,\ \delta_{i1} = 1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i1})} \\
& + \frac{\pi_i(t,t)^2 \cdot I\left(Y_{i1} \leq t,\ Y_{i2} \leq t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i2})} \\
& + \frac{[1 - \pi_i(t,t)]^2 \cdot I\left(Y_{i1} > t,\ Y_{i2} > t\right\}}{G_i(t)}
\end{aligned} \tag{2.8}
$$

With $G_i(t)$ known, the expectation of the IPCW-approximated Bivariate Brier Score is equal to the mean squared error, $MSE(t)$, plus a constant that is free of $\pi_i(t)$. This additional term represents the irreducible error incurred by approximating $S_i(t)$ using data (see Appendix A). As $G_i(t)$ is often unknown, we can replace it by $\hat{G}(t)$, a consistent estimate based on the Kaplan-Meier method.

2.5 Simulation Studies

2.5.1 Bivariate Brier Score

We first conducted a series of numerical experiments to study the performance of the proposed Bivariate Brier Score. We generated 1,000 independent datasets of size $n = 1,000$ based on the illness-death model. Across all simulated datasets and simulation settings, we assumed Weibull baseline hazards with shape parameter equal to 1.5 and scale parameter equal to 0.2, and a population frailty variance of $\theta = 0.5$. We considered four simulation settings, varying whether or not the semi-competing outcomes depended on a single, uniform random covariate, and varying the administrative censoring rate at 0% and 50%. We calculate the integrated Bivariate Brier Score for 1-year survival over a sequence of 100 evenly spaced time points in each simulation and compared the results from the model fit to a calculation which utilized the true model parameters. This was done to compare the fitted results to those results which signified the degree of irreducible error in the Bivariate Brier Score for each setting. These results are given in Table 2.1. As shown the results from the model fit were consistent with those calculated using the true model parameters.

2.5.2 Deep Neural Network Approach

We then conducted simulations to illustrate the feasibility of the proposed model. We simulated the observed data, $\mathcal{D} = \{(Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2},\ \boldsymbol{X}_i);\ i = 1, \ldots, n\}$ in a fully factorial design by varying the sample size, frailty variance, log-risk function, and censoring rates, a total of 24 settings (Table 2.2). Specifically, we simulated the shared frailty, γ_i, from $Gamma(1/\theta, 1/\theta)$ with $\text{Var}(\gamma_i) = \theta$ taking values of 0.5 and 2.0, corresponding to varying degrees of dependence between event times.

Table 2.1: Mean (SD) integrated Bivariate Brier Score under various data generation settings.

Simulation Settings		1-Year iBBS	
Covariate Generated?	Censoring Generated?	True	Estimated
No	No	0.0187 (0.0068)	0.0199 (0.0073)
Yes	No	0.0181 (0.0067)	0.0205 (0.0077)
No	Yes	0.0206 (0.0067)	0.0219 (0.0072)
Yes	Yes	0.0195 (0.0066)	0.0221 (0.0075)

The baseline hazard functions, λ_{01}, λ_{02}, and λ_{03}, were taken to be Weibull with shape and scale parameters equal to 1. We simulate two standard Normal random covariates, $X_1, X_2 \sim N(0,1)$, which were taken to be predictive of the morbidity and mortality hazards through either a linear and non-linear log-risk function. We first examined a linear log-risk function

$$h_g(\boldsymbol{X}_i) = x_i^\top \boldsymbol{\beta}_g,$$

with $\boldsymbol{\beta}_g = [1,1]^\top$ for $g = 1,2,3$, so that the requirements for the classical model is satisfied, facilitating a fair comparison with existing methods. We then considered a non-linear function

$$h_g(\boldsymbol{X}_i) = \log(|\boldsymbol{X}_i|^\top \boldsymbol{\beta}_g + 1),$$

with $\boldsymbol{\beta}_g = [1,1]^\top$ for $g = 1,2,3$. Censoring times were generated from an exponential distributions to yield approximate censoring rates of 0%, 25% and 50%. We varied the number of patients as $1,000$ and $10,000$. For each parameter configuration, 50 datasets were independently generated.

We compared our method to a classical MLE approach, which directly maximizes the log-likelihood function under the assumption of a semi-Markov model with Weibull baseline hazard functions. This approach assumes that the risk functions are linear combinations of the generated covariates. We compare the predictive performance of our method to the MLE approach using the average mean integrated squared error for estimating the log-risk surfaces, given by

$$\frac{1}{n}\sum_{i=1}^{n}[h_g(\boldsymbol{X}_i) - \hat{h}_g(\boldsymbol{X}_i)]^2; g = 1,2,3$$

for each state transition hazard, separately. As shown in Table 2.2, both methods accurately recover the log-risk surfaces when the true underlying function is linear. However, in the non-linear settings, our deep neural network approach has a much lower mean integrated squared error, on average, compared to the classical MLE method, indicating a good performance of the proposed method.

Table 2.2: Average (SD) mean integrated squared errors under varying simulated log-risk surfaces for each state transition hazard

Simulation Settings				Maximum Likelihood Estimation			Deep Neural Network		
n	θ	Risk	Censoring	h_1	h_2	h_3	h_1	h_2	h_3
1,000	0.50	Linear	0%	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.07 (0.05)	0.08 (0.08)	0.08 (0.05)
10,000	0.50	Linear	0%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.08 (0.07)	0.08 (0.05)	0.08 (0.07)
1,000	2.00	Linear	0%	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)	0.12 (0.07)	0.13 (0.07)	0.13 (0.09)
10,000	2.00	Linear	0%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.11 (0.06)	0.11 (0.08)	0.13 (0.10)
1,000	0.50	Non-Linear	0%	1.80 (0.33)	1.82 (0.39)	1.85 (0.34)	0.09 (0.05)	0.09 (0.04)	0.08 (0.04)
10,000	0.50	Non-Linear	0%	1.80 (0.13)	1.77 (0.13)	1.78 (0.11)	0.07 (0.03)	0.08 (0.03)	0.08 (0.05)
1,000	2.00	Non-Linear	0%	1.92 (0.53)	1.85 (0.54)	1.96 (0.53)	0.15 (0.05)	0.15 (0.06)	0.14 (0.05)
10,000	2.00	Non-Linear	0%	1.82 (0.17)	1.81 (0.18)	1.83 (0.18)	0.14 (0.04)	0.12 (0.03)	0.13 (0.06)
1,000	0.50	Linear	25%	0.01 (0.02)	0.02 (0.01)	0.02 (0.02)	0.10 (0.06)	0.10 (0.07)	0.13 (0.12)
10,000	0.50	Linear	25%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.10)	0.12 (0.09)	0.12 (0.10)
1,000	2.00	Linear	25%	0.03 (0.02)	0.02 (0.02)	0.04 (0.03)	0.15 (0.10)	0.13 (0.09)	0.18 (0.12)
10,000	2.00	Linear	25%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.14 (0.10)	0.12 (0.08)	0.14 (0.10)
1,000	0.50	Non-Linear	25%	1.96 (0.44)	2.01 (0.54)	2.24 (0.66)	0.10 (0.07)	0.10 (0.06)	0.10 (0.08)
10,000	0.50	Non-Linear	25%	1.95 (0.15)	1.91 (0.16)	2.16 (0.20)	0.07 (0.04)	0.09 (0.08)	0.09 (0.07)
1,000	2.00	Non-Linear	25%	2.06 (0.62)	1.92 (0.72)	2.25 (0.79)	0.15 (0.08)	0.15 (0.08)	0.13 (0.06)
10,000	2.00	Non-Linear	25%	1.88 (0.21)	1.88 (0.21)	2.04 (0.28)	0.10 (0.05)	0.11 (0.06)	0.11 (0.05)
1,000	0.50	Linear	50%	0.01 (0.02)	0.02 (0.02)	0.04 (0.03)	0.10 (0.07)	0.10 (0.06)	0.20 (0.15)
10,000	0.50	Linear	50%	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.10 (0.07)	0.11 (0.08)	0.17 (0.16)
1,000	2.00	Linear	50%	0.03 (0.03)	0.03 (0.02)	0.05 (0.05)	0.22 (0.13)	0.17 (0.13)	0.24 (0.17)
10,000	2.00	Linear	50%	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.14 (0.09)	0.14 (0.10)	0.16 (0.14)
1,000	0.50	Non-Linear	50%	2.06 (0.50)	2.20 (0.72)	2.61 (1.00)	0.09 (0.06)	0.13 (0.13)	0.18 (0.14)
10,000	0.50	Non-Linear	50%	2.03 (0.21)	2.00 (0.18)	2.36 (0.25)	0.06 (0.03)	0.09 (0.08)	0.10 (0.09)
1,000	2.00	Non-Linear	50%	2.16 (0.76)	2.00 (0.72)	2.41 (0.91)	0.18 (0.10)	0.18 (0.09)	0.16 (0.10)
10,000	2.00	Non-Linear	50%	1.92 (0.25)	1.95 (0.23)	2.22 (0.38)	0.10 (0.05)	0.11 (0.06)	0.15 (0.13)

2.6 Boston Lung Cancer Study

We then utilize our approach to study a subset of patients from the Boston Lung Cancer Study (BLCS), a large hospital-based cancer epidemiology cohort investigating the molecular mechanisms and clinical pathophysiology of lung cancer [28]. The subset includes 5,296 patients with non-small cell lung cancer, diagnosed between June 1983 and October 2021. Also included in the dataset are patients' characteristics, namely, age at diagnosis (years), sex (0: male; 1: female), race (0: other; 1: white), ethnicity (0: non-Hispanic; 1: Hispanic), height (meters), weight (kilograms), smoking status (0: never; 1: former; 2: current), pack-years, cancer stage (1-4), and two indicators of genetic mutations (EGFR and KRAS).

Semi-competing events of cancer progression and death were documented in the data; the date of progression is the date of the first source evidence, including exam, radiology report or pathology. Progression followed by death was observed in 111 (2%) patients, progression but alive at the last followup date was observed in 224 (4%) patients, and death prior to progression was observed among 1,916 (36%) patients. To investigate the dependence of disease progression on death and to predict the the hazards of transitioning between states based on patient risk factors, we fit models (2.4)-(2.6) via our proposed approach. Specifically, we assumed Weibull baseline hazards for $\lambda_{0g}(s), g = 1, 2, 3$, and $\gamma_i \overset{i.i.d}{\sim} \text{Gamma}(1/\theta, 1/\theta)$. We then fit our proposed model to optimize the objective function (2.7) and output estimates for the finite dimensional parameters (ϕ's and θ) and the predicted $h_g, g = 1, 2, 3$ (log-risk estimates), for any covariate values.

We estimated the frailty variance, θ, to be 3.15 (bootstrapped 95% CI: 3.02-3.29), suggesting that progression is indeed correlated with death. **Figure 2.4** depicts the log-risk (h) functions for the effect of patient age at diagnosis on each state transition, stratified by sex assigned at birth and initial cancer stage, and fixing the other covariates to be at their sample means or modes. As shown, there seems to exist a non-linear effect of age that differs by transition, cancer stage and sex. The left panel shows that younger age and more advanced stage is associated with higher hazards for progression; for the transitions from diagnosis or progression to death (the middle and right panels), older age is associated with higher hazards; interestingly, while sex does not seem to play a role in disease progression (the left panel), male patients are more likely to die than female patients after diagnosis (the middle panel) or after progression (the right panel). Finally, more advanced stage is associated with higher hazards for all the transitions.

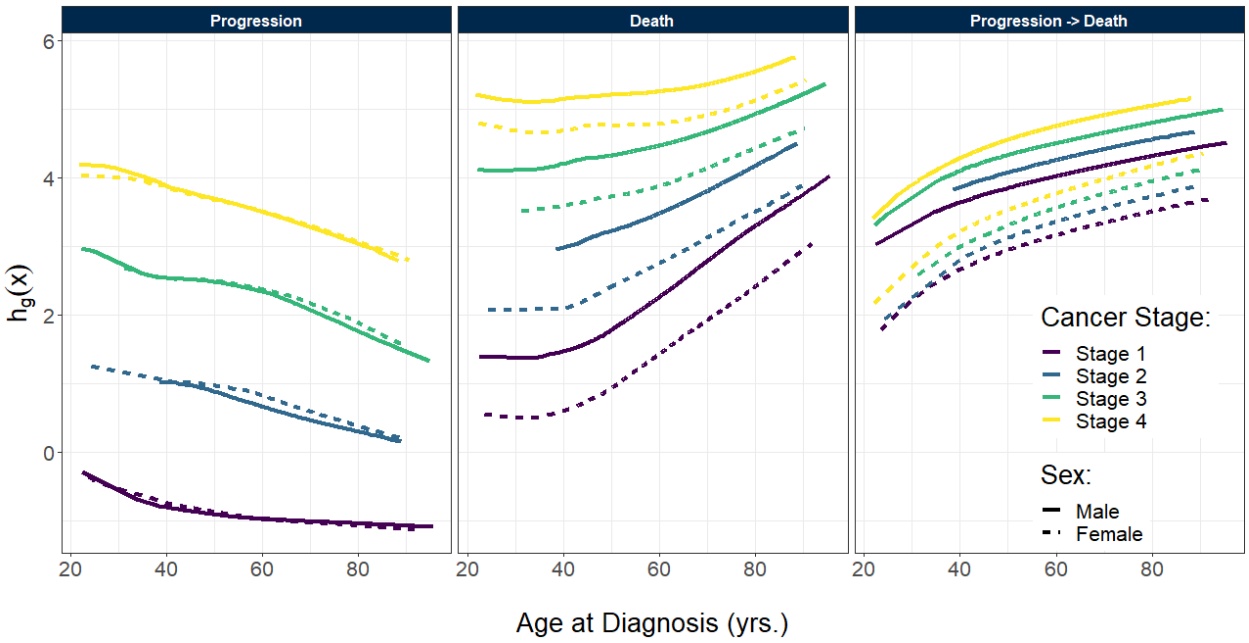


Figure 2.4: Log risk functions of age at diagnosis on each state transition, stratified by sex (solid versus dashed lines) and initial cancer stage (line color).

CHAPTER 3

Neural Expectation-Maximization Algorithm for Semi-Competing Risks

3.1 Background

In the previous chapter, baseline transition hazards were estimated via gradient methods under certain parametric assumptions. However, non-parametric baseline hazards may confer greater robustness, with much intensive computation. To address this, we propose a novel neural expectation-maximization algorithm, in which we hope to bridge the gap between classical statistical approaches and machine learning. Specifically, we extend our previous DNN-SCR method by proposing a hybrid approach to deep learning via our so-called neural expectation-maximization (NEM) algorithm. A common approach for estimating baseline hazard is through nonparametric maximum likelihood estimation (NPMLE), whereby the cumulative baseline hazards are taken to be non-decreasing step functions with jumps at unique observed failure times. This is solvable through gradient methods, however, the Hessian matrix for the NPMLE is not sparse. As its size increases linearly in n, computation is highly unstable. The estimation of a frailty parameter also adds complexity. To address these issues within the deep learning framework, we propose an NEM algorithm to estimate all the unknown parameters and functions. Our goal is to present an EM algorithm that is more numerically stable, especially for larger sample sizes. As deep learning can recover non-linear risk scores, we test our method by simulating risk surfaces of varying complexity. We then revisit the Boston Lung Cancer Study, where we investigate the impact of clinical and genetic predictors on disease progression and mortality.

3.2 Notation

Recall the notation we established in the previous chapter. We consider two events, a non-terminal event and a terminal event, with T_{i1} denoting the time to the non-terminal event, T_{i2}

denoting the time to the terminal event, and C_i denoting the censoring time for a given individual, $i = 1, \ldots, n$. Our observable data are $\mathcal{D} = (Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2}, \boldsymbol{X}_i)$, where $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = I(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = I(T_{i1} \leq Y_{i2})$, $\boldsymbol{X}_i$ is a p-vector of covariates, and $I(\cdot)$ denotes the indicator function. We model these outcomes in terms of three state transition hazards, $\lambda_{gi}(s) = \lambda_{0g}(s) \exp\left\{h_g(\boldsymbol{X}_i)\right\}$ and $\Lambda_{gi}(t) = \int_0^t \lambda_{gi}(s)ds$ for $g = 1, 2, 3$. Denoting the vector of frailties by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ and the collection of model parameters by $\boldsymbol{\psi} = \{\Lambda_{01}, \Lambda_{02}, \Lambda_{03}, \theta\}$, we can write the augmented data likelihood as

$$
\begin{aligned}
L(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} & \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \\
& \times \left[\lambda_{01}(Y_{i1})\, e^{h_1(\boldsymbol{X}_i)}\right]^{\delta_{i1}} \times \left[\lambda_{02}(Y_{i2})\, e^{h_2(\boldsymbol{X}_i)}\right]^{(1-\delta_{i1})\delta_{i2}} \\
& \times \left[\lambda_{03}(Y_{i2} - Y_{i1})\, e^{h_3(\boldsymbol{X}_i)}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1})\, e^{h_1(\boldsymbol{X}_i)} + \Lambda_{02}(Y_{i1})\, e^{h_2(\boldsymbol{X}_i)}\right.\right. \\
& \qquad \left.\left. + \delta_{i1}\Lambda_{03}(Y_{i2} - Y_{i1})\, e^{h_3(\boldsymbol{X}_i)}\right]\right\}.
\end{aligned}
\tag{3.1}
$$

Here, $\gamma_i \overset{i.i.d}{\sim} \Gamma(1/\theta, 1/\theta)$ (i.e., both shape and rate are $1/\theta$ so that the mean and variance are respectively 1 and θ), $i = 1, \ldots, n$, is a patient-specific *frailty* that models the dependence among the transition processes within subject i, that is, a larger value of θ reflects a stronger dependence.

3.3 Neural Expectation-Maximization Algorithm

In our previous work, we estimated the baseline hazard functions via gradient methods under certain parametric assumptions. For example, letting $\lambda_{0g}(s) = \phi_{1g}\phi_{2g}s^{\phi_{2g}-1}$, we specify Weibull baseline hazards with trainable parameters $\phi_{1g}, \phi_{2g}; g \in \{1, 2, 3\}$. Another common approach is non-parametric maximum likelihood estimation [145]. Under this approach, we specify the Λ_{0g} through non-decreasing step functions with jumps at unique observed failure times. This is also solvable through gradient methods, however, the Hessian matrix is not sparse, and its size increases linearly in n, making computation highly unstable. The estimation of the frailty parameter also adds complexity. To address these issues within a deep learning framework, we propose a novel neural expectation-maximization (EM) algorithm to estimate all the unknown parameters and risk functions. Our goal is to present an algorithm that is more numerically stable, especially for larger sample sizes [90], while allowing for flexible, non-parametric estimation of the covariate

risk functions. By viewing the subject-specific frailty as a missing variable, the algorithm iterates between three steps, namely the expectation (E) step, the maximization (M) step, and the neural (N) deep learning step. In the E-step, we update the conditional expectation the frailties, given the data and current values for the cumulative baseline hazards and risk functions. In the M-step, we estimate the jump sizes for the piece-wise constant baseline hazards by maximizing the expected log-likelihood found in the E-step, given the current estimates for the posterior expectations of the frailties. Then, fixing these quantities, we update our estimates of the log risk functions, $h_g(\boldsymbol{x}_i)$, and frailty variance, θ, as outputs of our neural network architectures in the N-step.

3.3.1 Conditional Frailty Distribution

To implement this approach, we need to derive the conditional distribution of γ_i based on the observed data and given the current estimates of the baseline hazards and risk functions. It can be shown that $\gamma_i|\mathcal{D}, \boldsymbol{\psi} \sim \text{Gamma}(\tilde{a}, \tilde{b})$, where

$$\tilde{a} = \frac{1}{\theta} + \delta_{i1} + \delta_{i2} \tag{3.2}$$

$$\tilde{b} = \frac{1}{\theta} + \Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}\left(Y_{i2} - Y_{i1}\right) e^{h_3(\boldsymbol{x}_i)}. \tag{3.3}$$

It follows that the posterior mean of γ_i is $\mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}] = \tilde{a}/\tilde{b}$, and the posterior mean of $\log(\gamma_i)$ is $\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}] = \text{digamma}(\tilde{a}) - \log(\tilde{b})$, where $\text{digamma}(\tilde{a}) = \partial \log[\Gamma(\tilde{a})]/\partial\tilde{a}$ and $\Gamma(\cdot)$ is the gamma function. Both quantities are needed for the E-Step. See Appendix B for details.

3.3.2 E-Step

The E-step calculates the expected log-conditional likelihood of the augmented data given the observed data, or our 'Q' function, which can be written as:

$$Q\left(\boldsymbol{\psi} \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}\right) = \mathbb{E}_{\boldsymbol{\gamma}}\left[\ell\left(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}\right) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}\right] = Q_1 + Q_2 + Q_3 + Q_4, \tag{3.4}$$

where $\ell\left(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}\right)$ is the logarithm of the likelihood in (3.1), $\boldsymbol{\psi}^{(m)}$ represents the current estimates of the parameters at the mth iteration, and Q_1, Q_2, Q_3, and Q_4 represent the additive pieces of the

Q function that are separable with respect to the model parameters:

$$
\begin{aligned}
Q_1 &= \sum_{i=1}^{n} \delta_{i1} \mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}^{(m)}] + \delta_{i1} \left\{ \log \left[\lambda_{01} \left(Y_{i1} \right) \right] + h_1(\boldsymbol{x}_i) \right\} \\
&\quad - \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \Lambda_{01} \left(Y_{i1} \right) e^{h_1(\boldsymbol{x}_i} \\
Q_2 &= \sum_{i=1}^{n} \delta_{i2} \mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}^{(m)}] + \left(1 - \delta_{i1}\right) \delta_{i2} \left\{ \log \left[\lambda_{02} \left(Y_{i2} \right) \right] + h_2(\boldsymbol{x}_i) \right\} \\
&\quad - \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \Lambda_{02} \left(Y_{i1} \right) e^{h_2(\boldsymbol{x}_i)} \\
Q_3 &= \sum_{i=1}^{n} \delta_{i1} \delta_{i2} \left\{ \log \left[\lambda_{03} \left(Y_{i2} \right) \right] + h_3(\boldsymbol{x}_i) \right\} - \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \delta_{i1} \left(\Lambda_{03} \left(Y_{i2} - Y_{i1} \right) \right) e^{h_3(\boldsymbol{x}_i)} \\
Q_4 &= \sum_{i=1}^{n} -\frac{1}{\theta} \log(\theta) + \left(\frac{1}{\theta} - 1 \right) \mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}^{(m)}] - \frac{1}{\theta} \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] - \log \Gamma \left(\frac{1}{\theta} \right).
\end{aligned}
$$

3.3.3 M-Step

The objective is to maximize the baseline hazard parameters given the expected log-likelihood and updated frailty estimates. As the maximizer of our objective function over the space of absolutely continuous cumulative baseline hazards does not exist [76], we restrict the parameter space of the cumulative baseline hazards, Λ_{01}, Λ_{02}, and Λ_{03}, to the one containing piecewise constant functions, with jumps occurring at observed event times. Maximizers over this discrete space are termed non-parametric maximum likelihood estimates of Λ_{01}, Λ_{02}, and Λ_{03}. Under this parameter space, $\lambda_{0g}(t)$ in (2.4) - (2.6) are replaced by $\Delta\Lambda_{0g}(t)$, the jump size at t for the baseline hazards of each state transition [91], and $\Lambda_{0g}(t) = \sum_{s=0}^{t} \Delta\Lambda_{0g}(s)$. Note that $\Delta\Lambda_{0g}(s) = 0$ if s is not one of the observed event times corresponding to state transition g. The M-step updates are

$$
\begin{aligned}
\Delta\Lambda_{01}^{(m+1)}(t) &= \frac{\sum_{i=1}^{n} \delta_{i1} I \left[Y_{i1} = t \right]}{\sum_{i=1}^{n} \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] I \left[Y_{i1} \geq t \right] \exp \left\{ h_1^{(m)}(\boldsymbol{x}_i) \right\}} \\
\Delta\Lambda_{02}^{(m+1)}(t) &= \frac{\sum_{i=1}^{n} \left(1 - \delta_{i1}\right) \delta_{i2} I \left[Y_{i2} = t \right]}{\sum_{i=1}^{n} \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] I[Y_{i2} \geq t] \exp \left\{ h_2^{(m)}(\boldsymbol{x}_i) \right\}} \\
\Delta\Lambda_{03}^{(m+1)}(t) &= \frac{\sum_{i=1}^{n} \delta_{i1} \delta_{i2} I \left[Y_{i2} - Y_{i1} = t \right]}{\sum_{i=1}^{n} \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \delta_{i1} I \left[Y_{i2} - Y_{i1} \geq t \right] \exp \left\{ h_3^{(m)}(\boldsymbol{x}_i) \right\}},
\end{aligned}
$$

where the numerators reflect the observed number of non-terminal events, the number of terminal events observed prior to non-terminal events, and the number of terminal events observed after non-terminal events, respectively These closed form updates in the M-step resemble Breslow-type estimators. As such, to seed the EM algorithm, we initialize Λ_{01}, Λ_{02}, and Λ_{03} with their respective,

Nelson-Aalen estimates. The frailty variance, θ, is involved in Q only through Q_4. Thus, in our N-step, we evaluate θ as a trainable parameter and take its update to be the direct maximizer from the neural network output. The starting value for θ can be taken to be any positive real number, however in practice, we use the estimated $\hat{\theta}$ arising from maximizing the objective function directly, assuming a linear log-risk function.

3.3.4 N-Step

Unlike approaches that parameterize $h_g(\boldsymbol{x}_i)$; $g \in \{1, 2, 3\}$, we opt for flexible, non-parametric versions of $h_g(\boldsymbol{x}_i)$ to better capture potential non-linear and higher-order dependencies between predictors and to maximize the predictive accuracy of our method. We propose a deep learning semi-competing risk model by estimating $\hat{h}_1(\cdot)$, $\hat{h}_2(\cdot)$, and $\hat{h}_3(\cdot)$ as outputs from three neural network sub-architectures. Specifically, we take negative log of Equation 3.1 to be the objective function for a multi-task deep neural network for modeling semi-competing outcomes based on potentially high-dimensional covariates. As before, our the N-step of our NEM approach consists of three risk-specific sub-networks, corresponding to the three transition hazards (see Figure 3.1).

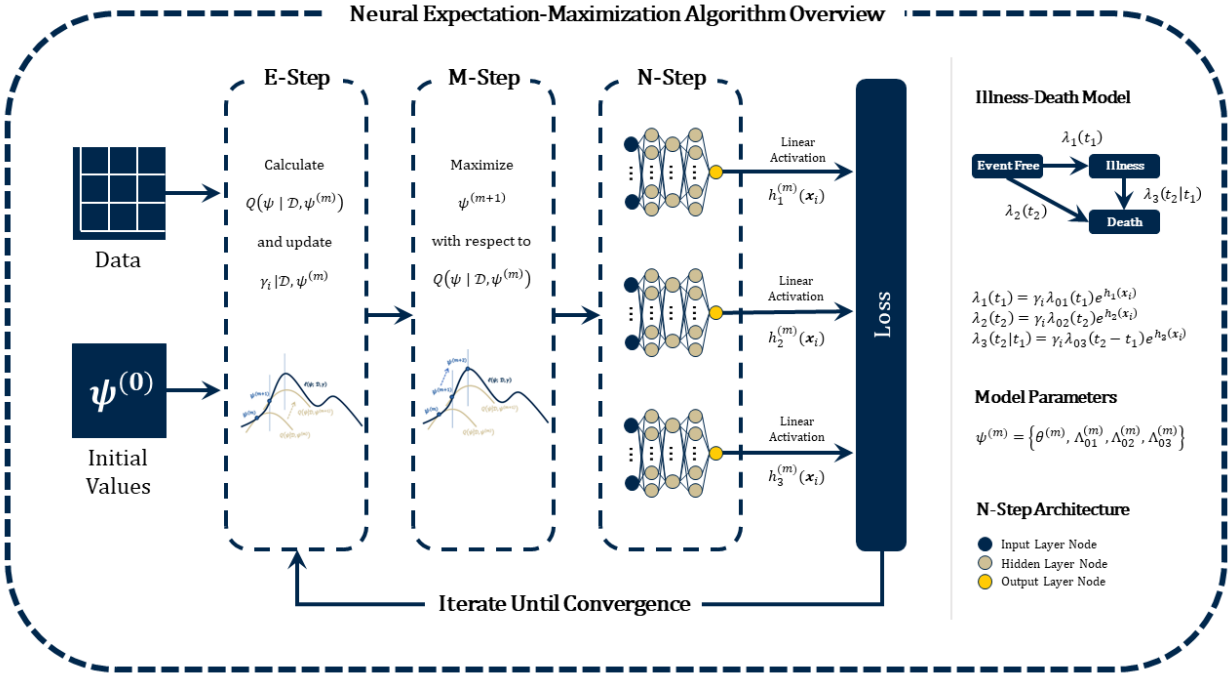


Figure 3.1: Overview of the neural expectation-maximization algorithm for semi-competing risks.

Each sub-network is made up of L layers, with k_l neurons in the lth layer ($l = 1, \ldots, L$). Sub-network predictions are based on an L-fold composite function, with ReLU activations in the hidden layers and a linear activation in the final layer. The number of hidden layers and nodes per layer,

as well as the dropout fraction, regularization, and learning rates are optimized as hyperparameters over a grid of values based on predictive performance.

We implement our approach using the `R` interface for the deep learning library TensorFlow, with model building and fitting done using Keras API [8, 7]. Taking advantage of the Keras paradigm for progressive disclosures of complexity, we implement our method as a custom Keras model, which has support for custom training, evaluation, and prediction methods within the context of a standard, user-friendly workflow. Finite parameter training for the frailty variance, θ, is done via the GradientTape API for automatic differentiation in a custom forward pass operation. Thus, the user need simply instantiate the DNN-SCR model as an R6 object with our custom model wrapper function and proceed with the typical workflow.

3.4 Simulation Study

We performed a series of simulation studies to validate our neural EM algorithm and illustrate the feasibility of our method. We simulated observed data, $\mathcal{D}$, from Equation 3.1, varying the sample size, population frailty variance, log-risk function, and censoring rates across 36 simulation settings. In particular, we simulated the shared frailty, γ_i, from a gamma distribution with mean 1 and variance θ, taking θ to be 0.5 or 2.0, corresponding to varying degrees of dependence between event times. The baseline hazard functions, λ_{01}, λ_{02}, and λ_{03}, were generated from Weibull distributions with shape and scale parameters ϕ_{g1} and ϕ_{g2}, respectively, $g \in \{1, 2, 3\}$. Across all simulations, we took $\phi_{11} = \phi_{21} = 2$, $\phi_{31} = 0.75$, $\phi_{12} = \phi_{22} = 2.25$, and $\phi_{32} = 2$. We generated two standard Normal random covariates, which were taken to be predictive of the morbidity and mortality hazards through one of three functions

- Linear: $h_g(\boldsymbol{X}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}_g;\ \boldsymbol{\beta}_g = \mathbf{1}_p = (1, 1, \ldots, 1);\ g = 1, 2, 3$
- Non-Linear: $h_g(\boldsymbol{X}_i) = \sum_{j=1}^{p} x_{ij}^3 \beta_{gj};\ \beta_{gj} = 1;\ g = 1, 2, 3;\ j = 1, \ldots, p$
- Non-Monotonic: $h_g(\boldsymbol{X}_i) = \log(|\boldsymbol{X}_i'\boldsymbol{\beta}_g| + 1);\ \boldsymbol{\beta}_g = \mathbf{1}_p = (1, 1, \ldots, 1);\ g = 1, 2, 3$

In the first scenario, we took a linear form so that the requirements for the classical models were satisfied, facilitating a fair comparison with a classical regression approach. In the second and third scenarios, we simulated the log-risk relationship as increasingly complex, non-linear functions of the data to highlight the utility of our method. Censoring times were generated from an exponential distributions to yield approximate censoring rates of 0%, 25% and 50%. Lastly, we vary the number of observations as $n = 1,000$ or $n = 10,000$. For each parameter configuration, a total of 500 datasets were independently generated. For our method, we varied the number of nodes per layer, the dropout fraction, the degree of regularization, and learning rate over a grid of

values to determine the setting with the best predictive performance. Performance was assessed via the bivariate Brier score integrated up to $t = 1$ year and the average mean integrated squared error (MISE) for estimating the log-risk surfaces for each state transition hazard, separately:

$$MISE_g = \frac{1}{n} \sum_{i=1}^{n} \left[h_g(\boldsymbol{X}_i) - \hat{h}_g(\boldsymbol{X}_i) \right]^2 ; \; g = 1, 2, 3$$

Tables 3.1 and 3.2 summarize the results of this simulation study. As shown in Table 3.1, estimation of the true frailty variance (θ) is consistent across all simulation settings, though the estimated $\hat{\theta}$ is slightly closer to the truth for smaller values of θ. Further, we integrated bivariate Brier score for one 'year' survival over a sequence of 100 evenly spaced time points in each simulation and compared the results from our method and the classical parametric regression model fit to a calculation which utilized the true model parameters. This was done to compare the fitted results to those results which signified the degree of irreducible error in the bivariate Brier score for each setting. Again, the results from our method are comparable with that of the 'true' bivariate Brier score across all simulation settings. Table 3.2 then compares our approach to the a standard regression in terms of the MISE for the predicted log-risk functions. In comparing the various simulation settings, it is shown that for both methods, the MISE increases slightly with the frailty variance and censoring rate. Further, the variability decreases with the increased sample size. Comparing the two methods under the different risk functions, it is shown that both methods accurately recover the log-risk surfaces for the respective state transitions when the true underlying function of the predictors is linear. However, in the non-linear settings, our neural EM approach has a much lower MISE, on average, compared to the classical approach, suggesting that our method out-performs the maximum likelihood approach when the functional form of the predictors is not truly linear. Lastly, in Figure 3.2, we graphically exemplify the estimation of the baseline cumulative hazard functions and population frailty variance under one simulation setting: $n = 1,000$, $\theta = 0.5$, log-risk function = non-monotonic, and censoring rates of 0%, 25%, and 50%.

3.5 Boston Lung Cancer Study

The work in this chapter is motivated by the Boston Lung Cancer Study (BLCS), one of the largest lung cancer survival cohorts in the world [28]. A primary objective of the BLCS is to better understand how risk factors influence a patient' disease trajectory, where they may experience adverse events such as a disease progression prior to death [70]. To address this, the BLCS has amassed a comprehensive database on patients enrolled at the Massachusetts General Hospital and the Dana-Farber Cancer Institute since 1992. The data collected by the BLCS contain demographics,

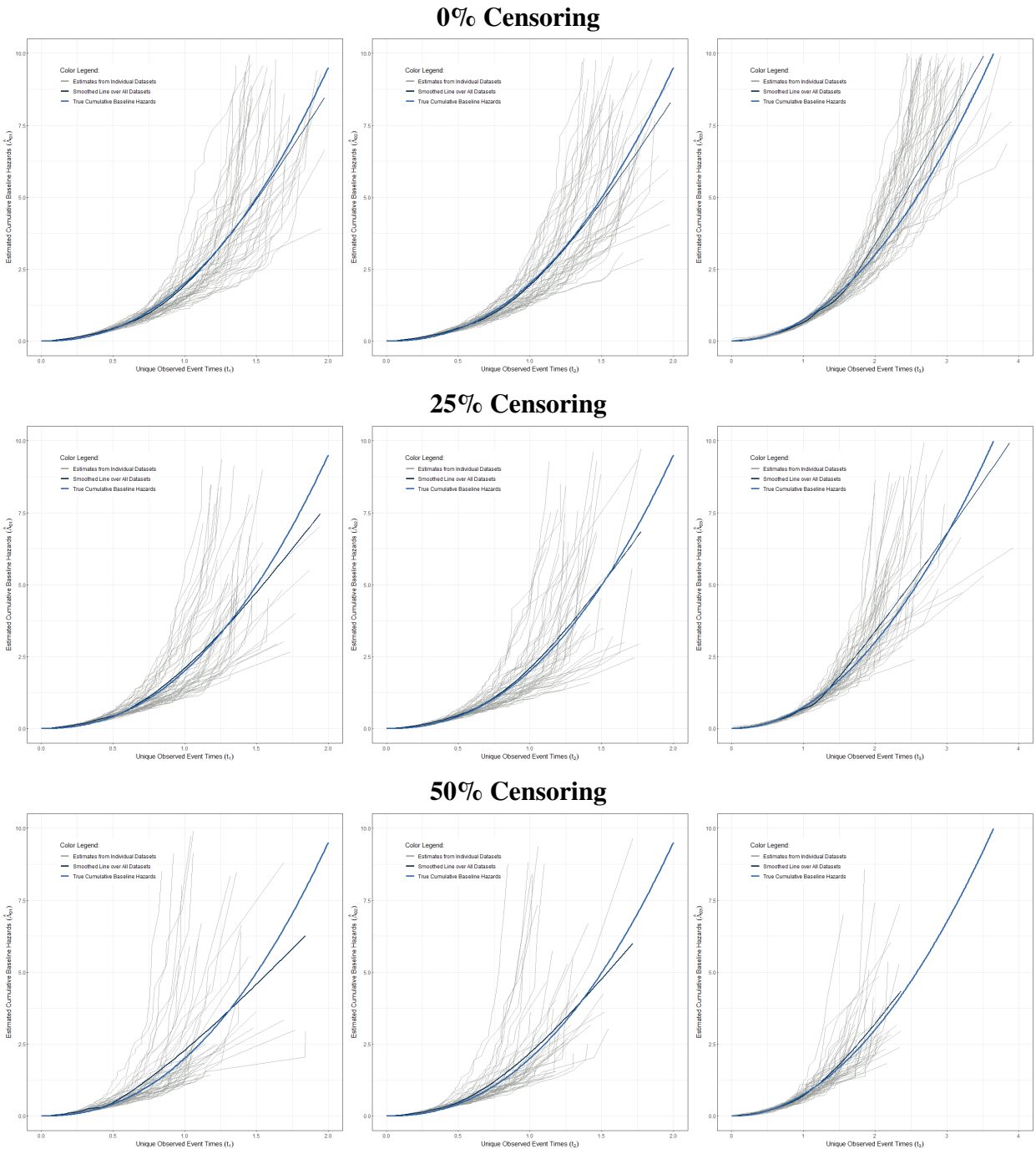


Figure 3.2: Estimated cumulative baseline hazard functions based on an example 50 generated datasets with $n = 1,000$, $\theta = 0.5$, log-risk function = non-monotonic, and censoring rates = 0%, 25%, and 50% (rows)

Table 3.1: Estimated frailty variance and one year integrated bivariate Brier score under various simulation settings

Simulation Settings			Frailty Variance Estimation			Integrated Bivariate Brier Score		
Sample Size	Risk Function	Censoring Rate	Truth	Parametric	Neural EM	Truth	Parametric	Neural EM
1,000	Linear	0%	0.5	0.49 (0.04)	0.49 (0.07)	0.1600 (0.0035)	0.1605 (0.0035)	0.1618 (0.0035)
10,000	Linear	0%	0.5	0.50 (0.01)	0.49 (0.03)	0.1584 (0.0013)	0.1585 (0.0013)	0.1594 (0.0014)
1,000	Linear	0%	2.0	1.96 (0.11)	1.92 (0.08)	0.1915 (0.0047)	0.1917 (0.0047)	0.1925 (0.0049)
10,000	Linear	0%	2.0	2.01 (0.03)	1.98 (0.04)	0.1911 (0.0016)	0.1911 (0.0016)	0.1921 (0.0015)
1,000	Non-Linear	0%	0.5	0.49 (0.05)	0.50 (0.08)	0.1822 (0.0038)	0.1841 (0.0036)	0.1855 (0.0036)
10,000	Non-Linear	0%	0.5	0.51 (0.01)	0.51 (0.02)	0.1821 (0.0011)	0.1836 (0.0011)	0.1858 (0.0014)
1,000	Non-Linear	0%	2.0	1.97 (0.11)	1.92 (0.07)	0.2244 (0.0022)	0.2258 (0.0022)	0.2271 (0.0024)
10,000	Non-Linear	0%	2.0	2.01 (0.03)	1.95 (0.03)	0.2245 (0.0009)	0.2251 (0.0009)	0.2276 (0.0015)
1,000	Non-Monotonic	0%	0.5	0.49 (0.05)	0.50 (0.08)	0.1822 (0.0038)	0.1841 (0.0036)	0.1855 (0.0036)
10,000	Non-Monotonic	0%	0.5	0.51 (0.01)	0.51 (0.02)	0.1821 (0.0011)	0.1836 (0.0011)	0.1858 (0.0014)
1,000	Non-Monotonic	0%	2.0	1.97 (0.11)	1.95 (0.07)	0.2244 (0.0022)	0.2258 (0.0022)	0.2271 (0.0024)
10,000	Non-Monotonic	0%	2.0	2.01 (0.03)	1.95 (0.03)	0.2245 (0.0009)	0.2251 (0.0009)	0.2276 (0.0015)
1,000	Linear	25%	0.5	0.47 (0.10)	0.47 (0.11)	0.1880 (0.0046)	0.1886 (0.0050)	0.1892 (0.0053)
10,000	Linear	25%	0.5	0.50 (0.05)	0.48 (0.02)	0.1899 (0.0030)	0.1900 (0.0029)	0.1914 (0.0029)
1,000	Linear	25%	2.0	2.00 (0.35)	1.95 (0.20)	0.2967 (0.0200)	0.2970 (0.0228)	0.2999 (0.0224)
10,000	Linear	25%	2.0	2.02 (0.12)	1.96 (0.08)	0.2979 (0.0074)	0.2979 (0.0069)	0.3030 (0.0079)
1,000	Non-Linear	25%	0.5	0.47 (0.14)	0.48 (0.12)	0.1851 (0.0045)	0.1866 (0.0040)	0.1879 (0.0048)
10,000	Non-Linear	25%	0.5	0.52 (0.05)	0.50 (0.04)	0.1858 (0.0010)	0.1874 (0.0012)	0.1893 (0.0025)
1,000	Non-Linear	25%	2.0	2.01 (0.23)	1.89 (0.21)	0.3042 (0.0176)	0.3065 (0.0220)	0.3088 (0.0201)
10,000	Non-Linear	25%	2.0	2.04 (0.10)	1.96 (0.10)	0.3032 (0.0034)	0.3044 (0.0048)	0.3113 (0.0093)
1,000	Non-Monotonic	25%	0.5	0.47 (0.14)	0.47 (0.12)	0.1851 (0.0045)	0.1866 (0.0040)	0.1879 (0.0048)
10,000	Non-Monotonic	25%	0.5	0.52 (0.05)	0.50 (0.04)	0.1858 (0.0010)	0.1874 (0.0012)	0.1893 (0.0025)
1,000	Non-Monotonic	25%	2.0	2.01 (0.23)	1.90 (0.21)	0.3042 (0.0176)	0.3065 (0.0220)	0.3088 (0.0201)
10,000	Non-Monotonic	25%	2.0	2.04 (0.10)	1.91 (0.10)	0.3032 (0.0034)	0.3044 (0.0048)	0.3113 (0.0093)
1,000	Linear	50%	0.5	0.47 (0.17)	0.47 (0.12)	0.1853 (0.0060)	0.1858 (0.0066)	0.1866 (0.0075)
10,000	Linear	50%	0.5	0.51 (0.05)	0.50 (0.05)	0.1912 (0.0042)	0.1912 (0.0046)	0.1934 (0.0044)
1,000	Linear	50%	2.0	1.96 (0.48)	1.87 (0.31)	0.3081 (0.0323)	0.3092 (0.0387)	0.3129 (0.0355)
10,000	Linear	50%	2.0	2.00 (0.16)	1.95 (0.06)	0.3050 (0.0133)	0.3050 (0.0129)	0.3090 (0.0142)
1,000	Non-Linear	50%	0.5	0.47 (0.20)	0.49 (0.12)	0.1794 (0.0037)	0.1808 (0.0032)	0.1830 (0.0056)
10,000	Non-Linear	50%	0.5	0.51 (0.06)	0.49 (0.03)	0.1819 (0.0022)	0.1833 (0.0026)	0.1853 (0.0072)
1,000	Non-Linear	50%	2.0	1.93 (0.28)	1.87 (0.21)	0.3074 (0.0211)	0.3090 (0.0239)	0.3127 (0.0264)
10,000	Non-Linear	50%	2.0	2.06 (0.11)	1.94 (0.09)	0.3076 (0.0065)	0.3092 (0.0077)	0.3179 (0.0146)
1,000	Non-Monotonic	50%	0.5	0.47 (0.20)	0.49 (0.13)	0.1794 (0.0037)	0.1808 (0.0032)	0.1830 (0.0056)
10,000	Non-Monotonic	50%	0.5	0.51 (0.06)	0.49 (0.04)	0.1819 (0.0022)	0.1833 (0.0026)	0.1853 (0.0072)
1,000	Non-Monotonic	50%	2.0	1.93 (0.28)	1.96 (0.21)	0.3074 (0.0211)	0.3090 (0.0239)	0.3127 (0.0264)
10,000	Non-Monotonic	50%	2.0	2.06 (0.11)	1.94 (0.09)	0.3076 (0.0065)	0.3092 (0.0077)	0.3179 (0.0146)

Table 3.2: Average (SD) mean integrated squared errors for the simulated log-risk surfaces, $h_g(\boldsymbol{X}_i)$, for each state transition hazard

Simulation Settings				Parametric Approach			Neural EM Algorithm		
Sample Size	Frailty Variance	Log-Risk Function	Censoring Rate	$h_1(\boldsymbol{X}_i)$	$h_2(\boldsymbol{X}_i)$	$h_3(\boldsymbol{X}_i)$	$h_1(\boldsymbol{X}_i)$	$h_2(\boldsymbol{X}_i)$	$h_3(\boldsymbol{X}_i)$
1,000	0.5	Linear	0%	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.07 (0.07)	0.08 (0.08)	0.07 (0.05)
10,000	0.5	Linear	0%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.08 (0.06)	0.07 (0.05)	0.07 (0.04)
1,000	2.0	Linear	0%	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)	0.12 (0.07)	0.11 (0.08)	0.12 (0.09)
10,000	2.0	Linear	0%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.11 (0.06)	0.11 (0.06)	0.13 (0.09)
1,000	0.5	Non-Linear	0%	0.17 (0.07)	0.15 (0.06)	0.19 (0.08)	0.07 (0.04)	0.09 (0.03)	0.07 (0.01)
10,000	0.5	Non-Linear	0%	0.17 (0.02)	0.19 (0.04)	0.18 (0.02)	0.08 (0.01)	0.08 (0.03)	0.08 (0.03)
1,000	2.0	Non-Linear	0%	0.22 (0.15)	0.27 (0.15)	0.22 (0.18)	0.15 (0.01)	0.10 (0.05)	0.11 (0.04)
10,000	2.0	Non-Linear	0%	0.20 (0.04)	0.19 (0.03)	0.20 (0.06)	0.14 (0.07)	0.14 (0.08)	0.12 (0.05)
1,000	0.5	Non-Monotonic	0%	1.79 (0.32)	1.79 (0.38)	1.83 (0.35)	0.09 (0.05)	0.09 (0.04)	0.09 (0.06)
10,000	0.5	Non-Monotonic	0%	1.82 (0.11)	1.81 (0.14)	1.76 (0.12)	0.07 (0.03)	0.08 (0.03)	0.08 (0.05)
1,000	2.0	Non-Monotonic	0%	1.89 (0.49)	1.86 (0.53)	1.97 (0.52)	0.15 (0.05)	0.13 (0.06)	0.14 (0.07)
10,000	2.0	Non-Monotonic	0%	1.82 (0.18)	1.80 (0.18)	1.85 (0.17)	0.14 (0.04)	0.12 (0.03)	0.14 (0.06)
1,000	0.5	Linear	25%	0.01 (0.02)	0.01 (0.01)	0.01 (0.02)	0.11 (0.10)	0.10 (0.07)	0.13 (0.12)
10,000	0.5	Linear	25%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.08)	0.12 (0.05)	0.12 (0.10)
1,000	2.0	Linear	25%	0.03 (0.02)	0.02 (0.02)	0.03 (0.03)	0.15 (0.10)	0.13 (0.08)	0.16 (0.11)
10,000	2.0	Linear	25%	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.14 (0.09)	0.12 (0.06)	0.14 (0.10)
1,000	0.5	Non-Linear	25%	0.18 (0.12)	0.15 (0.06)	0.32 (0.30)	0.10 (0.04)	0.08 (0.02)	0.07 (0.02)
10,000	0.5	Non-Linear	25%	0.19 (0.03)	0.21 (0.05)	0.21 (0.06)	0.08 (0.03)	0.08 (0.04)	0.09 (0.04)
1,000	2.0	Non-Linear	25%	0.29 (0.24)	0.31 (0.20)	0.38 (0.29)	0.14 (0.05)	0.11 (0.05)	0.12 (0.04)
10,000	2.0	Non-Linear	25%	0.21 (0.04)	0.20 (0.03)	0.19 (0.06)	0.12 (0.04)	0.14 (0.06)	0.19 (0.19)
1,000	0.5	Non-Monotonic	25%	1.97 (0.47)	2.02 (0.51)	2.18 (0.60)	0.10 (0.08)	0.10 (0.08)	0.12 (0.08)
10,000	0.5	Non-Monotonic	25%	1.92 (0.16)	1.91 (0.16)	2.16 (0.17)	0.09 (0.04)	0.09 (0.05)	0.11 (0.06)
1,000	2.0	Non-Monotonic	25%	2.00 (0.62)	1.97 (0.69)	2.25 (0.75)	0.13 (0.07)	0.15 (0.08)	0.13 (0.06)
10,000	2.0	Non-Monotonic	25%	1.85 (0.20)	1.85 (0.21)	2.12 (0.27)	0.10 (0.05)	0.11 (0.06)	0.11 (0.05)
1,000	0.5	Linear	50%	0.02 (0.02)	0.03 (0.02)	0.05 (0.03)	0.10 (0.07)	0.10 (0.09)	0.18 (0.17)
10,000	0.5	Linear	50%	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.10 (0.07)	0.11 (0.08)	0.17 (0.16)
1,000	2.0	Linear	50%	0.03 (0.03)	0.03 (0.02)	0.03 (0.05)	0.22 (0.13)	0.19 (0.13)	0.22 (0.17)
10,000	2.0	Linear	50%	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.14 (0.09)	0.14 (0.10)	0.16 (0.14)
1,000	0.5	Non-Linear	50%	0.16 (0.11)	0.20 (0.12)	0.43 (0.42)	0.11 (0.05)	0.09 (0.06)	0.08 (0.02)
10,000	0.5	Non-Linear	50%	0.19 (0.03)	0.22 (0.05)	0.22 (0.06)	0.08 (0.03)	0.08 (0.03)	0.11 (0.05)
1,000	2.0	Non-Linear	50%	0.30 (0.27)	0.31 (0.20)	0.37 (0.30)	0.14 (0.11)	0.12 (0.06)	0.13 (0.06)
10,000	2.0	Non-Linear	50%	0.22 (0.05)	0.20 (0.04)	0.19 (0.06)	0.16 (0.08)	0.14 (0.07)	0.14 (0.08)
1,000	0.5	Non-Monotonic	50%	2.04 (0.51)	2.00 (0.66)	2.57 (1.01)	0.11 (0.12)	0.13 (0.13)	0.18 (0.14)
10,000	0.5	Non-Monotonic	50%	2.04 (0.20)	2.05 (0.19)	2.33 (0.25)	0.06 (0.03)	0.09 (0.09)	0.14 (0.09)
1,000	2.0	Non-Monotonic	50%	2.13 (0.69)	2.00 (0.68)	2.43 (0.88)	0.18 (0.10)	0.18 (0.09)	0.16 (0.11)
10,000	2.0	Non-Monotonic	50%	1.94 (0.22)	1.95 (0.24)	2.25 (0.30)	0.10 (0.05)	0.11 (0.08)	0.15 (0.10)

social history, pathology, treatments, oncogenic mutation status, and other risk factors pertinent to these patient outcomes [97, 103]. Patients are recruited on a rolling basis to the BLCS upon initial lung cancer diagnosis and followed until death. During the course of follow-up, disease progression is recorded, which signifies a major non-terminal event in a patient's disease trajectory that modifies their risk of mortality. A combination of physical exam, imaging, and pathology data were used to determine the first data of progression. All-cause mortality was reported to the BLCS, with additional death information ascertained from the National Death Index and other sources. Thus, our semi-competing events are defined as Y_{i1} being the first instance of cancer progression or death, which could be censored by the end of follow-up, and Y_{i2} being the occurrence of death, either prior to or following progression, or censoring.

3.5.1 Study Sample

Among the 19,497 participants enrolled in the BLCS cohort, 7,585 were elibile for inclusion in this study. Eligibility was defined as having positive lung cancer diagnosis. Participants were ineligible if they were enrolled with esophageal cancer or other primary cancer, no cancer upon further study, or as a negative control in the case of spouses, friends, or other participants. Among those 7,585 eligible patients, we identified 7,462 (98.4%) with the temporal information necessary to define their semi-competing outcomes, namely (1) date of primary diagnosis, (2) progression and/or death date where applicable, and (3) last follow-up date or non-progression date. We further removed two patients with carcinoma *in situ*, i.e. stage 0. Thus, our final analytic cohort consisted of $n = 7,460$ patients diagnosed with lung cancer between June 1983 and October 2021. Disease progression was reported in 438 (5.9%) patients, with 143 (1.9%) patients experiencing progression followed by death and 295 (4.0%) patients alive by the end of follow up. In addition 2,720 (36.5%) patients died prior to progression (see Table 3.3).

Table 3.3: Semi-competing event rates among $n = 7,460$ patients in our analytic sample.

Progression Observed / Death Observed	Yes	No
Yes	143 (1.9%)	295 (4.0%)
No	2,720 (36.5%)	4,302 (57.7%)

Detailed information on patient demographics, smoking history, physiologic measurements, and genetic mutations were also collected. Potential demographic predictors included patient age at diagnosis (years), sex assigned at birth, self-identified race, ethnicity, and education level. Smoking status and pack-years of smoking also were included. Relevant clinical predictors included cancer

stage at diagnosis, histology, initial treatment, indications of chronic obstructive pulmonary disease (COPD) or asthma, and oncogenic (somatic driver) mutation status (EGFR or KRAS).

Table 3.4 reports summary statistics for these risk factors in our study sample. As shown, median (interquartile range (IQR)) age at diagnosis was 66 (58, 73) years, with 3,934 (53%) patients being female, 6,866 (92%) being white, and 6,419 (86%) being non-Hispanic. Clinically relevant features are as follows. Of the patients in our study sample, 6,003 (81%) had non-small cell lung cancer (NSCLC; adenocarcinoma, squamous cell carcinoma, or other/unspecified NSCLC), 301 (4%) had small cell lung cancer (SCLC), and 1,156 (15%) had lung cancer of other/unknown histologic type (e.g., mixed type). The majority of patients had a history of smoking (6,303; 84%), with a median (IQR) of 37 (12, 58) pack-years of smoking. Further, 1,512 patients (20%) were tested using the SNaPshot assay for the presence of genetic variants. The results of this testing revealed that 396 (5.3%) patients were postive for at least one KRAS variant and 285 (3.8%) patients were positive for at least one EGFR variant. Chronic obstructive pulmonary disease (COPD) was prevalent in 2,108 (28%) patient and 404 (5.4%) patients had asthma. Lastly, 4,350 (58%) of patients initially underwent surgery, while 1,841 (25%) patients initially received chemotherapy, 360 (4.8%) received radiation, and 909 (12%) received another form of initial treatment (Table 3.4). We note that these characteristics are similar to a recent study utilizing patient data from Massachusetts General Hospital, which draws comparisons to the BLCS cohort [150].

3.5.2 Univariate Associations

In an exploratory analysis, we first fit univariate Cox proportional hazards models to each event type (progression and death), separately, to understand the marginal associations between these events and our candidate predictors. We then fit fully-adjusted Cox models to the data. Hazard ratios (HR) and 95% confidence intervals (CI) are given in Table 3.5. As shown, initial cancer stage and treatment, as well as indications of COPD were significantly associated with disease progression in the fully adjusted model. This, however, does not account for death as a form of dependent censoring. In looking at overall mortality, we see that age, sex, education level, pack-years of smoking, histologic type, cancer stage, initial treatment, and indications of COPD were all associated with overall mortality (Table 3.5).

3.5.3 Predictive Modeling

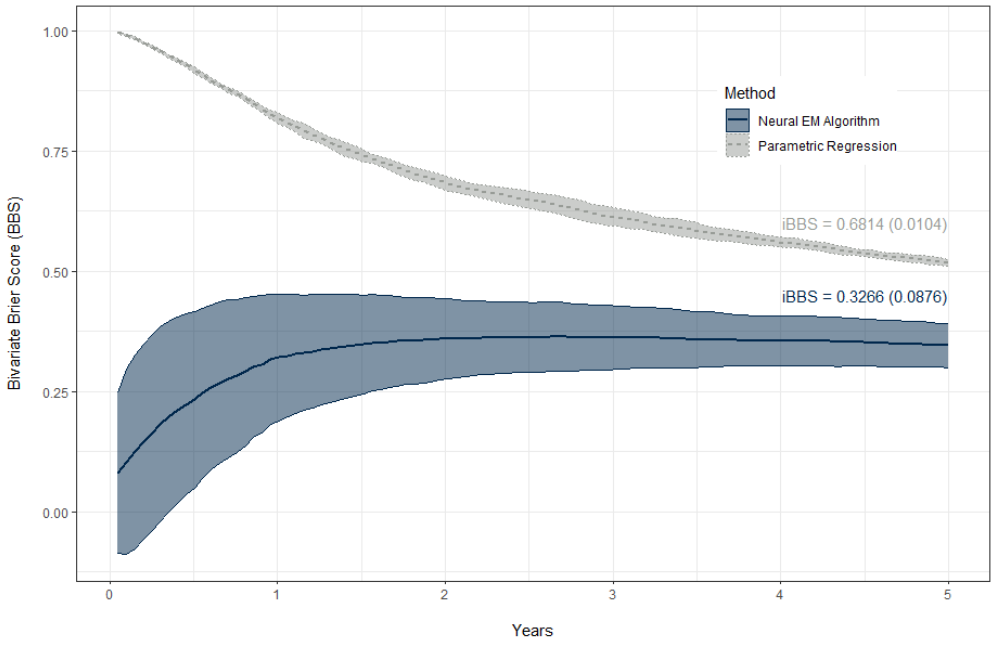
These variables were then used as candidate predictors in two modeling approaches. In the first approach, we fit our proposed neural EM algorithm, where we exemplify our method by estimating the hazards of cancer progression, mortality, and mortality following progression as non-linear functions of the predictors outlined above. We trained the model on a random 80% split of

Table 3.4: Demographic and clinical characteristics for the $n = 7,460$ patients diagnosed with lung cancer diagnosed between June 1983 and October 2021 in our analytic sample derived from the Boston Lung Cancer Study cohort. Summary statistics are reported as $n(\%)$ for categorical predictors and median (interquartile range) for continuous covariates.

Characteristic	**N = 7,460**[1]
Age at Diagnosis (yrs.)	66 (58, 73)
Unknown	723
Sex	
Female	3,934 (53%)
Male	3,408 (46%)
Unknown	118 (1.6%)
Race	
White	6,866 (92%)
Black	124 (1.7%)
Asian	142 (1.9%)
Other	102 (1.4%)
Unknown	226 (3.0%)
Ethnicity	
Non-Hispanic	6,419 (86%)
Hispanic	84 (1.1%)
Unknown	957 (13%)
Smoking Status	
Smoker	6,303 (84%)
Non-Smoker	988 (13%)
Unknown	169 (2.3%)
Pack-Years of Smoking	37 (12, 58)
Unknown	1,053
Histologic Type	
Adenocarcinoma	3,958 (53%)
Squamous Cell Carcinoma	1,175 (16%)
Non-Small Cell Lung Cancer, Unspecified	870 (12%)
Small Cell Lung Cancer	301 (4.0%)
Other/Unknown	1,156 (15%)
Stage	
1	2,926 (39%)
2	729 (9.8%)
3	1,438 (19%)
4	1,763 (24%)
Limited	86 (1.2%)
Extensive	96 (1.3%)
Unknown	422 (5.7%)
Initial Treatment	
Surgery	4,350 (58%)
Chemotherapy	1,841 (25%)
Radiation	360 (4.8%)
Other/Unknown	909 (12%)
EGFR Status	
Variant Negative	1,227 (16%)
Variant Positive	285 (3.8%)
Not Tested	5,948 (80%)
KRAS Status	
Variant Negative	1,116 (15%)
Variant Positive	396 (5.3%)
Not Tested	5,948 (80%)
Chronic Obstructive Pulmonary Disease	2,108 (28
Asthma	404 (5.4%)

[1]Median (IQR); n (%)

Figure 3.3: Average (SD) Bivariate Brier score (BBS) for our Neural EM Algorithm (blue, solid line) versus a semi-competing regression model (gray, dashed line), with 5-fold cross-validation. Integrated (BBS) was taken over 100 evenly spaced time points from time zero to five years.



the analytic sample, with 20% of this sample further used as a validation set during training. Hyperparameters, including the number of nodes per hidden layer, the learning rate, the dropout rate, and the regularization rate, were optimized over a grid of candidate values and chosen based on best predictive performance. We then tested our model on the remaining 20% of patients and calculated the Bivariate Brier score at one hundred evenly spaced time points from time zero to five years post-diagnosis. We compared our approach to a classical semi-competing regression model, where we again fit the model on a random 80% training split of the data and predicted the bivariate survival function on the holdout 20% testing set. For the traditional regression approach, we assumed the illness-death model that is semi-Markov with respect to the third transition hazard, as well as Weibull baseline hazards for each of the three state transitions. This procedure was repeated across each 80% and training and 20% testing split in 5-fold cross validation. Results from our prognostic modeling are given in Figure 3.3. As shown, the average five-year integrated Bivariate Brier score for our method was shown to be 0.3266 (0.0876), as compared to 0.6814 (0.0104) from the traditional regression model which assumes linear risk functions. This suggests that a model with linear risk function may not be predictive of progression or mortality, while a model that is agnostic to the form of the risk function may be better suited for survival prognostication.

Table 3.5: Results from unadjusted and adjusted Cox proportional hazards models studying the univariate associations between disease progression and death, separately, with our candidate predictors.

	Univariate Progression			Fully-Adjusted Progression			Univariate Mortality			Fully-Adjusted Mortality		
Characteristic	**HR[1]**	**95% CI[1]**	**p-value**	**HR[1]**	**95% CI[1]**	**p-value**	**HR[1]**	**95% CI[1]**	**p-value**	**HR[1]**	**95% CI[1]**	**p-value**
Age at Diagnosis (yrs.)	0.979	0.97, 0.99	< 0.001	1.00	1.00, 1.01	0.3	1.01	1.01, 1.01	< 0.001	1.02	1.01, 1.02	< 0.001
Sex												
Female	—	—		—	—		—	—		—	—	
Male	0.92	0.76, 1.11	0.4	1.03	0.84, 1.26	0.8	1.50	1.39, 1.61	< 0.001	1.37	1.26, 1.47	< 0.001
Unknown	1.32	0.65, 2.66	0.4	0.84	0.33, 2.14	0.7	0.78	0.51, 1.17	0.2	1.28	0.71, 2.30	0.4
Race												
White	—	—		—	—		—	—		—	—	
Black	1.26	0.65, 2.43	0.5	1.36	0.69, 2.68	0.4	0.84	0.62, 1.13	0.3	1.02	0.75, 1.37	> 0.9
Asian	2.73	1.76, 4.24	< 0.001	1.34	0.84, 2.13	0.2	0.61	0.43, 0.85	0.003	0.77	0.54, 1.08	0.13
Other	1.23	0.55, 2.75	0.6	0.956	0.42, 2.17	> 0.9	0.92	0.64, 1.32	0.7	1.06	0.73, 1.53	0.8
Unknown	2.24	1.50, 3.36	< 0.001	1.17	0.62, 2.19	0.6	0.71	0.55, 0.93	0.013	0.78	0.54, 1.14	0.2
Ethnicity												
Non-Hispanic	—	—		—	—		—	—		—	—	
Hispanic	1.49	0.71, 3.16	0.3	1.23	0.55, 2.75	0.6	0.57	0.37, 0.90	0.016	0.79	0.49, 1.26	0.3
Unknown	1.25	0.97, 1.62	0.085	1.18	0.85, 1.63	0.3	0.66	0.58, 0.75	< 0.001	0.981	0.85, 1.14	0.8
Smoking Status												
Smoker	—	—		—	—		—	—		—	—	
Non-Smoker	1.98	1.58, 2.47	< 0.001	0.91	0.68, 1.21	0.5	0.64	0.57, 0.72	< 0.001	0.91	0.79, 1.05	0.2
Unknown	2.75	1.75, 4.31	< 0.001	2.39	1.20, 4.75	0.013	0.57	0.40, 0.81	0.002	0.70	0.42, 1.18	0.2
Pack-Years of Smoking	0.987	0.98, 0.99	< 0.001	0.999	0.99, 1.00	0.5	1.01	1.01, 1.01	< 0.001	1.00	1.00, 1.00	< 0.001
Histologic Type												
Adenocarcinoma	—	—		—	—		—	—		—	—	
Squamous Cell Carcinoma	0.48	0.35, 0.67	< 0.001	0.93	0.66, 1.31	0.7	1.54	1.39, 1.71	< 0.001	1.31	1.17, 1.46	< 0.001
NSCLC, Unspecified[2]	0.953	0.72, 1.26	0.7	1.08	0.81, 1.45	0.6	0.92	0.80, 1.06	0.2	0.74	0.64, 0.85	< 0.001
Small Cell Lung Cancer	1.20	0.78, 1.85	0.4	1.20	0.50, 2.90	0.7	3.33	2.89, 3.84	< 0.001	1.25	1.03, 1.52	0.027
Other/Unknown	0.31	0.21, 0.47	< 0.001	0.46	0.30, 0.72	< 0.001	2.29	2.09, 2.52	< 0.001	1.26	1.13, 1.41	< 0.001
Stage												
1	—	—		—	—		—	—		—	—	
2	2.42	1.13, 5.21	0.024	2.30	1.07, 4.97	0.034	1.97	1.70, 2.28	< 0.001	1.81	1.56, 2.10	< 0.001
3	21.5	13.3, 34.7	< 0.001	19.2	11.7, 31.7	< 0.001	3.29	2.95, 3.67	< 0.001	2.56	2.27, 2.89	< 0.001
4	45.7	28.5, 73.4	< 0.001	22.1	13.0, 37.5	< 0.001	5.39	4.85, 6.00	< 0.001	4.80	4.19, 5.49	< 0.001
Limited	4.03	0.94, 17.3	0.061	3.12	0.58, 16.7	0.2	1.78	1.23, 2.56	0.002	0.92	0.61, 1.38	0.7
Extensive	73.4	38.3, 141	< 0.001	41.2	15.0, 114	< 0.001	10.4	8.17, 13.2	< 0.001	6.68	4.94, 9.02	< 0.001
Unknown	4.23	1.77, 10.1	0.001	11.4	4.50, 28.9	< 0.001	7.58	6.64, 8.65	< 0.001	3.01	2.53, 3.58	< 0.001
Initial Treatment												
Surgery	—	—		—	—		—	—		—	—	
Chemotherapy	10.2	8.17, 12.8	< 0.001	2.16	1.65, 2.85	< 0.001	3.25	2.97, 3.55	< 0.001	1.71	1.52, 1.91	< 0.001
Radiation	3.75	2.35, 6.00	< 0.001	1.50	0.92, 2.46	0.10	2.43	2.02, 2.93	< 0.001	1.93	1.59, 2.33	< 0.001
Other/Unknown	1.55	0.96, 2.50	0.072	0.67	0.39, 1.14	0.14	6.08	5.52, 6.69	< 0.001	2.81	2.47, 3.20	< 0.001
COPD[2]												
No	—	—		—	—		—	—		—	—	
Yes	0.40	0.31, 0.53	< 0.001	0.58	0.43, 0.77	< 0.001	0.91	0.84, 0.99	0.028	0.90	0.82, 0.98	0.017
Asthma												
No	—	—		—	—		—	—		—	—	
Yes	0.73	0.46, 1.18	0.2	0.65	0.40, 1.05	0.076	0.992	0.85, 1.16	> 0.9	1.02	0.87, 1.20	0.8
Genetic Variant Status												
Variants Negative	—	—		—	—		—	—		—	—	
Not Tested	0.20	0.16, 0.26	< 0.001	0.35	0.27, 0.44	< 0.001	1.71	1.48, 1.96	< 0.001	1.62	1.40, 1.87	< 0.001
KRAS or EGFR Positive	0.91	0.71, 1.17	0.5	1.04	0.80, 1.35	0.8	0.72	0.58, 0.89	0.002	0.88	0.71, 1.10	0.3

[1]HR = Hazard Ratio, CI = Confidence Interval

[2]NSCLC = Non-Small Cell Lung Cancer, COPD = Chronic Obstructive Pulmonary Disease

Figure 3.4: Average estimated cumulative baseline hazard functions and 95% bootstrap confidence intervals for each state transition based on 50 bootstrap samples of our data

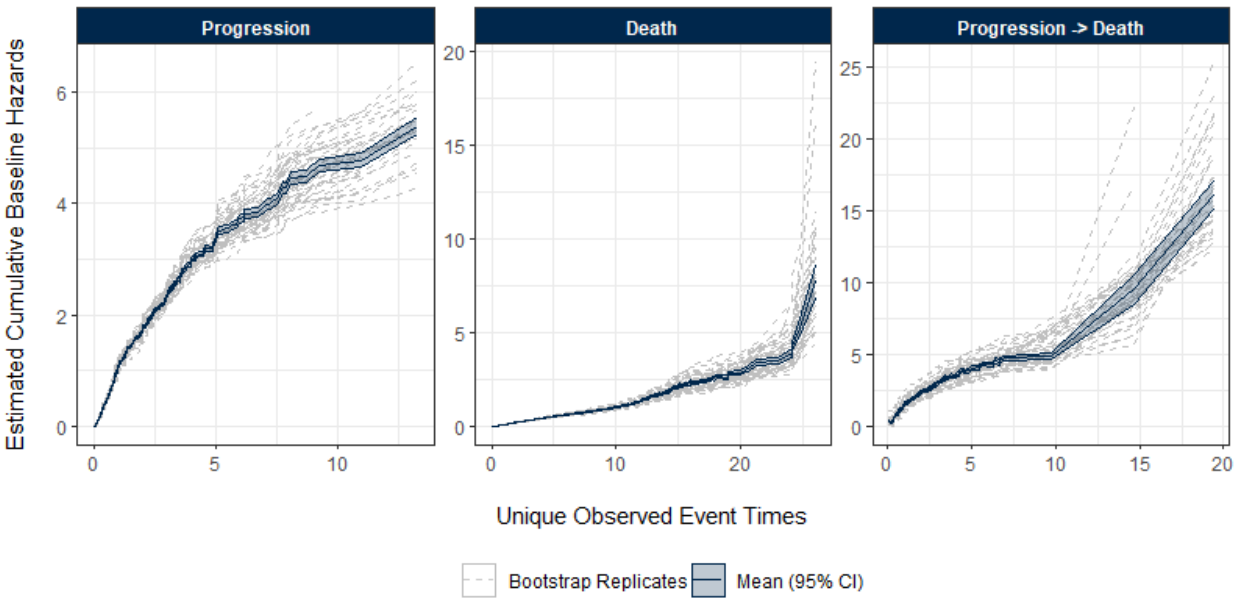


Figure 3.4 depicts the estimated cumulative baseline hazard functions for each state transition based on the 20% testing split of the cohort for each of the five-fold training splits. As shown, the baseline hazards are highest in the sojourn time between progression and death. Figure 3.5 depicts the log-risk (h) functions for the effect of patient age at diagnosis on each state transition, stratified by sex assigned at birth and smoking status. All other covariates were fixed to be at their sample means or modes for illustration. As shown, there is a non-linear relationship between age and the risk functions, which differ by transition, particularly in the transition from progression to death. Further, the risk of progression decreases with age but increases with death, both from diagnosis and from progression. Further, smoking status appears to have a strong effect on the risk of death from diagnosis, and to a lesser extent for the other state transitions. Lastly, across all state transitions, males have a higher risk of mortality, regardless of age and smoking status.

3.6 Discussion

In this chapter, we propose a neural expectation-maximization approach which, through a mixture of neural network architectures and trainable parameters, predicts time-to-event outcomes arising from a semi-competing risk framework (i.e., when an non-terminal event such as disease progression, modifies the risk of a patient's future survival). While previous work has developed machine

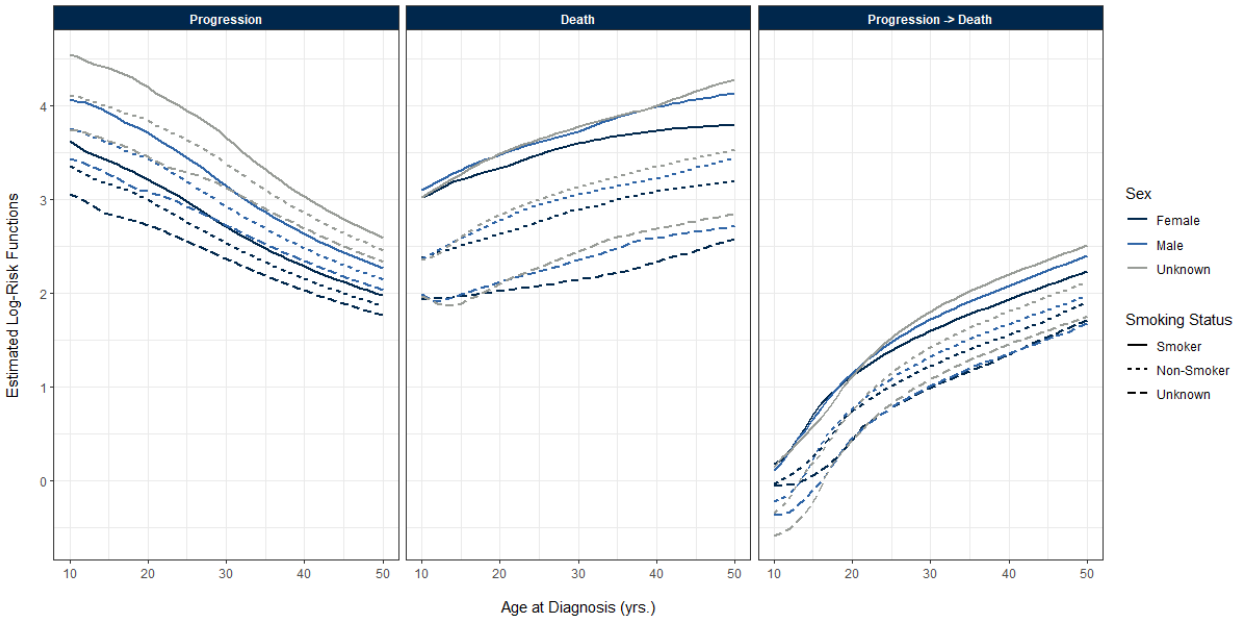


Figure 3.5: Example log-risk functions of age at diagnosis on each state transition, stratified by sex (line color) and smoking status (solid versus dashed lines).

learning approaches for multi-state or competing-risk settings [134, 89, 88, 1], in which progression and survival censor each other, we propose a new approach to further study the correlation between time to progression and death, and the modified hazards for mortality in the so-called 'sojourn time' between progression and death. In simulation, our results show high accuracy in estimating the relationship between predictors and the hazards of transitioning from disease onset to progression and death, particularly in situations where the risk relationship is increasingly complex.

Based on our analysis results, we detected several non-linear effects and interactions between commonly-studied risk factors such as age, sex assigned at birth, and smoking status. As shown in the predicted risk functions for our data, we note a potential interaction between sex assigned at birth and smoking status in the hazards for mortality. Such findings have been corroborated in [57] and more recently in [135]. [57] found that patterns of smoking differed by other well-known risk factors such as patient age, sex, tumor stage, and histology, with smoking and tumor stage being predictive of patient mortality. Further, they found significant interactions between smoking, clinical stage, and age with respect to progression. In our study, we note potential interactions between smoking status, age, and sex with respect to progression, as seen by the crossing of the covariate-specific risk functions. We also note that the difference in mortality hazards becomes more pronounced between males and females for each year higher age at diagnosis. The study by [135] examined the interacting effects between smoking status and other risk factors in patients with small-cell lung cancer, including age, sex, stage, and initial treatment. However, their study,

which focused on small-cell lung cancer, found that non-smokers had higher hazards for mortality, while the opposite is true in our cohort of small-cell and non-small cell cases. As opposed to these works, which considered the end points of progression and death without their shared dependence, our analysis treats these outcomes as semi-competing. In applying our method to the prediction of semi-competing outcomes in the Boston Lung Cancer Study Cohort, we found that our Neural EM approach had a much greater predictive accuracy than traditional semi-competing regression approaches. This is promising, as often, mortality is assessed without the consideration of progression as a competing event, or to avoid technical difficulties such as dependent censoring, composite endpoints such as progression-free survival are be constructed, which measure the time to the first of multiple possible events. However, the composite endpoints may mask the dependence of predictors on different endpoints, as the effects of certain clinical factors may differ across differing states in a patient's disease trajectory [9, 25]. Having a method which accurately predicts survival outcomes, while appropriately accounting for the dependence between multiple event types, will help improve clinical decision making.

Another major advantage of our approach over traditional regression models or other machine learning approaches is the use of deep learning for risk prediction. Deep neural networks have the ability to accommodate potentially high-dimensional predictors. Recent works have shown that estimates based on multilayer feedforward neural networks are able to circumvent the *curse of dimensionality* in nonparametric regression settings [14, 107]. Fully understanding this phenomenon is still a work in progress, but several authors reason that neural networks project the data into a much lower relevant representational space through weighting [3, 54].

While these results are promising, there are still several open problems to address. First, the implementation is computationally intensive, owing to the complex structure of the loss function and number of iterations required to achieve convergence. Future work will improve the efficiency of the proposed method. Further, our approach yields accuracy point estimates for the hazards of progression, mortality, and mortality following progression, however we do not yet have a means of quantifying the uncertainty surrounding these individualized risk predictions. We intend to extend the method in the framework of Bayesian neural networks to obtain prediction intervals. Lastly, our approach focuses on the joint distribution of the observed survival times for both event processes simultaneously. However, often in practice it is of interest to study the marginal distribution of the non-terminal event (e.g., disease progression) while appropriately accounting for the dependent censoring incurred by death. An alternate means of formulating this problem would focus on predicting the marginal survival function for disease progression in the presence of mortality as a semi-competing event. We will address these problems in subsequent work.

CHAPTER 4

A Pseudo-Value Approach to Causal Deep Learning of Semi-Competing Risks

4.1 Introduction

Lung cancer remains the leading cause of cancer-related deaths in the United States, accounting for one in five cancer-related deaths [122]. Significant progress has been made towards improving lung cancer prognosis, owing in part to better screening and advances in targeted therapies [92], however, the clinical course of patients with lung cancer is highly variable due to the complex genetic, environmental, and psycho-social risk factors which influence a patient's disease progression, and survival [129]. Furthering our understanding on the efficacy of patient-specific treatments is crucial when considering individualized approaches to care [140, 108].

More broadly, while mortality is often the primary endpoint when studying the effect of a particular treatment or exposure, non-fatal events may also impact illness trajectories and treatment decisions related to disease management. In the context of lung cancer, disease progression alter remaining available treatments, making lung cancer recurrence in patients who have undergone curative treatment an important endpoint [151, 44]. Thus, having a comprehensive understanding of a patient's event history, in particular, disease progression is important to inform clinical decision making. It is often of substantial interest to study the 'net' effect of an intervention or exposure on time to disease progression [38]. However, there are two challenges that hamper this analysis – how to evaluate causality in observational studies [52] and how to account for the the *semi-competing* relationship between disease progression and mortality [74].

As it is not always practical to conduct randomized controlled trials due to ethical or practical reasons, *causal inference* has emerged as a powerful tool for making statements about the etiology of an outcome based on changes in a *causal variable* of interest in the context of observational studies [104, 62]. The estimands can be the average risk difference (i.e., difference in survival probability between treatment groups) at a given point in time or the average difference in restricted mean life time [26, 112].

The presence of *semi-competing risks* can complicate causal inference by introducing dependent censoring, where the occurrence of death, or a *fatal* event, precludes recurrence, a *non-fatal* event. As a non-fatal event (recurrence) is often a precursor to the fatal event, this leads to informative censoring, which can bias estimates of treatment effects [74, 51, 99]. Much of the literature on causal methods for semi-competing risks are developed under a *potential outcomes* framework, using *principal stratification* to estimate causal effects [99, 32, 69, 146]. Principal stratification is a causal inference technique for handling post-treatment covariates in which patients are grouped based on post-treatment variables and causal effects are computed within these strata. For example, if we consider evaluating the causal effect of treatment, Z, on time to remission by time t_1, a principal stratification strategy would be to compute the survival average causal effect (SACE) among those individuals who would have survived as a member of either treatment or control group by some later time, t_2 [48]. Here, the interpretation of the survival average causal effect (SACE) is causal effect on remission among those individuals who would have survived as a member of either the treatment or control group until at least time t_2. However, in many contexts, it is unclear as to whether principal stratification is truly of scientific interest, or just a means of avoiding ill-defined counterfactual outcomes [105]. Further, many approaches for semi-competing survival functions use complicated objective functions, which require strong assumptions and are difficult to estimate with fidelity. Alternatively, when the outcome of interest is time to a non-fatal event, rather than the joint outcome of the non-fatal event and death, causal methods under the paradigm of 'truncation by death' have been developed [153, 96, 132, 152]. These approaches require special techniques to accommodate the presence of censoring.

Another promising approach to causal inference in survival analysis is through the use of pseudo-outcomes [11]. Here, the time-to-event outcome, which is subject to censoring, is replaced by a pseudo-survival probability, which represents a given individual's contribution to estimating the survival function of the study sample. This approach has several benefits. Firstly, using a discrete time survival approach avoids the need for common assumptions. Typical strategies involve the use of parametric families to characterize the distributions of the survival times, which may be too restrictive in practice, or utilize the Cox partial likelihood defined under the assumption of proportional hazards, which may not hold – particularly as the number of covariates increases. Further, pseudo-value based approaches replace the potentially censored survival times by jackknife-imputed survival probabilities. In the absence of censoring, standard loss functions can be utilized for optimization, rather than custom-designed approaches, and causal inference techniques such as inverse probability of treatment weighting (IPTW) or direct standardization via 'G-methods' are applicable.

Despite these advances, often, parametric and semi-parametric methods are limited in their ability to model complex relationships and interactions between covariates [77]. As such, there

has been a growing interest in applying machine learning to survival analysis, in order to improve the accuracy of models [142, 125]. Machine learning techniques, such as decision trees, random forests, and deep neural networks offer flexible and powerful approach for modeling survival data [120]. These methods can account for non-linear relationships and interactions between covariates and can handle high-dimensional datasets with many features. Several studies have demonstrated the effectiveness of machine learning approaches for survival analysis, including applications in cancer prognosis [156, 36, 40, 144]. Furthermore, the integration of causal inference into machine learning approaches has shown great promise for estimating the causal effects of treatments on survival outcomes. Several studies have proposed machine learning approaches for causal inference in survival analysis. For example, Hu et al. (2021) proposed an accelerated failure time Bayesian additive regression trees framework for estimating the heterogeneous survival treatment effects of lung cancer screening approaches [68], while Stitelman et al. (2012) proposed a general implementation of the targeted maximum likelihood estimator (TMLE) for longitudinal data in the context of a survival endpoint [130]. These studies and others demonstrate the potential of combining machine learning techniques with causal inference methods for survival analysis.

Many recent developments have been made towards applying deep learning approaches for estimation to survival analysis [148, 79, 80, 111]. However, while the potential effects of covariates are indeed estimated non-parametrically as outputs from neural network architectures in these settings, the construction of these loss function relies on an underlying Cox proportional hazards or Cox frailty model, which may carry strong assumptions, or the survival times themselves may be assumed to arise from a parametric family of distributions. In such cases, there is a disconnect between these likelihood-based loss functions and common deep learning algorithms [128]. Further applying deep learning to non-fatal event data with presents several challenges, including the need to account for dependent censoring, which requires careful modeling of the joint distribution of the semi-competing risks [119]. In an effort to address these issues, in this chapter, we propose a deep learning approach for estimating the causal effect of a given treatment on a non-fatal outcome in the presence of dependent censoring and potentially complex covariate relationships. In particular, we propose a three-stage approach. In the first stage, we estimate the marginal survival function for the non-fatal event based on a Clayton copula representation of the joint survival function. Following recent works by Andersen et al. (2017), Zhao et al. (2020), Sabathé et al. (2020), and Orenti et al. (2021), we propose a jackknife pseudo-value approach to circumvent the need for a complex loss function, whereby we estimate pseudo-survival probabilities at fixed time points as target values in the second stage [11, 154, 118, 102]. Estimation of pseudo-survival probabilities reduces the problem at hand to a straightforward minimization of the binary cross-entropy loss function. This approach further facilitates the development of causal estimators for such targets, which have been shown to be consistent and do not imposes common assumptions such as proportional hazards

across all time points. Lastly, we relate our pseudo outcomes to our causal variable of interest and additional confounders in a deep neural network to estimate survival average causal effect estimates via direct standardization.

The rest of this chapter is structured as follows. In Section 4.2, we review some notation and introduce concepts such as the Clayton copula, jackknife pseudo-values, deep learning, and our target estimand for causal inference before outlining our three-stage procedure and formulating our deep neural network. In Section 4.3, we provide a series of numerical studies to evaluate our proposed approach, and in Section 4.4, we apply our method to the Boston Lung Cancer Study, a large scale epidemiologic lung cancer cohort study. We conclude with some discussion on our current work and areas of future research.

4.2 Method

4.2.1 Notation

We consider two event types – a non-fatal event, such as disease recurrence, and a fatal event (i.e., death), and introduce the following notation. For a study consisting of n individuals, let T_{i1} and T_{i2} denote the times to the non-terminal and terminal events, respectively, for the ith individual; $i = 1, \ldots, n$. We observe Z_i, the causal variable of interest, and X_i, a p-vector of additional confounding variables. In the context of our data, Z_i is binary treatment indicator taking values $Z_i = 1$ if a patient underwent surgical resection and $Z_i = 0$ for other first-line treatment options. Further X_i include demographics, prevalent comorbidity conditions, or genetic variants for the ith subject. We assume $(T_{i1}, T_{i2}, Z_i, X_i)$ are i.i.d copies of (T_1, T_2, Z, X).

4.2.2 Bivariate Survival Function and the Clayton Copula

As a preamble, we consider a homogeneous situation, i.e., without covariates. We assume T_1 and T_2 are absolute, continuous random variables taking on non-negative values. Denote the marginal survival functions for the non-terminal and terminal events by $S_1(t_1) = Pr(T_1 > t_1)$ and $S_2(t_2) = Pr(T_2 > t_1)$, respectively. Note that the distribution of T_1 is non-parametrically identifiable only when the non-fatal event always precedes the fatal event [145]. Otherwise, as is the case in most practical settings, we assume a model for the joint survival distribution, given by

$$S(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2).$$

When the non-terminal and terminal events are positively correlated, it is natural to assume a Clayton copula model to express $S(t_1, t_2)$ as a functional of marginal survival functions, $S_1(t_1)$ and

$S_1(t_1)$ [31]:

$$S(t_1, t_2) = [S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1]^{-1/\theta} \tag{4.1}$$

where the copula dependence parameter, $\theta \geq 0$, measures the strength of the relationship between the non-fatal and fatal event times. Since the nonparametric function of (t_1, t_2) is only identifiable on the upper wedge, $0 < T_1 \leq T_2$, we assume model 4.1 on this upper wedge as well. Because model (4.1) may not hold in the lower wedge, the usual relationship that $\theta/(\theta + 2)$ = Kendall's τ may not hold [46].

4.2.3 Calculation of Distribution of Non-Fatal Event Time

Under the Clayton copula model, [46] show that the marginal survival function for the non-fatal event time is monotonic and estimable given the joint survival function in (4.1) and the marginal survival function for the fatal event. Specifically, for a fixed time point, t, the the joint survival function corresponds to the survival function for the first instance of either event, $S_*(t)$, which is often termed *progression free survival probability* in cancer research. The marginal survival function for the non-terminal event is related to the progression free survival probability and the survival function for the terminal event via

$$S_1(t) = [S_*(t)^{-\theta} - S_2(t)^{-\theta} + 1]^{-\frac{1}{\theta}}, \tag{4.2}$$

which constitutes the basis of estimating $S_1(t)$, as both $S_*(t)$ and $S_2(t)$ are estimable via the Kaplan-Meier method, because both the time to the terminal event and the time to either event are always observable. Moreover, several works have proposed estimates for θ, including the estimator given in Fine et al. (2001). In the setting where the marginal survival functions do not depend on covariates, We can estimate θ "ad hoc" via the concordance measure proposed by [101, 46]:

$$\frac{\sum_{i<j} W\left(Y_{ij1}, Y_{ij2}\right) D_{ij} \Delta_{ij}}{\sum_{i<j} W\left(Y_{ij1}, Y_{ij2}\right) D_{ij} \left(1 - \Delta_{ij}\right)} - 1 \tag{4.3}$$

where, for $1 \leq i \neq j \leq n$, we denote by $T_{ij1} = \min\left(T_{i1}, T_{j1}\right)$, $T_{ij2} = \min\left(T_{i2}, T_{j2}\right)$, and $C_{ij} = \min\left(C_i, C_j\right)$, and define $Y_{ij1} = \min\left(T_{ij1}, T_{ij2}, C_{ij}\right)$ and $Y_{ij2} = \min\left(T_{ij2}, C_{ij}\right)$ as the observable event times for the (i, j) pair. Further, $\Delta_{ij} = I\left[\left(T_{i1} - T_{j1}\right)\left(T_{i2} - T_{j2}\right) > 0\right]$ and $D_{ij} = I\left(T_{ij1} < T_{ij2} < C_{ij}\right)$, such that Δ_{ij} is estimable only when $D_{ij} = 1$. In contrast to the estimator of θ proposed in [46], we make a modification in (4.3) by subtracting 1. This is because the definition of θ in our formulation (4.2) corresponds to replacing θ in Equation (xx) of [46] by $\theta + 1$.

Lastly, let $Y_{i1} = \min(T_{i1}, T_{i2}, C_i)$ and $Y_{i2} = \min(T_{i2}, C_i)$ denote the observable event times for a

given individual. The weight function, $W_{a,b}(y_1, y_2)$, is defined as

$$W_{a,b}^{-1}(y_1, y_2) = n^{-1} \sum_i \{I\left(Y_{i1} \geq \min(a, y_1), Y_{i2} \geq \min(b, y_2)\right)\}$$

where constants a and b may be selected to dampen $W(\cdot)$ for large y_1 and y_2. Theoretically, Fine et al. (2001) show that $\hat{\theta}$ is a consistent estimator of θ, leading to the estimation of the non-fatal survival function in the absence of covariates.

4.2.4 Extension to the Distribution of Non-Fatal Event Time with Covariates

With covariates Z, X, the copula model (4.1) can be extended to

$$S(t_1, t_2|Z, X) = C_\theta[S_1(t_1|Z, X), S_2(t_2|Z, X)] = [S_1(t_1|Z, X)^{-\theta} + S_2(t_2|Z, X)^{-\theta} - 1]^{-1/\theta}, \quad (4.4)$$

where $S(t_1, t_2|Z, X) = \Pr(T_1 > t_1, T_2 > t_2|Z, X)$, $S_1(t_1|Z, X) = \Pr(T_1 > t_1|Z, X)$ and $S_2(t_2|Z, X) = \Pr(T_2 > t_2|Z, X)$. In this case, θ quantifies the correlation of T_1 and T_2 conditional on Z, X. Similarly, model (4.4) implies

$$S_1(t \mid Z, X) = [S_*(t \mid Z, X)^{-\theta} - S_2(t \mid Z, X)^{-\theta} + 1]^{-\frac{1}{\theta}},$$

which is the basis of estimating $S_1(t \mid Z, X)$. However, in this case, the estimator (4.3) of θ may not work as it was designed for a homogeneous population without considering covariates. Our idea is to extend estimator (4.3) by conditioning on Z, X. In particular, we propose to estimate $\hat{\theta}$ conditional on Z_i, X_i by focusing on the nearest k neighbors to subject i, using the Euclidean distance of covariates. We run through all the subjects and average these estimates to achieve an overall estimate of $\hat{\theta}$. We term the procedure a 'leave-one-in' approach. More specifically, let X denote the matrix of covariates, including the treatment variable, Z, where each sample $X_i \in X$ is a $(p+1)$-dimensional vector. We consider the Euclidean distance between X_i and $X_{i'}$ for $1 \leq i \neq i' \leq n$: $||(X_i - X_{i'}||_2 = \{\sum_{j=1}^{p+1}(x_{ij} - x_{i'j})^2\}^{1/2}$, where x_{ij} and $x_{i'j}$ are the jth components of X_i and $X_{i'}$, respectively; the Mahalanobis distance can also be used. In our numerical experience, both distances work almost equally well.

Then, for each individual, say, subject $i \in \{1, \ldots, n\}$, we identify the k nearest neighbors, among the n individuals, based on the their distances from this individual and denote them by $\mathcal{N}(i, k)$. We then estimate $\hat{\theta}$ based on subjects from $\mathcal{N}(i, k)$ via

$$\hat{\theta}^{(i)} = \frac{\sum_{j,l \in \mathcal{N}(i,k); j<l} W\left(Y_{jl1}, Y_{jl2}\right) D_{jl} \Delta_{jl}}{\sum_{j,l \mathcal{N}(i,k); j<l} W\left(Y_{jl1}, Y_{jl2}\right) D_{jl}\left(1 - \Delta_{jl}\right)} - 1.$$

Here, j and l index individuals in $\mathcal{N}(i,k)$. An overall estimate of θ is then given by

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} \hat{\theta}^{(i)}.$$

The number of neighbors, k, is chosen by visual examination of the estimated $\hat{\theta}$ values over a range of values for k. See Figure 4.1 for the results of this calculation over 50 generated datasets (black line = average value, grey ribbon = standard deviation) corresponding to Setting 2 in Section 4.3.

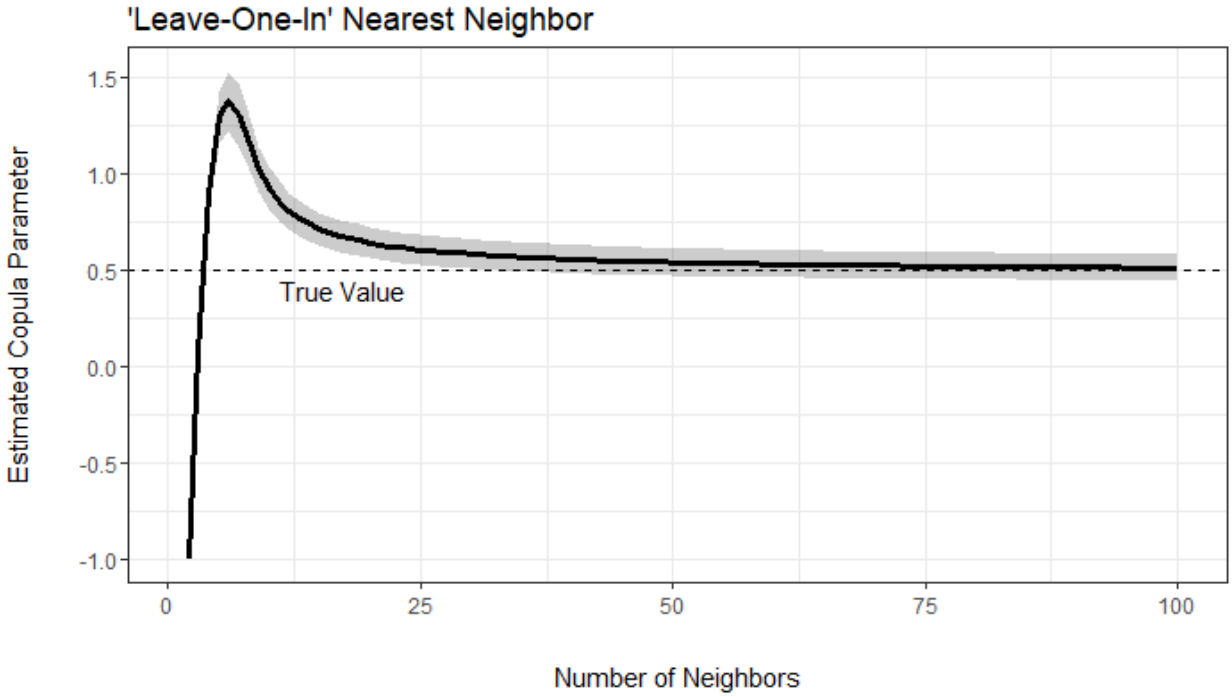


Figure 4.1: Example calculation across 50 simulated datasets with correlated covariates

4.2.5 Potential Outcomes Framework for Causal Inference

Under a potential outcomes framework, T_{i1}^{z} denotes the potential time to recurrence that would occur had $Z_i = z \in \{0,1\}$ for the ith individual. Causal inference infers the 'true' effect of an intervention on time to disease recurrence by comparing T_{i1}^{1} vs T_{i1}^{0} [104]. Before proceeding we make several common assumptions:

1. **Consistency**: $\exists\ \{T_{i1}^{1}, T_{i1}^{0}\}$ s.t. $T_{i1} = T_{i1}^{Z_i}$ almost surely. In other words, an individual's potential outcome under their assigned treatment group is the outcome that will actually be observed.

2. **Positivity**: $Z_i \in \{0, 1\}\ \forall X_i$, or the assumption that every individual has a non-zero probability of being assigned to either treatment group.

3. **No Interference**: T_{i1}^z is unaffected by the value of z for another subject, j.

4. **Exchangeability**: $T_{i1}^1, T_{i1}^0 \perp Z_i \mid X_i$, i.e., 'no unmeasured confounding.'

We are interested in the average causal effect of Z_i on the time to recurrence, T_{i1}. With the assumptions above, a common causal quantity of interest given the counterfactual potential outcomes is the average treatment effect (ATE), or the expected difference in potential outcomes over all individuals in the study. We can consider the average causal difference in the risk of recurrence at time t as

$$\mathbb{E}[I(T_{i1}^1 > t)] - \mathbb{E}[I(T_{i1}^0 > t)], \tag{4.5}$$

For Equation (4.5), note that $\mathbb{E}[I(T_{i1} \leq t)] = 1 - S_1(t)$. Thus, given a consistent estimator of $S_1(t)$, $\hat{S}_1(t)$, we can construct an 'S-learner' to estimate the ATE by training a deep neural network for $S_{i1}(t|X_i, Z_i)$ and predicting the potential outcomes $\hat{S}_{i1}(t|X_i, z); z \in \{0, 1\}$. An estimate of the ATE for the average causal risk difference is then given by

$$\hat{\text{ATE}} = n^{-1} \sum_{i=1}^{n} \{\hat{S}_{i1}(t|X_i, 1) - \hat{S}_{i1}(t|X_i, 0)\}. \tag{4.6}$$

4.2.6 A Pseudo-Values Approach for Causal Estimation

Our goal is to construct a model to study the difference in risk of recurrence at a given point in time. As the efficacy of a given treatment may change over time, common approaches to causal survival analysis such as the Cox Q-model may impose certain structures across all time points, e.g., proportional hazards, that are not realistic. Pseudo-values provide an intuitive means of circumventing the proportional hazards assumption, while also replacing potentially incompletely observed outcomes with a real-valued function of our outcome for each individual [11]. In general, for any function $f(t)$, pseudo-responses can be generated as $\hat{f}_i(t) = n\hat{f}(t) - (n-1)\hat{f}^{-i}(t)$, where $\hat{f}(t)$ is the overall estimate of $f(t)$ and $\hat{f}^{-i}(t)$ is an estimate omitting the ith subject. In our setting, consider J discrete time points, indexed by $j = 1, \ldots, J$. The probability of no recurrence by time t_j is given by $S_1(t_j) = \Pr(T_{i1} > t_j)$. A pseudo-outcome for individual i at time point t_j can be constructed as

$$\hat{S}_{i1}(t_j) = n \times \hat{S}_1(t_j) - (n-1) \times \hat{S}_1^{-i}(t_j)$$

where $\hat{S}_1(t_j)$ and $\hat{S}_1^{-i}(t_j)$ are the overall estimate of $S_1(t_j)$ using all n subjects and the 'leave-one-out' estimate excluding the ith subject, respectively, based on (4.2). Intuitively, this estimator for $S_1(t_j)$ represents the contribution of the ith individual in estimating $\mathbb{E}[S_1(t_j)]$ in a sample of n subjects. Further, because we have a consistent estimate of $S_1(t)$, $\hat{S}_{i1}(t_j)$ is approximately independent of $S_{i'1}(t_j)$ for $i \neq i'$ as $n \to \infty$ and

$$\lim_{n\to\infty} E[\hat{S}_{i1}(t_j) \mid Z_i, X_i] = S_1(t_j \mid Z_i, X_i)$$

for any i [6, 95]. With these results, the pseudo-values, $\hat{S}_{i1}(t)$, can then be used as numeric responses, similar to a logistic model fit to $I\left(T_{i1} > t_j\right)$ if the data were fully observed. However, as $I(T_i > t)$ is not observed for all subjects due to censoring, we must estimate the pseudo-responses for both the censored and uncensored individuals. We carry forward a design matrix of size $n \times (p + J)$, where p denotes the number of covariates in X and we include $J - 1$ dummy variables encoding time t_j.

4.2.7 Neural Network Architecture

The pseudo-value approach facilitates direct estimation of the target quantity of interest, without needing to optimizing the joint likelihood of the survival times directly. This circumvents the need for complex loss functions as part of the neural network architecture. Our deep neural network (DNN) directly minimizes the binary cross-entropy loss between the pseudo-survival probabilities, $\hat{S}_{i1}(t_j)$ and the predicted survival probabilities from the neural network output, $\pi_i(t_j)$, such that

$$\text{Binary Cross Entropy Loss} = \frac{1}{n}\left\{\sum_{i=1}^{n} -\hat{S}_{i1}(t_j)\log[\pi_i(t_j)] - [1-\hat{S}_{i1}(t_j)]\log[1-\pi_i(t_j)]\right\}.$$

Our proposed DNN is an S-learner consisting of a single fully-connected feed-forward neural network with an input layer, L hidden layers with k_l neurons in the lth layer; $l = 1, \ldots, L$, and an output layer [155, 83]. Hidden layers are connected via a non-linear activation function such as the rectified linear unit activation functions (ReLU; $\sigma_l(x) = \max(0, x)$), while the output layer's activation function is specified based on the target quantity. For example, as our target values are survival probabilities, a sigmoidal activation function ($\sigma_l(x) = \{1 + e^{-x}\}^{-1}$) is used for the final layer to constrain the output probabilities between 0 and 1. Estimation is based on an L-fold composite function

$$F_L(\cdot) = f_L \circ f_{L-1} \circ \cdots \circ f_1(\cdot) \text{ where } (g \circ f)(\cdot) = g(f(\cdot)),$$

$$f_l(x) = \sigma_l\left(\mathbf{W}_l x + \mathrm{b}_l\right) \in \mathbb{R}^{k_{l+1}},$$

where σ_l is an activation function, $\mathbf{W}_l$ are weights, and b_L are biases. Our network output is optimized under the binary cross-entropy loss function, which has a faster convergence rate than the traditional mean squared error due to its steeper gradient when the predicted output is far from the true output. Our final layer outputs a representation of the data, Ψ, which is used to then predict the counterfactual outcomes for each individual, $\hat{S}_{i1}(t|X_i, z); z \in \{0, 1\}$, before calculating

$$\hat{\hat{ATE}} = n^{-1} \sum_{i=1}^{n} \{\hat{S}_{i1}(t|X_i, 1) - \hat{S}_{i1}(t|X_i, 0)\}.$$

Hyperparameters needed to fully specify the neural network architecture include the number of hidden layers and number of nodes per hidden layer, the dropout fraction, and learning rate. In practice, these quantities are optimized over a Cartesian grid search based on predictive performance. We implement our approach with the `R` interface for Keras, using the deep learning library TensorFlow as the backend [7, 8].

4.3 Simulations

We next performed a series of simulations to assess the accuracy of our proposed approach against standard methods. In particular we varied the sample size, copula dependence parameter, censoring rates, and covariate-dependent risk functions in a fully factorial design. We considered two cases for the sample sizes, letting $n = 500$ or $n = 1,000$. Further, we let the copula dependence parameter, θ, equal 0.5, or 2, corresponding to a Kendall's τ value of 0.2, or 0.5. Dependent on each data generation model, we varied the parameters used to generate censoring times to achieve approximate censoring rates of 0% or 50%. Lastly, we considered two different generative model settings, described further below.

We considered two generative models of varying complexity. In the first setting, we simulated data from a proportional hazards model with a risk function that is linear in terms of the covariates, facilitating a fair comparison between the competing methods. In the second setting, we again simulated the data from a proportional hazards model, but we introduced a non-linear risk function through the use of higher order terms and correlated covariates.

Setting 1: Linear Risk Function

We first generated the data following the simulation scheme proposed in Peng and Fine (2007), Hsieh and Huang (2012), and Orenti et al. (2021) [106, 66, 102]. Specifically, we generated

non-fatal (T_{i1}) and fatal (T_{i2}) event times from marginal models specified by

$$\log(T_{i1}/3) = -(\beta_1 Z_i + \beta_1 X_{i1} + \beta_1 X_{i2}) + \varepsilon_{i1}$$
$$\log(T_{i2}/3) = -(\beta_2 Z_i + \beta_2 X_{i1} + \beta_2 X_{i2}) + \varepsilon_{i2},$$

where Z_i is a Bernoulli random variable with a success probability of 0.5, X_{i1} and X_{i2} are independent truncated normal random variables with mean 1, variance 0.5, and truncation bounds of [0, 2], and $(\varepsilon_{i1}, \varepsilon_{i2})$ are correlated random errors. To induce dependence between the simulated event times, we simulate ε_{i1} and ε_{i2} from the Clayton copula model,

$$\left[\Pr\left(\varepsilon_{i1} > t_1\right)^{-\theta} + \Pr\left(\varepsilon_{i2} > t_2\right)^{-\theta} - 1\right]^{-\frac{1}{\theta}},$$

where ε_{i1} and ε_{i2} follow the extreme value distribution, i.e., $\Pr\left(\varepsilon_{i1} > t_1\right) = \exp\{-\exp(t_1)\}$ and $\Pr\left(\varepsilon_{i2} > t_2\right) = \exp\{-\exp(t_2)\}$ [116]. Across all simulation settings, we fixed $\beta_1 = 1$ and $\beta_2 = 0.2$. In settings where the event times may be censored, we generated independent censoring times, C_i, from a mixture of uniforms, where $C_i \sim \text{Unif}(0, 1)$ with probability 0.2 and from $\text{Unif}(1, 1.2)$ with probability 0.8, yielding an approximate censoring rate of 50%.

Setting 2: Non-Linear Risk Function, Correlated Covariates

In our second data generation scenario, we adopted a similar framework as described previously, but we have modified the covariate risk functions to include higher-order terms and correlations. We generated three covariates, $\boldsymbol{X} = (X_1, X_2, X_3)'$, from a multivariate normal distribution with $\boldsymbol{X} \sim N_3(\boldsymbol{0}, \Sigma)$, where the covariance matrix, Σ, is AR(1) with elements $(\sigma_{ij}) = 0.5^{|i-j|}$. We then dichotomized $Z_i = \mathbb{I}(X_{i1} \geq 0)$ to be a binary covariate representing our causal variable of interest. We generated the event times, (T_{i1}) and (T_{i2}), from marginal models specified by

$$\log(T_{i1}/3) = -(\beta_1 Z_i + \beta_1 X_{i1}^2 + \beta_1 X_{i2}^2) + \varepsilon_{i1}$$
$$\log(T_{i2}/3) = -(\beta_2 Z_i + \beta_2 X_{i1}^2 + \beta_2 X_{i2}^2) + \varepsilon_{i2},$$

to understand the performance differences between our non-parametric approach and approaches which are mis-specified when assuming a linear form with independent covariates.

Across all scenarios, we independently generated 50 datasets and calculated the average bias and mean squared error (MSE) for the estimated average treatment effect (ATE) for our proposed approach against a causal Q-model, which was fit using generalized estimating equations with a complementary log-log mean link, corresponding to the proportional hazards model [102]. For the methods which relied on the calculation of pseudo-values, we first estimated the copula dependence parameter using the 'leave-one-in' approach described previously, applied to the entire sample of

n observations. We carried forward the estimated $\hat{\theta}$ to calculate the pseudo- non-fatal survival probabilities at fixed time points $t = 0.2$, 0.4, 0.6, 0.8, and 1.0. For our method, we hypertuned our DNN parameters once per simulation setting and carried forward the best configuration of hyperparameters across all 50 datasets. Lastly, for each method, we randomly split each dataset into an 80% training set and a 20% testing set. We fit the respective models on the training set and calculated the ATE at $t = 1.0$ in the testing set.

Table 4.1 summarize the results of this simulation study. As shown, model performance was similar in the first data generation setting where the parametric Q-model is correctly specified, though the correct model is slightly less biased and more efficient. This is to be expected, as we are fitting the true model to the data, while the DNN represents a stochastic approximation of the true data generation function. In the second setting, however, the performance for our proposed approach is better, as the true covariate risk function contains correlated covariates and higher-order terms. While the degree of bias for the proposed approach remains fairly consistent with the first data generation setting, the bias increases for the parametric Q-model. We also note that for both methods, performance was typically better in settings with a larger sample size ($n = 1,000$ versus 500), a smaller degree of dependence between the event times ($\theta = 0.5$ versus 2.0), and when the data were fully observed versus censored, as expected.

4.4 Boston Lung Cancer Study

The Boston Lung Cancer Study is a collaborative research effort between Dana-Farber Cancer Institute and Massachusetts General Hospital which focuses on improving the understanding and treatment of lung cancer, one of the leading causes of cancer-related deaths worldwide [27].

4.4.1 Study Population

Among all participants in the Boston Lung Cancer Study (BLCS) cohort, 7,755 were initially eligible for inclusion in this analysis. Eligibility was defined as having a positive lung cancer diagnosis. Participants were ineligible if they were enrolled with esophageal cancer or other primary cancer, no cancer upon further study, or as a negative control in the case of spouses, friends, or other participants. Among those 7,755 eligible patients, we identified 7,697 (99%) with the temporal information necessary to define their semi-competing outcomes, namely (1) date of primary diagnosis, (2) recurrence, progression, and/or death date where applicable, and (3) last follow-up date or non-progression date. We further removed 56 patients diagnosed in the past 6 months, 25 patients with negative survival times, 212 patients with small-cell lung cancer, and 6 patients with carcinoma *in situ* , i.e., stage 0 (Figure 4.2). As available treatment options are

Table 4.1: Average bias and mean squared error (MSE) for estimated vs. true ATE comparing our proposed method to the parametric Q-Model. Results are averaged over 50 independently generated datasets for each setting.

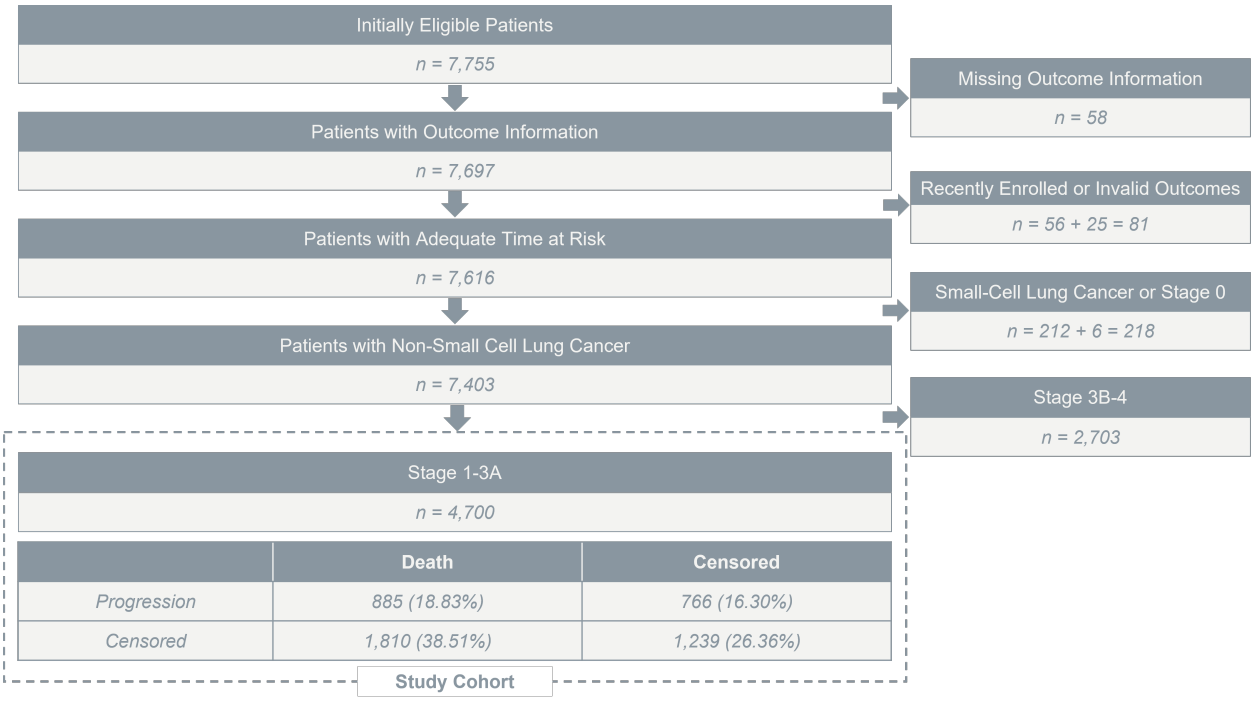
Simulation Settings			**Bias**		**Mean Squared Error**	
n	θ	**Censoring**	**Q-Model**	**Proposed**	**Q-Model**	**Proposed**
Setting 1						
500	0.5	50%	0.0025	0.0060	0.0020	0.0063
500	0.5	0%	0.0025	0.0045	0.0022	0.0042
500	2.0	50%	0.0025	0.0057	0.0022	0.0053
500	2.0	0%	0.0018	0.0069	0.0019	0.0011
1000	0.5	50%	0.0018	0.0025	0.0013	0.0028
1000	0.5	0%	0.0023	0.0035	0.0014	0.0028
1000	2.0	50%	0.0019	0.0048	0.0014	0.0037
1000	2.0	0%	0.0018	0.0030	0.0012	0.0021
Setting 2						
500	0.5	50%	0.0483	0.0043	0.0076	0.0032
500	0.5	0%	0.0520	0.0030	0.0078	0.0031
500	2.0	50%	0.0444	-0.0083	0.0081	0.0045
500	2.0	0%	0.0476	-0.0030	0.0079	0.0046
1000	0.5	50%	0.0485	-0.0043	0.0036	0.0028
1000	0.5	0%	0.0518	-0.0034	0.0038	0.0024
1000	2.0	50%	0.0444	-0.0040	0.0046	0.0032
1000	2.0	50%	0.0475	-0.0035	0.0042	0.0033

predicated on a patient's cancer stage, we considered two subgroups of patients – those who were diagnosed with stages 1-3a NSCLC (4,700; 63.5%) and those who were diagnosed with stages 3b-4 NSCLC (2,703; 36.5%). As stages 1-3a are widely considered to be operable, we focused on understanding the average treatment effect of first-line surgical resection on time-to-relapse among this subset of patients (Figure 4.2).

4.4.2 Patient Characteristics

Descriptive statistics for the study cohort are given in Table 4.2. As shown, median age among all patients with NSCLC was 66 years old [interquartile range (IQR): 59-74], with a majority of patients identifying as female (54%), White/Caucasian (92%) and non-Hispanic (87%). Further, the majority of study participants were former smokers (57%) with a median 40 pack-years of smoking (IQR: 16-53). Among all patients, the majority underwent surgical resection (4,444;

Figure 4.2: Flowchart of inclusion and exclusion criteria for the Boston Lung Cancer Study analytic sample and distributions of observed outcomes (progression and/or death).



67%) as first-line treatment. However, stratifying by stage at diagnosis, we found that patients with earlier-stage diagnoses were slightly older (median, IQR age: 68, 61-74 years versus 64, 56-72 years), with a higher proportion being female (55% versus 50%) and White/Caucasian (93% versus 92%), and a lower proportion identifying as non-Hispanic (85% versus 90%). Social history differed between these two groups as well, with more former smokers (60% versus 53%) as compared to current smokers (25% versus 30%) in the earlier-stage group, though a higher median number of pack-years of smoking (40 versus 37 pack-years). Lastly, rates of testing for two common genetic variants, EGFR and KRAS, differed between these groups, with more patients (81% versus 76%) tested in the earlier-stage group. Among those tested, we observed a higher proportion of patients in the earlier-stage group with a KRAS mutation (30% versus 21%), though a higher proportion in the late-stage group with an EGFR mutation (18% versus 21%). We then carried forward our final analytic cohort of 4,700 patients diagnosed with non-small cell lung cancer (NSCLC), stages 1-3a. Disease recurrence was reported in 1,651 (35.13%) patients, with 885 (18.83%) patients experiencing recurrence followed by death and 1,810 (38.51%) patients who died prior to recurrence (Figure 4.2).

Table 4.2: Characteristics of the $n = 7,403$ patients in the Boston Lung Cancer Study cohort, overall and stratified by stage at diagnosis.

Characteristic	**Overall, $n = 7,403$[1]**	**Stage at Diagnosis**	
		1-3A, $n = 4,700$[1]	**3B-4, $n = 2,703$[1]**
First-Line Treatment			
Chemotherapy	1,851 (28%)	365 (8.0%)	1,486 (70%)
Other	7 (0.1%)	2 (<0.1%)	5 (0.2%)
Radiation	366 (5.5%)	194 (4.3%)	172 (8.1%)
Surgery	4,444 (67%)	3,994 (88%)	450 (21%)
Unknown	735	145	590
Age at Diagnosis (yrs.)	66 (59, 74)	68 (61, 74)	64 (56, 72)
Body Mass Index	26.4 (23.0, 31.1)	26.6 (23.3, 31.1)	25.7 (22.6, 30.1)
Sex			
Male	3,431 (46%)	2,093 (45%)	1,338 (50%)
Female	3,966 (54%)	2,603 (55%)	1,363 (50%)
Unknown	6 (<0.1%)	4 (<0.1%)	2 (<0.1%)
Race			
White/Caucasian	6,834 (92%)	4,349 (93%)	2,485 (92%)
Other	364 (4.9%)	212 (4.5%)	152 (5.6%)
Unknown	205 (2.8%)	139 (3.0%)	66 (2.4%)
Ethnicity			
Non-Hispanic	6,410 (87%)	3,990 (85%)	2,420 (90%)
Hispanic	87 (1.2%)	57 (1.2%)	30 (1.1%)
Unknown	906 (12%)	653 (14%)	253 (9.4%)
Education			
Some Grade School	438 (5.9%)	276 (5.9%)	162 (6.0%)
Some High School	976 (13%)	589 (13%)	387 (14%)
High School Graduate	1,451 (20%)	946 (20%)	505 (19%)
Vocational/Technical School	279 (3.8%)	156 (3.3%)	123 (4.6%)
Some College or Associate's Degree	1,469 (20%)	940 (20%)	529 (20%)
College Graduate	962 (13%)	604 (13%)	358 (13%)
Graduate or Professional School	831 (11%)	514 (11%)	317 (12%)
Other	997 (13%)	675 (14%)	322 (12%)
Smoking Status			
Never Smoker	1,009 (14%)	592 (13%)	417 (15%)
Former Smoker	4,251 (57%)	2,821 (60%)	1,430 (53%)
Current Smoker	1,979 (27%)	1,171 (25%)	808 (30%)
Smoker, Status Unknown	164 (2.2%)	116 (2.5%)	48 (1.8%)
Pack-Years of Smoking	40 (16, 53)	40 (19, 53)	37 (12, 54)
EGFR Mutation			
No	1,255 (17%)	737 (16%)	518 (19%)
Yes	298 (4.0%)	158 (3.4%)	140 (5.2%)
Not Tested	5,850 (79%)	3,805 (81%)	2,045 (76%)
KRAS Mutation			
No	1,148 (16%)	630 (13%)	518 (19%)
Yes	405 (5.5%)	265 (5.6%)	140 (5.2%)
Not Tested	5,850 (79%)	3,805 (81%)	2,045 (76%)

[1] n (%); Median (IQR)

4.4.3 Time-to-Recurrence

In line with our proposed analytic framework, we first calculated the survival function for recurrence based on the joint survival function and the survival function for death under the assumed Clayton copula. We carried this for our study sample, as well as stratified by patient sex (male versus female). The copula dependence parameter, θ, captures the strength of the relationship between progression and death, with larger values corresponding to a higher degree of dependence between these two events. We estimated the value of this parameter using our 'leave-one-in' modification to the extended concordance-based estimator proposed in Fine et al. (2001) [46]. Among all patients in our study, we estimated the dependence between progression and death to be 5.60, corresponding to a Kendall's τ value of 0.737. This suggests a high degree of correlation between progression and death. Further stratified by patient sex, we estimated this dependence to be higher among females (5.93) than males (4.85), corresponding to a τ of 0.748 versus 0.708, respectively.

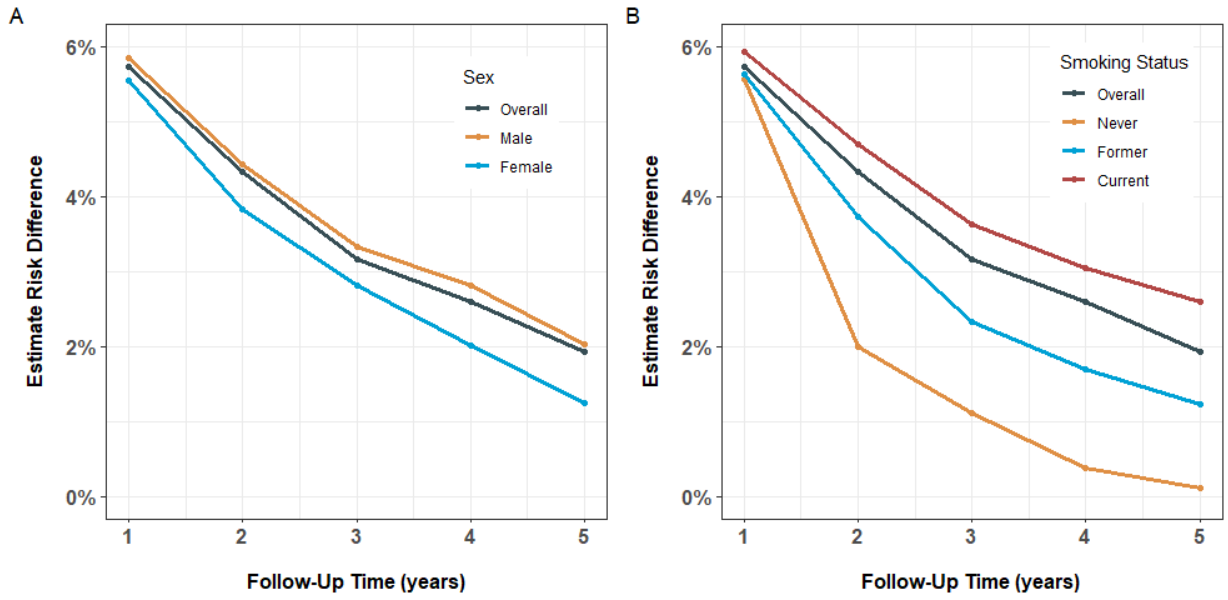
4.4.4 Risk Difference between First-Line Therapies

We calculated pseudo-recurrence probabilities at one-year benchmarks from one- to five-years follow up. We carried forward these pseudo-outcomes to our S-learner, where we estimated the average causal difference in the risk of recurrence between surgery and other first-line treatments overall, and stratified by sex and smoking status. These results are presented in Figure 4.3. As shown, the overall difference in risk of recurrence between first-line therapies was estimated to vary over time, with a 5.7% difference at one year, attenuating to 1.9% after five years. Stratified by patient sex, we see that among male patients, the risk difference is slightly higher, with a one-year difference of 5.9, attenuating to 2.0%, as compared to female patients, among whom we estimated the risk difference to be between 5.6% and 1.3% over five years. Larger differences were observed when stratifying by patient smoking status. As shown, treatment differences were slightly higher among current smokers, ranging from 5.9% to 2.5%, while among former (range: 5.6% to 1.2%) and never smokers (range: 5.6% to 0.1%) these differences were less.

4.5 Discussion

In this work, we propose a deep learning framework for causal inference in time-to-event data with dependent censoring due to semi-competing risks, with a focus on non-fatal events such as time-to-recurrence. We demonstrate the performance of our approach on simulated data and apply it to a real-world dataset from a large epidemiologic lung cancer cohort. Our findings highlight the importance of accounting for semi-competing risks and provide new insights into the causal relationship between first line surgical resection and and the risk of recurrence. As shown, this

Figure 4.3: Estimated average causal difference in the risk of recurrence between surgery and other first-line treatments among patients with stage 1-3A non-small cell lung cancer, over time and (A) stratified by sex; (B) stratified by smoking status



approach provides an accurate method for estimating the causal average treatment effect on the probability of disease recurrence, particularly in settings where the true relationship between the non-fatal outcome, treatment, and other confounding variables is complex.

A specific aim of this study was to focus on the effect of treatment on time to recurrence, rather than alternatives such as overall survival or progression-free survival, for several reasons. First, time to recurrence provides a more precise and clinically meaningful measure of the duration of response to treatment. Time to recurrence measures the time from diagnosis to the point where disease progression is observed, while composites such as progression-free survival measures the time to either disease progression or death. As a result, time to recurrence can more accurately capture the effect of treatment on disease progression, while progression-free survival can be confounded by the effect of treatment on survival. As remaining treatment options are dictated by the monitoring of disease progression, directly studying recurrence is less susceptible to bias than progression-free survival [151, 44].

In the context of the Boston Lung Cancer Study data, we observed differences in the efficacy of surgical resection compared to other first-line therapies, which attenuated over time. While there is limited literature on this topic, several studies suggest that surgical resection has better prognostic outcomes in patients with stage 1-3A NSCLC, particularly in the first five years of follow up [143, 136]. Further, advances in surgical techniques have led to safer, less invasive procedures,

which make surgery an important intervention, potentially in addition to other therapeutic regimens [98]. Additional, we note a modest difference in the effect of surgery versus other first-line therapies when comparing male and female patient subgroups. While previous studies have reported similar rates of recurrence between these sub-populations [82], the timing of recurrence differs [39]. There is also evidence that female patients have a significantly better response to neoadjuvant chemotherapy than male patients [24]. With respect to smoking status, we note many other individualized factors may contribute to greater perceived treatment benefits for current smokers versus former or never smokers, including stage and genetic mutations [33, 110], warranting further study. Further, much of the literature on NSCLC prognosis points to a lack of emphasis on predictors of other clinical endpoints besides overall survival [22]. Namely, research has shown that patients and providers are interested in endpoints such as disease recurrence and response to therapy, which impact quality of life and guide treatment decisions [37].

There are also several open problems and areas of future direction. A primary concern is how to conduct inference in this setting. Our approach yields accurate point estimates for our causal estimand, but we do not yet have a means of quantifying the uncertainty surrounding these estimates. While uncertainty quantification in causal deep learning is still relatively new, it is an important step in developing methods that have practical clinical applicability [2]. Other approaches such as Bayesian neural networks may lead to valid inference for testing for the significance of the causal effect estimates. Further, the implementation is computationally intensive, owing to the intermediate steps needed to calculate the marginal survival functions and pseudo-responses before training our deep neural network. Future work will improve the efficiency of the proposed method. We also consider extending this approach to other useful target values, such as restricted mean survival times. We will address these problems in subsequent work. Overall, however, we demonstrate the performance of this approach on simulated and real-world data, highlighting its ability to accurately estimate the causal effect in the presence of semi-competing risks. Our findings demonstrate the importance of accounting for dependent censoring due to semi-competing risks when estimating the causal effect of treatment on time-to-non fatal events.

APPENDIX A

Technical Details for Chapter 2

A.1 Illness-Death Model Notation

Let T_1 and T_2 denote the times to a non-terminal and terminal event, respectively. Let $\lambda_1(t_1)$ denote the hazard of the non-terminal event at time t_1, $\lambda_2(t_2)$ denote the hazard of the terminal event at t_2 without experiencing the non-terminal event, and $\lambda_3(t_2 \mid t_1)$ denote the hazard of the terminal event at t_2 given the observation of the non-terminal event at $t_1 \leq t_2$. These hazard rates, corresponding to the transitions between states are defined as

$$\lambda_1(t_1) = \lim_{\Delta \to 0} \Pr\left[T_1 \in [t_1, t_1 + \Delta) \mid T_1 \geq t_1, T_2 \geq t_1\right] / \Delta; \quad t_1 > 0 \tag{A.1}$$

$$\lambda_2(t_2) = \lim_{\Delta \to 0} \Pr\left[T_2 \in [t_2, t_2 + \Delta) \mid T_1 \geq t_2, T_2 \geq t_2\right] / \Delta; \quad t_2 > 0 \tag{A.2}$$

$$\lambda_3(t_2 \mid t_1) = \lim_{\Delta \to 0} \Pr\left[T_2 \in [t_2, t_2 + \Delta) \mid T_1 = t_1, T_2 \geq t_2\right] / \Delta; \quad t_2 \geq t_1 > 0. \tag{A.3}$$

Note that the definitions of $\lambda_1(t_1)$ and $\lambda_2(t_2)$ mirror that of the cause-specific hazards under a competing risks framework, where they describe the hazards of first observing either the non-terminal or terminal event. Under semi-competing risks, observing the non-terminal event is subject to observing the terminal event, but not vice-versa. Hence, $\lambda_3(t_2 \mid t_1)$ describes the hazards of observing the terminal event at t_2 after having observed the non-terminal event at t_1. As we cannot observe the non-terminal event after the terminal event has been observed, the space of (T_1, T_2) is restricted to the so-called 'upper wedge' of the first quadrant where $t_1 \leq t_2$, and the non-terminal event is said to be dependently censored by the terminal event. To incorporate this dependence, we model (A.1) - (A.3) by extending the Cox proportional hazards model [34] to a shared gamma-frailty conditional Markov model

$$\lambda_1 (t_1 \mid \gamma, \boldsymbol{X}) = \gamma \lambda_{01} (t_1) \exp \{h_1(\boldsymbol{X})\}; \quad t_1 > 0 \tag{A.4}$$

$$\lambda_2 (t_2 \mid \gamma, \boldsymbol{X}) = \gamma \lambda_{02} (t_2) \exp \{h_2(\boldsymbol{X})\}; \quad t_2 > 0 \tag{A.5}$$

$$\lambda_3 (t_2 \mid t_1, \gamma, \boldsymbol{X}) = \gamma \lambda_{03} (t_2 - t_1) \exp \{h_3(\boldsymbol{X})\}; \quad t_2 \geq t_1 > 0, \tag{A.6}$$

where γ is a random effect, referred to as a subject's frailty, $\lambda_{0g}(t); g \in \{1, 2, 3\}$ are the baseline hazards for the three state transitions, $\boldsymbol{X}$ is a p-vector of covariates, and $h_g(\boldsymbol{X})$ are log-risk functions which relate the covariates to the hazard rates for each potential transition. This model is considered semi-Markov, as the time to the terminal event after having observed the non-terminal event is conditional on the 'sojourn time' between events. Based on (A.4) - (A.5), we can also write out the following survival functions

$$S(t_1, t_1 \mid \gamma) = e^{-\gamma[\Lambda_{01}(t_1) \exp\{h_1(\boldsymbol{X})\} + \Lambda_{02}(t_1) \exp\{h_2(\boldsymbol{X})\}]} \tag{A.7}$$

$$S_{2|1}(t_2 \mid t_1, \gamma) = e^{-\gamma \Lambda_{03}(t_2 - t_1) \exp\{h_3(\boldsymbol{X})\}}; \quad t_2 \geq t_1, \tag{A.8}$$

where $\Lambda_{0g}(t) = \int_0^t \lambda_{0g}(u)du$ are the cumulative baseline hazards for each transition. We can see that the joint survival function conditional on γ and evaluated at $t_2 = t_1$ takes the form in (A.7) due to the competing nature of the state transitions from no event to the first of either the non-terminal or the terminal event. The survival function of t_2 conditional on t_1 and the frailty in (A.8) is defined in terms of transition in the sojourn time between events.

A.2 Conditional Likelihood Function

In practice, we do not fully observe (T_1, T_2), as both events are subject to administrative censoring. Let C denote the censoring time. We observe

$$\mathcal{D} = \{(Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2},\ \boldsymbol{X}_i);\ i = 1, \ldots, n\},$$

where $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = I(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = I(T_{i1} \leq Y_{i2})$, $\boldsymbol{X}_i$ is a p-vector of covariates, $I(\cdot)$ is the indicator function, and i indexes an individual subject in the study, $i = 1, \ldots, n$. There are four potential event progressions we can observe for an individual during a finite period of follow-up (Table A.1). To construct the likelihood conditional on the subject-specific frailties, we multiply the likelihood contributions under each case in Table A.1, raised to the appropriate event indicators, and taken over the n subjects. Define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$

to be the n-vector of latent frailties, and let $\boldsymbol{\psi} = \{\Lambda_{01}, \Lambda_{02}, \Lambda_{03}, \theta\}$ be the collection of model parameters to be learned. The likelihood function is

Table A.1: Potential event progressions and corresponding observed data

	Observed Event Progression		Observed Data			
Case	Non-Terminal	Terminal	Y_{i1}	δ_{i1}	Y_{i2}	δ_{i2}
1	✓	✓	T_{i1}	1	T_{i2}	1
2	✗	✓	T_{i2}	0	T_{i2}	1
3	✓	✗	T_{i1}	1	C_i	0
4	✗	✗	C_i	0	C_i	0

$$
\begin{aligned}
L(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \big[& S(Y_{i1}, Y_{i1}) \times \gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} \times S_{2|1}(Y_{i2} \mid Y_{i1}) \times \\
& \times \gamma_i \lambda_{03}(Y_{i2} - Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}]^{\delta_{i1}\delta_{i2}} \times [S(Y_{i1}, Y_{i1}) \times \gamma_i \lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{X}_i)\}]^{(1-\delta_{i1})\delta_{i2}} \\
& \times \big[S(Y_{i1}, Y_{i1}) \times \gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} \times S_{2|1}(Y_{i2} \mid Y_{i1})\big]^{\delta_{i1}(1-\delta_{i2})} \times [S(Y_{i1}, Y_{i1})]^{(1-\delta_{i1})(1-\delta_{i2})} \\
= \prod_{i=1}^{n} & \{\gamma_i^2 \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} \lambda_{03}(Y_{i2} - Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\} \\
& \times e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\} + \Lambda_{03}(Y_{i2} - Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}]} \Big\}^{\delta_{i1}\delta_{i2}} \\
& \times \Big\{ e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\}]} \\
& \times \gamma_i \lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{X}_i)\} \Big\}^{(1-\delta_{i1})\delta_{i2}} \times \{\gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} \\
& \times e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\} + \Lambda_{03}(Y_{i2} - Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}]} \Big\}^{\delta_{i1}(1-\delta_{i2})} \\
& \times \Big\{ e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\}]} \Big\}^{(1-\delta_{i1})(1-\delta_{i2})} \\
= \prod_{i=1}^{n} & \{\gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\}\}^{\delta_{i1}} \times \{\gamma_i \lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{X}_i)\}\}^{(1-\delta_{i1})\delta_{i2}} \\
& \times \{\gamma_i \lambda_{03}(Y_{i2} - Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}\}^{\delta_{i1}\delta_{i2}} \\
& \times \Big\{ e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\}]} \Big\}^{(1-\delta_{i1})} \\
& \times \Big\{ e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\} + \Lambda_{03}(Y_{i2} - Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}]} \Big\}^{\delta_{i1}}
\end{aligned}
$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} \gamma_i^{\delta_{i1}+\delta_{i2}} \times \{\lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i)\}\}^{\delta_{i1}} \\
&\quad \times \{\lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{X}_i)\}\}^{(1-\delta_{i1})\delta_{i2}} \times \{\lambda_{03}(Y_{i2}-Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}\}^{\delta_{i1}\delta_{i2}} \\
&\quad \times \left\{ e^{-\gamma_i [\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{X}_i\}+\Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{X}_i)\}+\delta_{i1}\Lambda_{03}(Y_{i2}-Y_{i1}) \exp\{h_3(\boldsymbol{X}_i)\}]} \right\}
\end{aligned} \tag{A.9}
$$

A.3 Likelihood Function

In this section, we provide the necessary steps to derive our objective function, namely the marginal data log-likelihood function under the assumed illness-death model. Note that, as given in Table A.1, the observable times for each patient, Y_{i1} and Y_{i2}, arise from T_{i1}, T_{i2}, and C_i in one of four cases: (1) the patient experiences both the non-terminal and the terminal event, (2) the patient experiences only the terminal event, (3) the patient experience only the non-terminal event, or (4) the patient experiences neither event. The likelihood contribution for a given patient under each of these four cases can be derived, starting with the hazard rates defined in (2.4) - (2.6) and utilizing the joint density of the event times as follows.

Case 1: We first derive the likelihood contribution for an individual who experiences both events. Conditional on one's frailty, (γ), this is given by

$$
\begin{aligned}
&\exp\left\{-\int_0^{y_1} [\lambda_1(s)+\lambda_2(s)]\, ds\right\} \lambda_1(y_1) \times \exp\left\{-\int_0^{y_2-y_1} \lambda_3(s) ds\right\} \lambda_3(y_2-y_1) \\
&\quad = \exp\left\{-\int_0^{y_1} [\gamma\lambda_{01}(s) \exp\{h_1(x)\} + \gamma\lambda_{02}(s) \exp\{h_2(x)\}]\, ds\right\} \times \gamma\lambda_{01}(y_1) \exp\{h_1(x)\} \\
&\qquad \times \exp\left\{-\int_0^{y_2-y_1} \gamma\lambda_{03}(s) \exp\{h_3(x)\}\, ds\right\} \gamma\lambda_{03}(y_2-y_1) \exp\{h_3(x)\} \\
&\quad = \exp\{-\gamma [\exp\{h_1(x)\} \Lambda_{01}(y_1) + \exp\{h_2(x)\} \Lambda_{02}(y_1)]\} \gamma\lambda_{01}(y_1) \exp\{h_1(x)\} \\
&\qquad \times \exp\{-\gamma \exp\{h_3(x)\} \Lambda_{03}(y_2-y_1)\} \gamma\lambda_{03}(y_2-y_1) \exp\{h_3(x)\} \\
&\quad = \exp\left\{-\gamma \left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2-y_1)\right]\right\} \\
&\qquad \times \gamma e^{h_1(x)} \lambda_{01}(y_1) \times \gamma e^{h_3(x)} \lambda_{03}(y_2-y_1),
\end{aligned}
$$

where $\Lambda_{01}(t) = \int_0^t \lambda_{01}(s) ds$, $\Lambda_{02}(t) = \int_0^t \lambda_{02}(s) ds$, and $\Lambda_{03}(s,t) = \Lambda_{03}(t) - \Lambda_{03}(s)$ denote the cumulative conditional hazard functions. Marginalizing over γ, we take

$$\int_0^\infty \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2 - y_1)\right]\right\}$$
$$\times \gamma e^{h_1(x)}\lambda_{01}(y_1) \times \gamma e^{h_3(x)}\lambda_{03}(y_2 - y_1) \times \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma.$$

For simplicity, let

$$A = e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2 - y_1)$$
$$B = e^{h_1(x)}\lambda_{01}(y_1) \times e^{h_3(x)}\lambda_{03}(y_2 - y_1),$$

such that

$$= \int_0^\infty \gamma^2 B \times \exp\{-\gamma A\} \times \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma$$
$$= B \times \left(\frac{1/\theta}{1/\theta + A}\right)^{1/\theta} \times \int_0^\infty \gamma^2 \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\left(\frac{1}{\theta} + A\right)^{1/\theta}\gamma^{1/\theta-1}\exp\left\{-\gamma\left(\frac{1}{\theta} + A\right)\right\}d\gamma.$$

Recognizing this as the second moment of a Gamma random variable, this expression reduces to

$$= B \times \left(\frac{1/\theta}{1/\theta + A}\right)^{1/\theta} \times \left\{\mathrm{Var}[\gamma] + \mathbb{E}[\gamma]^2\right\} = B \times \left(\frac{1/\theta}{1/\theta + A}\right)^{1/\theta} \times \left\{\frac{1/\theta}{(1/\theta + A)^2} + \left(\frac{1/\theta}{1/\theta + A}\right)^2\right\}$$
$$= B \times \left(\frac{1}{1 + \theta A}\right)^{1/\theta} \times \frac{1 + \theta}{(1 + \theta A)^2} = B \times (1 + \theta) \times (1 + \theta A)^{-1/\theta - 2}.$$

Thus, the likelihood contribution for an individual under Case 1 is

$$e^{h_1(x)}\lambda_{01}(y_1) \times e^{h_3(x)}\lambda_{03}(y_2 - y_1) \times (1 + \theta)$$
$$\times \left\{1 + \theta\left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2 - y_1)\right]\right\}^{-1/\theta - 2}. \tag{A.10}$$

Case 2: Next, we derive the likelihood contribution for an individual who experiences just the terminal event. Conditional on γ, is given by

$$\exp\left\{-\int_0^{y_1}\left[\lambda_1(s) + \lambda_2(s)\right]ds\right\}\lambda_2(y_1)$$
$$= \exp\left\{-\int_0^{y_1}\left[\gamma\lambda_{01}(s)\exp\{h_1(x)\} + \gamma\lambda_{02}(s)\exp\{h_2(x)\}\right]ds\right\} \times \gamma\lambda_{02}(y_1)\exp\{h_2(x)\}$$
$$= \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1)\right]\right\} \times \gamma e^{h_2(x)}\lambda_{02}(y_1).$$

Marginalizing over γ, we take

$$\int_0^\infty \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}(y_1)+e^{h_2(x)}\Lambda_{02}(y_1)\right]\right\}\times\gamma e^{h_2(x)}\lambda_{02}(y_1)$$
$$\times\frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma.$$

For simplicity, let

$$A = e^{h_1(x)}\Lambda_{01}(y_1)+e^{h_2(x)}\Lambda_{02}(y_1)$$
$$B = e^{h_2(x)}\lambda_{02}(y_1),$$

such that

$$= \int_0^\infty \gamma B\times\exp\{-\gamma A\}\times\frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma$$
$$= B\times\left(\frac{1/\theta}{1/\theta+A}\right)^{1/\theta}\times\int_0^\infty\gamma\frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\left(\frac{1}{\theta}+A\right)^{1/\theta}\gamma^{1/\theta-1}\exp\left\{-\gamma\left(\frac{1}{\theta}+A\right)\right\}d\gamma.$$

Recognizing this as the first moment of a Gamma random variable, this expression reduces to

$$= B\times\left(\frac{1/\theta}{1/\theta+A}\right)^{1/\theta}\times\mathbb{E}[\gamma] = B\times\left(\frac{1}{1+\theta A}\right)^{1/\theta}\times\frac{1}{1+\theta A} = B\times(1+\theta A)^{-1/\theta-1}.$$

Thus, the likelihood contribution for an individual under Case 2 is

$$e^{h_2(x)}\lambda_{02}(y_1)\times\left\{1+\theta\left[e^{h_1(x)}\Lambda_{01}(y_1)+e^{h_2(x)}\Lambda_{02}(y_1)\right]\right\}^{-1/\theta-1}. \tag{A.11}$$

Case 3: Next, we derive the likelihood contribution for an individual who experiences only the non-terminal event. Conditional on γ, this is given by

$$\exp\left\{-\int_0^{y_1}[\lambda_1(s)+\lambda_2(s)]\,ds\right\}\lambda_1(y_1)\times\exp\left\{-\int_0^{y_2-y_1}\lambda_3(s)ds\right\}$$
$$= \exp\left\{-\int_0^{y_1}[\gamma\lambda_{01}(s)\exp\{h_1(x)\}+\gamma\lambda_{02}(s)\exp\{h_2(x)\}]\,ds\right\}\times\gamma\lambda_{01}(y_1)\exp\{h_1(x)\}$$
$$\times\exp\left\{-\int_0^{y_2-y_1}\gamma\lambda_{03}(s)\exp\{h_3(x)\}\,ds\right\}$$
$$= \exp\{-\gamma[\exp\{h_1(x)\}\Lambda_{01}(y_1)+\exp\{h_2(x)\}\Lambda_{02}(y_1)]\}\gamma\lambda_{01}(y_1)$$
$$\times\exp\{h_1(x)\}\exp\{-\gamma\exp\{h_3(x)\}\Lambda_{03}(y_2-y_1)\}$$

$$= \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2 - y_1)\right]\right\} \times \gamma e^{h_1(x)}\lambda_{01}(y_1).$$

Marginalizing this over γ, we take

$$\int_0^\infty \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2 - y_1)\right]\right\} \times \gamma e^{h_1(x)}\lambda_{01}(y_1) \times \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma.$$

For simplicity, let

$$\begin{aligned} A &= e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1) + e^{h_3(x)}\Lambda_{03}(y_2 - y_1) \\ B &= e^{h_1(x)}\lambda_{01}(y_1), \end{aligned}$$

such that

$$\begin{aligned} &= \int_0^\infty \gamma B \times \exp\{-\gamma A\} \times \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma \\ &= B \times \left(\frac{1/\theta}{1/\theta + A}\right)^{1/\theta} \times \int_0^\infty \gamma \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\left(\frac{1}{\theta} + A\right)^{1/\theta}\gamma^{1/\theta-1}\exp\left\{-\gamma\left(\frac{1}{\theta} + A\right)\right\}d\gamma. \end{aligned}$$

Recognizing this as the first moment of a Gamma random variable, this expression reduces to:

$$= B \times \left(\frac{1/\theta}{1/\theta + A}\right)^{1/\theta} \times \mathbb{E}[\gamma] = B \times \left(\frac{1}{1 + \theta A}\right)^{1/\theta} \times \frac{1}{1 + \theta A} = B \times (1 + \theta A)^{-1/\theta - 1}.$$

Thus, the likelihood contribution for an individual under Case 3 is

$$\begin{aligned} &e^{h_1(x)}\lambda_{01}(y_1) \times \left\{1 + \theta\left[e^{h_1(x)}\Lambda_{01}(y_1) + e^{h_2(x)}\Lambda_{02}(y_1)\right.\right. \\ &\quad \left.\left. + e^{h_3(x)}\Lambda_{03}(y_2 - y_1)\right]\right\}^{-1/\theta - 1}. \end{aligned} \tag{A.12}$$

Case 4: Finally, we derive the likelihood contribution for an individual who experiences neither event. Conditional on γ, this is given by

$$\exp\left\{-\int_0^{y_1}\left[\lambda_1(s) + \lambda_2(s)\right]ds\right\}\exp\left\{-\int_0^{y_1}\left[\gamma\lambda_{01}(s)\exp\{h_1(x)\} + \gamma\lambda_{02}(s)\exp\{h_2(x)\}\right]ds\right\}$$

$$= \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}\left(y_1\right) + e^{h_2(x)}\Lambda_{02}\left(y_1\right)\right]\right\}.$$

Marginalizing this over γ, we take

$$\int_0^\infty \exp\left\{-\gamma\left[e^{h_1(x)}\Lambda_{01}\left(y_1\right) + e^{h_2(x)}\Lambda_{02}\left(y_1\right)\right]\right\} \times \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma.$$

For simplicity, let $A = e^{h_1(x)}\Lambda_{01}\left(y_1\right) + e^{h_2(x)}\Lambda_{02}\left(y_1\right)$, such that

$$\begin{aligned}
&= \int_0^\infty \exp\left\{-\gamma A\right\} \times \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\theta^{-\frac{1}{\theta}}\gamma^{\frac{1}{\theta}-1}\exp\left\{-\frac{\gamma}{\theta}\right\}d\gamma \\
&= \left(\frac{1/\theta}{1/\theta + A}\right)^{1/\theta} \times \int_0^\infty \frac{1}{\Gamma\left(\frac{1}{\theta}\right)}\left(\frac{1}{\theta} + A\right)^{1/\theta}\gamma^{1/\theta-1}\exp\left\{-\gamma\left(\frac{1}{\theta} + A\right)\right\}d\gamma.
\end{aligned}$$

Recognizing this as that of a Gamma random variable's density function take over its support, this reduces to $(1 + \theta A)^{-1/\theta}$. Thus, the likelihood contribution for an individual under Case 4 is

$$\left\{1 + \theta\left[e^{h_1(x)}\Lambda_{01}\left(y_1\right) + e^{h_2(x)}\Lambda_{02}\left(y_1\right)\right]\right\}^{-1/\theta}. \tag{A.13}$$

Likelihood: Given the likelihood contributions under each of the four cases derived above and our event indicators, δ_1 and δ_2 as denoted in Table A.1, we can write out the full likelihood contribution of the ith sample individual as follows

$$\begin{aligned}
\mathcal{L}_i &= \text{Case}_1(y_{i1})^{\delta_{i1}\delta_{i2}} \times \text{Case}_2(y_{i1})^{(1-\delta_{i1})\delta_{i2}} \times \text{Case}_3(y_{i1})^{\delta_{i1}(1-\delta_{i2})} \times \text{Case}_4(y_{i1})^{(1-\delta_{i1})(1-\delta_{i2})} \\
&= \lambda_{01}\left(y_{i1}\right)^{\delta_{i1}}\lambda_{02}\left(y_{i2}\right)^{\delta_{i2}(1-\delta_{i1})}\lambda_{03}\left(y_{i2}\right)^{\delta_{i1}\delta_{i2}} \\
&\quad \times \exp\left\{\delta_{i1}\cdot h_1(x_i) + \delta_{i2}\left(1-\delta_{i1}\right)\cdot h_2(x_i) + \delta_{i1}\delta_{i2}\cdot h_3(x_i)\right\} \times (1+\theta)^{\delta_{i1}\delta_{i2}} \\
&\quad \times \left\{1 + \theta\left[\Lambda_{01}\left(y_{i1}\right)e^{h_1(x_i)} + \Lambda_{02}\left(y_{i1}\right)e^{h_2(x_i)} + \Lambda_{03}\left(y_{i2}-y_{i1}\right)e^{h_3(x_i)}\right]\right\}^{-\frac{1}{\theta}-\delta_{i1}-\delta_{i2}}.
\end{aligned}$$

The full likelihood is then the product of each individual's likelihood contribution

$$\begin{aligned}
\mathcal{L} = \prod_{i=1}^n \mathcal{L}_i &= \prod_{i=1}^n \lambda_{01}\left(y_{i1}\right)^{\delta_{i1}}\lambda_{02}\left(y_{i2}\right)^{\delta_{i2}(1-\delta_{i1})}\lambda_{03}\left(y_{i2}\right)^{\delta_{i1}\delta_{i2}} \\
&\quad \times \exp\left\{\delta_{i1}\cdot h_1(x_i) + \delta_{i2}\left(1-\delta_{i1}\right)\cdot h_2(x_i) + \ \delta_{i1}\delta_{i2}\cdot h_3(x_i)\right\} \times (1+\theta)^{\delta_{i1}\delta_{i2}} \\
&\quad \times \left\{1 + \theta\left[\Lambda_{01}\left(y_{i1}\right)e^{h_1(x_i)} + \Lambda_{02}\left(y_{i1}\right)e^{h_2(x_i)} + \ \Lambda_{03}\left(y_{i2}-y_{i1}\right)e^{h_3(x_i)}\right]\right\}^{-\frac{1}{\theta}-\delta_{i1}-\delta_{i2}}
\end{aligned} \tag{A.14}$$

A.4 Objective Function under Weibull Baseline Hazards

In order to fully define our objective function, we must specify a form for the baseline hazards. We assume the baseline hazards follow a Weibull(ϕ_1, ϕ_2) distribution, with

$$\lambda_{0g}(t) = \phi_{g1}\phi_{g2}t^{\phi_{g1}-1}; \quad \Lambda_{0g}(t) = \phi_{g2}t^{\phi_{g1}}; g \in \{1, 2, 3\}.$$

We can then specify the full likelihood function as

$$\begin{aligned}
\mathcal{L} = \prod_{i=1}^{n} &\left(\phi_{11}\phi_{12}y_{i1}^{\phi_{11}-1}\right)^{\delta_{i1}} \left(\phi_{21}\phi_{22}y_{i2}^{\phi_{21}-1}\right)^{\delta_{i2}(1-\delta_{i1})} \left(\phi_{31}\phi_{32}y_{i2}^{\phi_{31}-1}\right)^{\delta_{i1}\delta_{i2}} \\
&\times \exp\left\{\delta_{i1}\cdot h_1(x_i) + \delta_{i2}\left(1-\delta_{i1}\right)\cdot h_2(x_i) + \delta_{i1}\delta_{i2}\cdot h_3(x_i)\right\} \\
&\times (1+\theta)^{\delta_{i1}\delta_{i2}} \times \left\{1+\theta\left[\phi_{12}y_{i1}^{\phi_{11}}e^{h_1(x_i)} + \phi_{22}y_{i1}^{\phi_{21}}e^{h_2(x_i)} + \phi_{32}(y_{i2}-y_{i1})^{\phi_{31}}e^{h_3(x_i)}\right]\right\}^{-\frac{1}{\theta}-\delta_{i1}-\delta_{i2}},
\end{aligned}$$

and the log-likelihood function as:

$$\begin{aligned}
\ell = \log[\mathcal{L}] = \log\Bigg[\prod_{i=1}^{n} &\left(\phi_{11}\phi_{12}y_{i1}^{\phi_{11}-1}\right)^{\delta_{i1}} \left(\phi_{21}\phi_{22}y_{i2}^{\phi_{21}-1}\right)^{\delta_{i2}(1-\delta_{i1})} \left(\phi_{31}\phi_{32}y_{i2}^{\phi_{31}-1}\right)^{\delta_{i1}\delta_{i2}} \\
&\times \exp\left\{\delta_{i1}\cdot h_1(x_i) + \delta_{i2}\left(1-\delta_{i1}\right)\cdot h_2(x_i) + \delta_{i1}\delta_{i2}\cdot h_3(x_i)\right\} \\
&\times (1+\theta)^{\delta_{i1}\delta_{i2}} \times \left\{1+\theta\left[\phi_{12}y_{i1}^{\phi_{11}}e^{h_1(x_i)} + \phi_{22}y_{i1}^{\phi_{21}}e^{h_2(x_i)} + \phi_{32}(y_{i2}-y_{i1})^{\phi_{31}}e^{h_3(x_i)}\right]\right\}^{-\frac{1}{\theta}-\delta_{i1}-\delta_{i2}}\Bigg] \\
= \sum_{i=1}^{n} &\delta_{i1}\cdot\left[\log(\phi_{11}) + \log(\phi_{12}) + (\phi_{11}-1)\cdot\log(y_{i1})\right] + \delta_{i2}\left(1-\delta_{i1}\right)\cdot\left[\log(\phi_{21}) + \log(\phi_{22})\right. \\
&\left. + (\phi_{21}-1)\cdot\log(y_{i2})\right] + \delta_{i1}\delta_{i2}\cdot\left[\log(\phi_{31}) + \log(\phi_{32}) + (\phi_{31}-1)\cdot\log(y_{i2})\right] \\
&+ \ \delta_{i1}\cdot h_1(x_i) + \delta_{i2}\left(1-\delta_{i1}\right)\cdot h_2(x_i) + \delta_{i1}\delta_{i2}\cdot h_3(x_i) + \delta_{i1}\delta_{i2}\cdot(1+\theta) \\
&- \left(\frac{1}{\theta}+\delta_{i1}+\delta_{i2}\right)\cdot\log\left\{1+\theta\left[\phi_{12}y_{i1}^{\phi_{11}}e^{h_1(x_i)} + \phi_{22}y_{i1}^{\phi_{21}}e^{h_2(x_i)} + \phi_{32}(y_{i2}-y_{i1})^{\phi_{31}}e^{h_3(x_i)}\right]\right\}.
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
\ell = \sum_{i=1}^{n} &\delta_{i1}\cdot\left[\log(\phi_{11}) + \log(\phi_{12}) + (\phi_{11}-1)\cdot\log(y_{i1}) + h_1(x_i)\right] \\
&+ \delta_{i2}\left(1-\delta_{i1}\right)\cdot\left[\log(\phi_{21}) + \log(\phi_{22}) + (\phi_{21}-1)\cdot\log(y_{i2}) + h_2(x_i)\right] \\
&+ \delta_{i1}\delta_{i2}\cdot\left[\log(\phi_{31}) + \log(\phi_{32}) + (\phi_{31}-1)\cdot\log(y_{i2}) + h_3(x_i) + \log(1+\theta)\right] \\
&- \left(\frac{1}{\theta}+\delta_{i1}+\delta_{i2}\right)\log\left\{1+\theta\left[\phi_{12}y_{i1}^{\phi_{11}}e^{h_1(x_i)} + \phi_{22}y_{i1}^{\phi_{21}}e^{h_2(x_i)} + \phi_{32}(y_{i2}-y_{i1})^{\phi_{31}}e^{h_3(x_i)}\right]\right\}
\end{aligned} \tag{A.15}$$

A.5 Bivariate Brier Score

A.5.1 Brier Score with Fully-Observed Data

First, consider a single event of interest, and let T_i denote the event time for individual i; $i = 1, \ldots, n$. Let $f_i(t)$ denote the corresponding density function, such that

$$S_i(t) = \Pr(T_i > t) = \int_t^\infty f_i(u)du$$

is the survival function for the ith individual at time t. Let $\pi_i(t)$ denote a valid estimate for the survival function. Assuming the event times are fully observed, define the Brier Score as

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[I\left\{T_i > t\right\} - \pi_i(t)\right]^2,$$

where $I\{\cdot\}$ is the indicator function. This provides an approximation to the true, unknown survival functions through step functions with jumps at event times. As the survival functions must be approximated by $I\left\{T_i > t\right\}$, we can see that

$$\begin{aligned}
\mathbb{E}\left[BS(t)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[I\left\{T_i > t\right\} - \pi_i(t)\right]^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\pi_i(t)^2 I\left\{T_i \le t\right\} + \left\{1-\pi_i(t)\right\}^2 I\left\{T_i > t\right\}\right]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\pi_i(t)^2 I\left\{T_i \le t\right\} + \left\{1-\pi_i(t)\right\}^2 I\left\{T_i > t\right\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2\Pr\left(T_i \le t\right) + \left[1-\pi_i(t)\right]^2\Pr\left(T_i > t\right) = \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2\left[1-S_i(t)\right] + \left[1-\pi_i(t)\right]^2 S_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left[S_i(t)-\pi_i(t)\right]^2 + S_i(t)\left[1-S_i(t)\right]\right\} = MSE(t) + \frac{1}{n}\sum_{i=1}^{n}\left\{S_i(t)\left[1-S_i(t)\right]\right\},
\end{aligned}$$

where the additional piece is constant with respect to $\pi_i(t)$ and represents the irreducible error in approximating $S_i(t)$ by these step functions.

A.5.2 Brier Score with Right Censoring

In practice, it is often the case that not all events are observed. In this situation, it is known that the event time, T_i, occurs after some censoring time, C_i. We observe $\mathcal{D} = \{(Y_i, \delta_i);\ i = 1, \ldots, n\}$, where $Y_i = \min\{T_i, C_i\}$ is the observation time and $\delta_i = I\{T_i \le C_i\}$ is the event indicator for the

ith individual. With censoring, we must adopt inverse probability of censoring weighting (IPCW) ([56, 50]). The IPCW-approximated Brier Score is

$$BS_C(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\pi_i(t)^2 \cdot I\{Y_i \le t, \delta_i = 1\}}{G_i(Y_{i-})} + \frac{\{1-\pi_i(t)^2\} \cdot I\{Y_i > t\}}{G_i(t)}\right],$$

where $G_i(t) = \Pr(C_i > t) > 0$ is the survival function of the censoring distribution for the ith individual. With $G_i(t)$ known, note that the expectation of the IPCW-approximated Brier Score is equivalent to that of the Brier Score without censoring

$$\begin{aligned}
\mathbb{E}\left[\mathrm{BS}_C(t)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\pi_i(t)^2 I\{Y_i \le t, \delta_i = 1\}}{G_i(Y_i)} + \frac{[1-\pi_i(t)]^2\, 1\{Y_i > t\}}{G_i(t)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2 \cdot \mathbb{E}\left[\frac{1\{T_i \le t, T_i \le C_i\}}{G_i(Y_i)}\right] + [1-\pi_i(t)]^2\,\frac{\Pr(T_i > t, C_i > t)}{G_i(t)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2 \int_0^t \frac{G_i(u-)f_i(u)}{G_i(u-)}du + [1-\pi_i(t)]^2\,\frac{G_i(t)S_i(t)}{G_i(t)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2\,[1-S_i(t)] + [1-\pi_i(t)]^2\,S_i(t) \\
&= \mathrm{MSE}(t) + \frac{1}{n}\sum_{i=1}^{n}S_i(t)\,[1-S_i(t)]\,.
\end{aligned}$$

A.5.3 Bivariate Brier Score with Right Censoring

Following the framework outlined in the previous sections, we now provide additional detail on the derivation of the Bivariate Brier Score outlined in Section 2.4. As described previously, consider two events of interest which form a semi-competing relationship. Let T_{i1} and T_{i2} denote the non-terminal and terminal event times, respectively, and C_i the censoring time for the ith individual. We observe $\mathcal{D} = \{(Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2});\ i = 1, \ldots, n\}$ where $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = I(T_{i2} \le C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = I(T_{i1} \le Y_{i2})$, and $I(\cdot)$ denotes the indicator function.

We show that the expectation of the Bivariate Brier Score is equal to the mean squared error of the predictor, $\pi_i(t)$, plus a constant. To proceed, we compute the expectation in additive pieces. In the first piece, we consider the region where at least the non-terminal event is observed by time t, and Y_{i1} is less than or equal to Y_{i2}, but the terminal event may or may not be observed. In the second piece, we consider the region where the terminal event is observed prior to the non-terminal event. In the third piece, we consider the region where neither event has been observed by time t.

Piece 1: At least the non-terminal event is observed by time t, and Y_{i1} is less than or equal to Y_{i2},

but the terminal event may or may not be observed.

$$\begin{aligned}
&\mathbb{E}\left[\frac{\pi_i(t)^2 \times I\left(Y_{i1} \leq t,\ \delta_{i1} = 1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i1})}\right] = \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq C_i,\ T_{i1} \leq T_{i2}\right\}}{G_i(T_{i1})}\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq C_i,\ T_{i1} \leq T_{i2}\right\}}{G_i(Y_{i1})} \mid T_{i1}, T_{i2}\right]\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right\}}{G_i(T_{i1})} \times \mathbb{E}\left[I\left(T_{i1} \leq C_i\right\} \mid T_{i1}, T_{i2}\right]\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right\}}{G_i(T_{i1})} \times \Pr\left(T_{i1} \leq C_i\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right\}}{G_i(T_{i1})} \times G_i\left(T_{i1}\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right\}\right] = \pi_i(t)^2 \times \Pr\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right).
\end{aligned}$$

Piece 2: The terminal event is observed prior to the non-terminal event occurring.

$$\begin{aligned}
&\mathbb{E}\left[\frac{\pi_i(t)^2 \times I\left(Y_{i1} \leq t,\ Y_{i2} \leq t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i2})}\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2},\ T_{i2} \leq C_i\right\}}{G_i(Y_{i2})}\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2},\ T_{i2} \leq C_i\right\}}{G_i(Y_{i2})} \mid T_{i1}, T_{i2}\right]\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right\}}{G_i(Y_{i2})} \times \mathbb{E}\left[I\left(T_{i2} \leq C_i\right\} \mid T_{i1}, T_{i2}\right]\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right\}}{G_i(Y_{i2})} \times \Pr\left(T_{i2} \leq C_i\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right\}}{G_i(Y_{i2})} \times G_i\left(T_{i2}\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right\}\right] = \pi_i(t)^2 \times \Pr\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right).
\end{aligned}$$

Piece 3: Neither event has been observed by time t.

$$\begin{aligned}
&\mathbb{E}\left[\frac{[1-\pi_i(t)]^2 \times I\left(Y_{i1} > t,\ Y_{i2} > t\right\}}{G_i(t)}\right] = \frac{[1-\pi_i(t)]^2}{G_i(t)} \times \mathbb{E}\left[I\left(T_{i1} > t,\ T_{i2} > t,\ C_i > t\right\}\right] \\
&= \frac{[1-\pi_i(t)]^2}{G_i(t)} \times \Pr\left(T_{i1} > t,\ T_{i2} > t,\ C_i > t\right) \\
&= \frac{[1-\pi_i(t)]^2}{G_i(t)} \times \Pr\left(T_{i1} > t,\ T_{i2} > t\right) \times \Pr\left(C_i > t\right) \\
&= \frac{[1-\pi_i(t)]^2}{G_i(t)} \times S_i\left(t, t\right) \times G_i\left(t\right) = [1-\pi_i(t)]^2 \times S_i\left(t, t\right).
\end{aligned}$$

Combining these pieces, and summing over the n individuals, we can see that

$$\begin{aligned}
\mathbb{E}\left[BBS_c(t)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \leq t,\ \delta_{i1}=1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i1})}\right. \\
&\quad + \frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \leq t,\ Y_{i2} \leq t,\ \delta_{i1}=0,\ \delta_{i2}=1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i2})} \\
&\quad \left. + \frac{\left[1-\pi_i(t)\right]^2 \cdot I\left(Y_{i1} > t,\ Y_{i2} > t\right\}}{G_i(t)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \leq t,\ \delta_{i1}=1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i1})}\right] \\
&\quad + \mathbb{E}\left[\frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \leq t,\ Y_{i2} \leq t,\ \delta_{i1}=0,\ \delta_{i2}=1,\ Y_{i1} \leq Y_{i2}\right\}}{G_i(Y_{i2})}\right] \\
&\quad + \mathbb{E}\left[\frac{\left[1-\pi_i(t)\right]^2 \cdot I\left(Y_{i1} > t,\ Y_{i2} > t\right\}}{G_i(t)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2 \cdot \Pr\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right) \\
&\quad + \pi_i(t)^2 \cdot \Pr\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right) + \left[1-\pi_i(t)\right]^2 \cdot S_i\left(t,t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\pi_i(t)^2 \cdot \left[1-S_i(t)\right] + \left[1-\pi_i(t)\right]^2 \cdot S_i\left(t,t\right) \\
&= \mathrm{MSE}(t) + \frac{1}{n}\sum_{i=1}^{n}S_i\left(t,\ t\right) \cdot \left[1-S_i\left(t,\ t\right)\right].
\end{aligned}$$

In expectation, the Bivariate Brier Score is equivalent to the mean squared error of the predictor, $\pi_i(t)$, plus an additional piece that is constant with respect to $\pi_i(t)$. This additional term represents the irreducible error incurred by approximating $S_i(t)$ by step functions, $I\left(Y_{i1} > t, Y_{i2} > t\right)$.

APPENDIX B

Technical Details for Chapter 3

B.1 Neural Expectation-Maximization Algorithm

In the following, we provide additional detail on the neural expectation-maximization algorithm outlined in Section 3.3. Viewing the subject-specific frailties as a latent, random effects, the algorithm iterates between three steps, namely the expectation (E) step, the maximization (M) step, and the neural (N) step. In the E-step, the frailties are estimated given the data and current values for the model parameters by taking the expectation of the augmented conditional log-likelihood in (3.1). In the M-step, the model parameters corresponding to the baseline hazard functions are estimated by maximizing the expected log-likelihood in the E-step, given the current estimates for the frailties. Then, fixing these quantities, the log risk functions for a patient's covariates and the population frailty variance are updated as outputs of the neural network architectures in the N-step.

B.1.1 Conditional Frailty Distribution

The conditional density of $\boldsymbol{\gamma}$ given the data is proportional to the product of the conditional likelihood, $L\left(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}\right)$, which was derived in Appendix A, and the marginal density of $\boldsymbol{\gamma}$ by Bayes rule. We assume that each γ_i independently follows a Gamma distribution with a density function $f(\gamma_i) = \frac{\theta^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})}\gamma_i^{\frac{1}{\theta}-1}e^{-\frac{\gamma_i}{\theta}}$ so that $\mathbb{E}[\gamma_i] = 1$ and $\text{Var}(\gamma_i) = \theta$. The marginal density of $\boldsymbol{\gamma}$ is the product over the n independent γ_i densities. Thus, for a fixed value of θ, the posterior distribution of $\boldsymbol{\gamma}$ is

$$
\begin{aligned}
f(\boldsymbol{\gamma}|\mathcal{D}, \boldsymbol{\psi}) \propto f(\boldsymbol{\gamma}) \times L\left(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}\right) = \prod_{i=1}^{n} \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \\
\times \left[\lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)}\right]^{\delta_{i1}} \times \left[\lambda_{02}\left(Y_{i2}\right) e^{h_2(\boldsymbol{x}_i)}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\lambda_{03}\left(Y_{i2} - Y_{i1}\right) e^{h_3(\boldsymbol{x}_i)}\right]^{\delta_{i1}\delta_{i2}} \\
\times \exp\left\{-\gamma_i \left[\Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}\left(Y_{i2} - Y_{i1}\right) e^{h_3(\boldsymbol{x}_i)}\right]\right\}.
\end{aligned}
$$

Considering only the terms which involve γ_i, we can reduce the above expression to

$$f(\boldsymbol{\gamma}|\mathcal{D},\boldsymbol{\psi}) \propto \prod_{i=1}^{n} \gamma_i^{\frac{1}{\theta}+\delta_{i1}+\delta_{i2}-1} \times \exp\left\{-\gamma_i\left[\frac{1}{\theta} + \Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}\left(Y_{i2}-Y_{i1}\right) e^{h_3(\boldsymbol{x}_i)}\right]\right\}.$$

which we recognize to also be the kernel of a Gamma distribution, apart from a constant. Conditional on the data, the γ_i's follow a Gamma$(\tilde{a},\ \tilde{b})$ distribution with

$$\begin{aligned}\tilde{a} &= \frac{1}{\theta} + \delta_{i1} + \delta_{i2},\\ \tilde{b} &= \frac{1}{\theta} + \Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \ \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}\left(Y_{i2}-Y_{i1}\right) e^{h_3(\boldsymbol{x}_i)}.\end{aligned}$$

From this result, we have that the posterior means of the γ_i are given by $\mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}] = \tilde{a}/\tilde{b}$. The posterior means of $\log(\gamma_i)$ can also be derived. Without loss of generality, let the rate parameter, $\tilde{b}$, equal 1, as its effect on the logarithm of γ_i is a negative linear shift by a factor of $\log(\tilde{b})$. Thus, the density of $\gamma_i \sim$ Gamma$(\tilde{a}, 1)$ is given by

$$f(\gamma_i|\mathcal{D},\boldsymbol{\psi}) = \frac{1}{\Gamma(\tilde{a})}\gamma_i^{\tilde{a}-1}\exp\{-\gamma_i\}d\gamma_i = \frac{1}{\Gamma(\tilde{a})}\gamma_i^{\tilde{a}}\exp\{-\gamma_i\}\frac{d\gamma_i}{\gamma_i}.$$

Substituting $\gamma_i = \exp\{\log(\gamma_i)\}$ and noting that $d\gamma_i/\gamma_i = d\log(\gamma_i)$, we have that

$$f(\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}) = \frac{1}{\Gamma(\tilde{a})}\exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\}d\log(\gamma_i).$$

As $f(\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi})$ is a probability density function, and therefore must integrate to unity, and the support of $\log(\gamma_i)$ is in $\mathbb{R}$, we have that

$$\Gamma(\tilde{a}) = \int_{\mathbb{R}} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\}d\log(\gamma_i).$$

Differentiating under the integral with respect to $\tilde{a}$, we have that

$$\begin{aligned}&\frac{\partial}{\partial\tilde{a}}\left[\exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\}d\log(\gamma_i)\right]\\ &\quad= \log(\gamma_i)\exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\}d\log(\gamma_i)\\ &\quad= \Gamma(\tilde{a})\log(\gamma_i)f(\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}).\end{aligned}$$

Finally, dividing by $\Gamma(\tilde{a})$ and integrating over $\mathbb{R}$ with respect to $\log(\gamma_i)$ yields an expression for the

posterior expectation of $\log(\gamma_i)$ as follows

$$\begin{aligned}
\mathbb{E}[\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}] &= -\log(\tilde{b}) + \int_{\mathbb{R}} \log(\gamma_i) f(\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}) \\
&= -\log(\tilde{b}) + \frac{1}{\Gamma(\tilde{a})}\int_{\mathbb{R}} \frac{\partial}{\partial \tilde{a}} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i) \\
&= -\log(\tilde{b}) + \frac{1}{\Gamma(\tilde{a})}\frac{\partial}{\partial \tilde{a}}\int_{\mathbb{R}} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i) \\
&= -\log(\tilde{b}) + \frac{1}{\Gamma(\tilde{a})}\frac{\partial}{\partial \tilde{a}}\Gamma(\tilde{a}) = -\log(\tilde{b}) + \frac{\partial}{\partial \tilde{a}}\log\left[\Gamma(\tilde{a})\right] = \text{digamma}(\tilde{a}) - \log(\tilde{b})
\end{aligned}$$

B.1.2 E-Step

In the E-step, we calculate the expected log-conditional likelihood of the augmented data given the observed data, or our 'Q' function. Q can be written as

$$\begin{aligned}
Q(\boldsymbol{\psi} \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}) &= \mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}] \\
&= \mathbb{E}_{\boldsymbol{\gamma}}[\log(\textstyle\prod_{i=1}^n \gamma_i^{\delta_{i1}+\delta_{i2}} \times \frac{\theta^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \\
&\quad \times [\lambda_{01}(Y_{i1})e^{h_1(\boldsymbol{x}_i)}]^{\delta_{i1}} \times [\lambda_{02}(Y_{i2})e^{h_2(\boldsymbol{x}_i)}]^{(1-\delta_{i1})\delta_{i2}} \times [\lambda_{03}(Y_{i2}-Y_{i1})e^{h_3(\boldsymbol{x}_i)}]^{\delta_{i1}\delta_{i2}} \\
&\quad \times \exp\{-\gamma_i[\Lambda_{01}(Y_{i1})e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}(Y_{i1})e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}(Y_{i2}-Y_{i1})e^{h_3(\boldsymbol{x}_i)}]\}) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}] \\
&= \mathbb{E}_{\boldsymbol{\gamma}}[\textstyle\sum_{i=1}^n \delta_{i1}\log(\gamma_i) + \delta_{i2}\log(\gamma_i) + \delta_{i1}\log[\lambda_{01}(Y_{i1})] + \delta_{i1}h_1(\boldsymbol{x_i}) \\
&\quad + \ (1-\delta_{i1})\delta_{i2}\log[\lambda_{02}(Y_{i2})] + (1-\delta_{i1})\delta_{i2}h_2(\boldsymbol{x_i}) \\
&\quad + \ \delta_{i1}\delta_{i2}\log[\lambda_{03}(Y_{i2}-Y_{i1})] + \delta_{i1}\delta_{i2}h_3(\boldsymbol{x_i}) \\
&\quad - \ \gamma_i[\Lambda_{01}(Y_{i1})e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}(Y_{i1})e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}(Y_{i2}-Y_{i1})e^{h_3(\boldsymbol{x}_i)}] \\
&\quad - \frac{1}{\theta}\log(\theta) + (\frac{1}{\theta}-1)\log(\gamma_i) - \frac{1}{\theta}\gamma_i - \log\Gamma(\frac{1}{\theta}) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}] \\
&= \textstyle\sum_{i=1}^n \delta_{i1}\mathbb{E}_{\boldsymbol{\gamma}}[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}] + \delta_{i2}\mathbb{E}_{\boldsymbol{\gamma}}[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}] \\
&\quad + \ \delta_{i1}\log[\lambda_{01}(Y_{i1})] + \delta_{i1}h_1(\boldsymbol{x_i}) \\
&\quad + \ (1-\delta_{i1})\delta_{i2}\log[\lambda_{02}(Y_{i2})] + (1-\delta_{i1})\delta_{i2}h_2(\boldsymbol{x_i}) \\
&\quad + \ \delta_{i1}\delta_{i2}\log[\lambda_{03}(Y_{i2}-Y_{i1})] + \delta_{i1}\delta_{i2}h_3(\boldsymbol{x_i}) \\
&\quad - \ \mathbb{E}_{\boldsymbol{\gamma}}[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}^{(m)}][\Lambda_{01}(Y_{i1})e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}(Y_{i1})e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\Lambda_{03}(Y_{i2}-Y_{i1})e^{h_3(\boldsymbol{x}_i)}] \\
&\quad - \ \frac{1}{\theta}\log(\theta) + (\frac{1}{\theta}-1)\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}^{(m)}] - \frac{1}{\theta}\mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] - \log\Gamma(\frac{1}{\theta}) \\
&= Q_1 + Q_2 + Q_3 + Q_4,
\end{aligned}$$

where

$$
\begin{aligned}
Q_1 &= \sum_{i=1}^{n} \delta_{i1} \mathbb{E}_{\boldsymbol{\gamma}} \left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)} \right] + \delta_{i1} \log \left[\lambda_{01} \left(Y_{i1} \right) \right] + \delta_{i1} h_1(\boldsymbol{x_i}) \\
&\quad - \mathbb{E}_{\boldsymbol{\gamma}} \left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}^{(m)} \right] \Lambda_{01} \left(Y_{i1} \right) e^{h_1(\boldsymbol{x}_i)} \\
Q_2 &= \sum_{i=1}^{n} \delta_{i2} \mathbb{E}_{\boldsymbol{\gamma}} \left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}^{(m)} \right] + \left(1 - \delta_{i1} \right) \delta_{i2} \log \left[\lambda_{02} \left(Y_{i2} \right) \right] \\
&\quad + \left(1 - \delta_{i1} \right) \delta_{i2} h_2(\boldsymbol{x_i}) - \mathbb{E}_{\boldsymbol{\gamma}} \left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}^{(m)} \right] \Lambda_{02} \left(Y_{i1} \right) e^{h_2(\boldsymbol{x}_i)} \\
Q_3 &= \sum_{i=1}^{n} \delta_{i1} \delta_{i2} \log \left[\lambda_{03} \left(Y_{i2} - Y_{i1} \right) \right] + \delta_{i1} \delta_{i2} h_3(\boldsymbol{x_i}) \\
&\quad - \mathbb{E}_{\boldsymbol{\gamma}} \left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}^{(m)} \right] \delta_{i1} \Lambda_{03} \left(Y_{i2} - Y_{i1} \right) e^{h_3(\boldsymbol{x}_i)} \\
Q_4 &= \sum_{i=1}^{n} -\frac{1}{\theta} \log(\theta) + \left(\frac{1}{\theta} - 1 \right) \mathbb{E}[\log(\gamma_i) | \mathcal{D}, \boldsymbol{\psi}^{(m)}] - \frac{1}{\theta} \mathbb{E}[\gamma_i | \mathcal{D}, \boldsymbol{\psi}^{(m)}] - \log \Gamma \left(\frac{1}{\theta} \right).
\end{aligned}
$$

B.1.3 M-Step

In the M-step, the objective is to maximize the baseline hazard functions in the expected log-likelihood with the updated frailty estimates. Note that our objective function, Q, can be written as the sum of Q_1, Q_2, Q_3, and Q_4. Each of the first three involves only the baseline hazard for a state transition, and the last one involves only the frailty variance. Thus, the M-step updates for $\Lambda_{01}, \Lambda_{02}$, and Λ_{03} can be defined utilizing Q_1, Q_2, and Q_3, separately, and the frailty variance, θ, with Q_4. As the maximizer of our objective function over the space of absolutely continuous cumulative baseline hazards does not exist [76], we restrict the parameter space of the cumulative baseline hazards, $\Lambda_{01}, \Lambda_{02}$, and Λ_{03}, to the one containing piecewise constant functions, with jumps occurring at observed event times. Maximizers over this discrete space are termed nonparametric maximum likelihood estimates of $\Lambda_{01}, \Lambda_{02}$, and Λ_{03}. Under this parameter space, $\lambda_{0g}(t)$ in (2.4) - (2.6) are replaced by $\Delta\Lambda_{0g}(t)$, the jump size at t for the baseline hazards of each state transition [91], and $\Lambda_{0g}(t) = \sum_{s=0}^{t} \Delta\Lambda_{0g}(s)$. Note that $\Delta\Lambda_{0g}(s) = 0$ if s is not one of the observed event times corresponding to state transition g. As such, the M-step updates for these jump sizes, $\Delta\Lambda_{0g}(t)$, can be derived as follows.

Update for $\Delta\Lambda_{01}(t)$: The M-step involves maximize the values of the baseline hazard parameters given the expected log-likelihood and updated frailty estimates. Substituting the discretized jump sizes, $\Delta\Lambda_{01}$, for λ_{01}, we rewrite Q_1 as

$$Q_1 = \sum_{i=1}^{n} \delta_{i1} \mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}^{(m)}] + \delta_{i1} \log\left[\Delta\Lambda_{01}\left(Y_{i1}\right)\right] + \delta_{i1} h_1(\boldsymbol{x}_i)$$
$$- \ \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \Lambda_{01}\left(Y_{i1}\right) \exp\{h_1(\boldsymbol{x}_i)\}.$$

For a fixed t, differentiating Q_1 with respect to $\Delta\Lambda_{01}(t)$, we have the score function for $\Delta\Lambda_{01}(t)$,

$$\frac{\partial Q_1}{\partial \Delta\Lambda_{01}(t)} = \sum_{i=1}^{n} \frac{\delta_{i1} I(Y_{i1} = t)}{\Delta\Lambda_{01}(t)} - \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] I[Y_{i1} \geq t] \exp\{h_1(\boldsymbol{x}_i)\}.$$

Setting this equal to zero, we can show that the update, $\Delta\Lambda_{01}^{(m+1)}(t)$ is

$$\Delta\Lambda_{01}^{(m+1)}(t) = \frac{\sum_{i=1}^{n} \delta_{i1} I\left[Y_{i1} = t\right]}{\sum_{i=1}^{n} \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] I\left[Y_{i1} \geq t\right] \exp\left\{h_1^{(m)}(\boldsymbol{x}_i)\right\}},$$

where the numerator reflects the observed number of non-terminal events.

Update for $\boldsymbol{\Delta\Lambda_{02}(t)}$: As before, substituting $\Delta\Lambda_{02}$ for λ_{02}, we rewrite Q_2 as

$$Q_2 = \sum_{i=1}^{n} \delta_{i2} \mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}^{(m)}] + (1 - \delta_{i1})\, \delta_{i2} \log\left[\Delta\Lambda_{02}\left(Y_{i2}\right)\right]$$
$$+ \ (1 - \delta_{i1})\, \delta_{i2} h_2(\boldsymbol{x}_i) - \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \Lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{x}_i)\}.$$

Differentiating Q_2 with respect to $\Delta\Lambda_{02}$, we have the score function

$$\frac{\partial Q_2}{\partial \Delta\Lambda_{02}(t)} = \sum_{i=1}^{n} \frac{(1 - \delta_{i1})\, \delta_{i2} I[Y_{i2} = t]}{\Delta\Lambda_{02}(t)} - \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] I[Y_{i2} \geq t] \exp\{h_2(\boldsymbol{x}_i)\}.$$

Setting this equal to zero, we can show that the update, $\Delta\Lambda_{02}^{(m+1)}(t)$, is

$$\Delta\Lambda_{02}^{(m+1)}(t) = \frac{\sum_{i=1}^{n} (1 - \delta_{i1})\, \delta_{i2} I\left[Y_{i2} = t\right]}{\sum_{i=1}^{n} \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] I[Y_{i2} \geq t] \exp\left\{h_2^{(m)}(\boldsymbol{x}_i)\right\}},$$

where the numerator is the number of terminal events observed prior to non-terminal events.

Update for $\boldsymbol{\Delta\Lambda_{03}(t)}$: Lastly, substituting $\Delta\Lambda_{03}$ for λ_{03}, we rewrite Q_3 as

$$Q_3 = \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \log\left[\Delta\Lambda_{03}\left(Y_{i2} - Y_{i1}\right)\right] + \delta_{i1}\delta_{i2} h_3(\boldsymbol{x}_i)$$
$$- \ \mathbb{E}[\gamma_i|\mathcal{D}, \boldsymbol{\psi}^{(m)}] \delta_{i1} \Lambda_{03}\left(Y_{i2} - Y_{i1}\right) \exp\{h_3(\boldsymbol{x}_i)\}.$$

Differentiating Q_3 with respect to $\Delta\Lambda_{03}$, we have the score function

$$\frac{\partial Q_3}{\partial \Delta\Lambda_{03}(t)} = \sum_{i=1}^{n} \frac{\delta_{i1}\delta_{i2} I\left[Y_{i2} - Y_{i1} = t\right]}{\Delta\Lambda_{03}\left(t\right)} - \mathbb{E}[\gamma_i | \mathcal{D}, \boldsymbol{\psi}^{(m)}] \delta_{i1} I\left[Y_{i2} - Y_{i1} \geq t\right] \exp\{h_3(\boldsymbol{x}_i)\},$$

and equating this to zero, we have that the update, $\Delta\Lambda_{03}^{(m+1)}(t)$, is

$$\Delta\Lambda_{03}^{(m+1)}(t) = \frac{\sum_{i=1}^{n} \delta_{i1}\delta_{i2} I\left[Y_{i2} - Y_{i1} = t\right]}{\sum_{i=1}^{n} \mathbb{E}[\gamma_i | \mathcal{D}, \boldsymbol{\psi}^{(m)}] \delta_{i1} I\left[Y_{i2} - Y_{i1} \geq t\right] \exp\left\{h_3^{(m)}(\boldsymbol{x}_i)\right\}},$$

where the numerator reflects the number of terminal events observed after non-terminal events. These closed form updates in the M-step are Breslow-type estimators. As such, to seed the EM algorithm, we initialize Λ_{01}, Λ_{02}, and Λ_{03} with their respective, unadjusted Nelson-Aalen estimators.

APPENDIX C

Technical Details for Chapter 4

C.1 Supplemental Simulation Information

C.1.1 Data Generation Procedure

In the following, we detail the the data generation procedure for our simulation studies.

Proportional Hazards Model, Linear Risk Function

Given the formulation of the Clayton copula, we can express the bivariate surival function of the non-fatal, T_{i1}, and fatal, T_{i2}, event times as

$$S(t_1, t_2) = \Pr(T_{i1} > t_1, T_{i2} > t_2) = \left[S_1(t_1)^{1-\theta} + S_2(t_2)^{1-\theta} - 1\right]^{\frac{1}{1-\theta}}; 0 \leq t_1 \leq t_2,$$

where $S_1(t_1)$ is the marginal survival function of the non-fatal event, $S_2(t_2)$ is the marginal survival function of the fatal event, and θ is the copula parameter which measures the dependence between the non-fatal and fatal event times. In the first simulation setting, we generated non-fatal (T_{i1}) and fatal (T_{i2}) event times from marginal models specified by

$$\log(T_{i1}/3) = -(\beta_1 Z_i + \beta_1 X_{i1} + \beta_1 X_{i2}) + \varepsilon_{i1}$$
$$\log(T_{i2}/3) = -(\beta_2 Z_i + \beta_2 X_{i1} + \beta_2 X_{i2}) + \varepsilon_{i2},$$

where Z_i is a Bernoulli random variable with a success probability of 0.5, X_{i1} and X_{i12} are independent truncated normal random variables with mean 1, variance 0.5, and truncation bounds of [0, 2], and $(\varepsilon_{i1}, \varepsilon_{i2})$ are correlated random errors. To induce dependence between the simulated event times, we simulate ε_{i1} and ε_{i2} from the Clayton copula model,

$$\left[\Pr\left(\varepsilon_{i1} > t_1\right)^{-\theta} + \Pr\left(\varepsilon_{i2} > t_2\right)^{-\theta} - 1\right]^{-\frac{1}{\theta}},$$

where ε_{i1} and ε_{i2} follow the extreme value distribution, i.e., $\Pr\left(\varepsilon_{i1} > t_1\right) = \exp\{-\exp(t_1)\}$ and $\Pr\left(\varepsilon_{i2} > t_2\right) = \exp\{-\exp(t_2)\}$ [116]. The data generation procedure is as follows:

1. Draw two independent uniform random variables, $U_{i1}, V_{i2} \sim \text{Unif}(0, 1)$
2. Set $\varepsilon_{i1} = \log\{-\log(U_{i1})\}$
3. Set $U_{i2} = \left[\left(V_{i2}^{-\theta/(1+\theta)} - 1\right) \times \exp\{\theta \exp(\varepsilon_{i1})\} + 1\right]^{-1/\theta}$
4. Set $\varepsilon_{i2} = \log\{-\log(U_{i2})\}$
5. Draw a Bernoulli random variable, Z_i, with success probability 0.5
6. Draw X_{i1}, X_{i2} from independent N(1, 0.5) distributions with truncation bounds [0, 2]
7. Set $T_{i1} = 3 \times \exp\{-(\beta_1 Z_i + \beta_1 X_{i1} + \beta_1 X_{i2}) + \varepsilon_{i1}\}$ with $\beta_1 = 1$
8. Set $T_{i2} = 3 \times \exp\{-(\beta_2 Z_i + \beta_2 X_{i1} + \beta_2 X_{i2}) + \varepsilon_{i2}\}$ with $\beta_2 = 0.2$
9. Draw C_i, from a mixture of uniforms, where $C_i \sim \xi_i \text{Unif}(0, 1) + (1 - \xi_i)\text{Unif}(1, 1.2)$ with $\xi_i \sim \text{Bern}(0.2)$
10. Set $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = \mathbb{I}(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = \mathbb{I}(T_{i1} \leq Y_{i2})$
11. Repeat steps (1) - (10) for $i = 1, \ldots, n$
12. Return $\{(Y_{i1}, \delta_{i1}, Y_{i2}, \delta_{i2}, Z_i, X_{i1}, X_{i2}); i = 1, \ldots, n\}$

Proportional Hazards Model, Non-Linear Risk Function

In this setting, we repeat the same data generation procedure as listed above, except

- In step (6), we draw X_{i1}, X_{i2} from independent N(0, 0.5) distributions with truncation bounds [-1, 1]
- In step (7), we set $T_{i1} = 3 \times \exp\{-(\beta_1 Z_i + \beta_1 X_{i1}^2 + \beta_1 X_{i2}^2) + \varepsilon_{i1}\}$
- In step (8), we set $T_{i2} = 3 \times \exp\{-(\beta_2 Z_i + \beta_2 X_{i1}^2 + \beta_2 X_{i2}^2) + \varepsilon_{i1}\}$.

Accelerated Failure Time Model, Proportional Hazards Violated

In the third setting, we generated data from a complex model which violates the proportional hazard assumption and includes non-linear effects and interactions as follows:

1. Draw two independent uniform random variables, $\varepsilon_{i1}, V_{i2} \sim \text{Unif}(0, 1)$
2. Set $\varepsilon_{i2} = \left[\varepsilon_{i1}^{-\theta}\left(V_{i2}^{-\theta/(1+\theta)} - 1\right) + 1\right]^{-1/\theta}$

3. Draw $\boldsymbol{X} = (X_1, \ldots, X_{12})' \sim N_{12}(\mathbf{0}, \Sigma)$, where

   $\Sigma_{12\times 12} = \begin{bmatrix} D_{6\times 6} & 0_{6\times 6} \\ 0_{6\times 6} & D_{6\times 6} \end{bmatrix}$ with elements $(d_{ij}) = 0.5^{|i-j|}$

4. Set $Z_i = \mathbb{I}(X_{i1} \geq 0)$

5. Set $a_1 = 0.5 + 2Z_i$

6. Set $b_1 = 10 + |5Z_i + (X_{i2} - 0.5)^2 + 2Z_i X_{i2} + X_{i3} + X_{i4} + X_{i5} + X_{i6} + X_{i7} + X_{i8} + X_{i9}|$

7. Set $a_2 = 2a_1$ and $b_2 = 2b_1$

8. Given $S(t|Z, X) = \exp\{-(t/b)^a\}$ and $F(t|Z, X) = 1 - S(t|X) = 1 - \exp\{-(t/b)^a\}$

   (a) Set $T_{i1} = F^{-1}(\varepsilon_{i1}|Z, X) = b_1 \times [-\log(1 - \varepsilon_{i1})]^{1/a_1}$

   (b) Set $T_{i2} = F^{-1}(\varepsilon_{i2}|Z, X) = b_2 \times [-\log(1 - \varepsilon_{i2})]^{1/a_2}$

9. Draw $C_i \sim \text{Exp}(\lambda)$ with $\lambda = 1$

10. Set $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = \mathbb{I}(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = \mathbb{I}(T_{i1} \leq Y_{i2})$

11. Repeat steps (1) - (10) for $i = 1, \ldots, n$

12. Return $\{(Y_{i1}, \delta_{i1}, Y_{i2}, \delta_{i2}, Z_i, X_{i1}, X_{i2}); i = 1, \ldots, n\}$

BIBLIOGRAPHY

[1] Pengyu Huang Aastha and Yan Liu. Deepcompete: A deep learning approach to competing risks in continuous time domain. In *AMIA Annual Symposium Proceedings*, volume 2020, page 177. American Medical Informatics Association, 2020.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[3] Anees Abrol, Zening Fu, Mustafa Salman, Rogers Silva, Yuhui Du, Sergey Plis, and Vince Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):1–17, 2021.

[4] Akshay Agrawal, Akshay Modi, Alexandre Passos, Allen Lavoie, Ashish Agarwal, Asim Shankar, Igor Ganichev, Josh Levenberg, Mingsheng Hong, Rajat Monga, et al. Tensorflow eager: A multi-stage, python-embedded dsl for machine learning. *Proceedings of Machine Learning and Systems*, 1:178–189, 2019.

[5] Kwang Woo Ahn, Anjishnu Banerjee, Natasha Sahr, and Soyoung Kim. Group and within-group variable selection for competing risks data. *Lifetime Data Analysis*, 24(3):407–424, 2018.

[6] Kwang Woo Ahn and Franco Mendolia. Pseudo-value approach for comparing survival medians for dependent data. *Statistics in medicine*, 33(9):1531–1538, 2014.

[7] JJ Allaire and François Chollet. *keras: R Interface to 'Keras'*, 2022. R package version 2.9.0.

[8] JJ Allaire and Yuan Tang. *tensorflow: R Interface to 'TensorFlow'*, 2022. R package version 2.9.0.

[9] Eitan Amir, Bostjan Seruga, Ryan Kwong, Ian F Tannock, and Alberto Ocaña. Poor correlation between progression-free and overall survival in modern clinical trials: are composite endpoints the answer? *European Journal of Cancer*, 48(3):385–388, 2012.

[10] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[11] Per K Andersen, Elisavet Syriopoulou, and Erik T Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, 36(17):2669–2681, 2017.

[12] Allison B Ashworth, Suresh Senan, David A Palma, Marc Riquet, Yong Chan Ahn, Umberto Ricardi, Maria T Congedo, Daniel R Gomez, Gavin M Wright, Giulio Melloni, et al. An individual patient data metaanalysis of outcomes and prognostic factors after treatment of oligometastatic non–small-cell lung cancer. *Clinical lung cancer*, 15(5):346–355, 2014.

[13] Brett C Bade and Charles S Dela Cruz. Lung cancer 2020: epidemiology, etiology, and prevention. *Clinics in chest medicine*, 41(1):1–24, 2020.

[14] Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

[15] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, 2008.

[16] Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

[17] Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

[18] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

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

[20] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[21] Stephen F Brown, Alan J Branford, and William Moran. On the use of artificial neural networks for the analysis of survival data. *IEEE transactions on neural networks*, 8(5):1071–1077, 1997.

[22] Michael D Brundage, Diane Davies, and William J Mackillop. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest*, 122(3):1037–1057, 2002.

[23] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[24] Robert James Cerfolio, Ayesha S Bryant, Ethan Scott, Manisha Sharma, Francisco Robert, Sharon A Spencer, and Robert I Garver. Women with pathologic stage i, ii, and iii non-small cell lung cancer have better survival than men. *Chest*, 130(6):1796–1802, 2006.

[25] Aloka Chakravarty and Rajeshwari Sridhara. Use of progression-free survival as a surrogate marker in oncology trials: some regulatory issues. *Statistical Methods in Medical Research*, 17(5):515–518, 2008.

[26] Pei-Yun Chen and Anastasios A Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

[27] David C Christiani. The boston lung cancer survival cohort, 2017.

[28] David C. Christiani. The Boston lung cancer survival cohort. http://grantome.com/grant/NIH/U01-CA209414-01A1, 2017. [Online; accessed November 12, 2022].

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

[30] Antonio Ciampi, Johanne Thiffault, Jean-Pierre Nakache, and Bernard Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, 4(3):185–204, 1986.

[31] David G Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[32] Leah Comment, Fabrizia Mealli, Sebastien Haneuse, and Corwin Zigler. Survivor average causal effects for continuous time: a principal stratification approach to causal inference with semicompeting risks. *arXiv preprint arXiv:1902.09304*, 2019.

[33] Alessio Cortellini, Andrea De Giglio, Katia Cannita, Diego L Cortinovis, Robin Cornelissen, Cinzia Baldessari, Raffaele Giusti, Ettore D'Argento, Francesco Grossi, Matteo Santoni, et al. Smoking status during first-line immunotherapy and chemotherapy in nsclc patients: A case–control matched analysis from a large multicenter study. *Thoracic Cancer*, 12(6):880–889, 2021.

[34] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[35] Colin J Crilly, Sebastien Haneuse, and Jonathan S Litt. Predicting the outcomes of preterm neonates beyond the neonatal intensive care unit: What are we missing? *Pediatric Research*, 89(3):426–445, 2021.

[36] Lei Cui, Hansheng Li, Wenli Hui, Sitong Chen, Lin Yang, Yuxin Kang, Qirong Bo, and Jun Feng. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC bioinformatics*, 21:1–14, 2020.

[37] Judith R Davidson, Michael D Brundage, and Deb Feldman-Stewart. Lung cancer treatment decisions: patients' desires for participation and information. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, 8(6):511–520, 1999.

[38] Amanda Delgado and Achuta Kumar Guddati. Clinical endpoints in oncology-a primer. *American journal of cancer research*, 11(4):1121, 2021.

[39] Romano Demicheli, Marco Fornili, Federico Ambrogi, Kristin Higgins, Jessamy A Boyd, Elia Biganzoli, and Chris R Kelsey. Recurrence dynamics for non–small-cell lung cancer: effect of surgery on the development of metastases. *Journal of Thoracic Oncology*, 7(4):723–730, 2012.

[40] Shreyesh Doppalapudi, Robin G Qiu, and Youakim Badr. Lung cancer survival period prediction and understanding: Deep learning approaches. *International Journal of Medical Informatics*, 148:104371, 2021.

[41] Antonio Eleuteri, Roberto Tagliaferri, Leopoldo Milano, Sabino De Placido, and Michele De Laurentiis. A novel neural network-based survival analysis model. *Neural Networks*, 16(5-6):855–864, 2003.

[42] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995.

[43] Mahtab Jahanbani Fard, Ping Wang, Sanjay Chawla, and Chandan K Reddy. A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3126–3139, 2016.

[44] David Fedor, W Rainey Johnson, and Sunil Singhal. Local recurrence following lung cancer surgery: incidence, risk factors, and outcomes. *Surgical oncology*, 22(3):156–161, 2013.

[45] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[46] Jason P Fine, Hongyu Jiang, and Rick Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, 2001.

[47] Evelyn Fix and Jerzy Neyman. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, 23(3):205–241, 1951.

[48] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[49] Laurie E Gaspar, Erica J McNamara, E Greer Gay, Joe B Putnam, Jeffrey Crawford, Roy S Herbst, and James A Bonner. Small-cell lung cancer: prognostic factors and changing treatment over 15 years. *Clinical lung cancer*, 13(2):115–122, 2012.

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

[51] Debashis Ghosh. A causal framework for surrogate endpoints with semi-competing risks data. *Statistics & probability letters*, 82(11):1898–1902, 2012.

[52] Emilio AL Gianicolo, Martin Eichler, Oliver Muensterer, Konstantin Strauch, and Maria Blettner. Methods for evaluating causality in observational studies: Part 27 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 117(7):101, 2020.

[53] Ashish Goel, Alpana Raizada, Ananya Agrawal, Kamakshi Bansal, Saurabh Uniyal, Pratima Prasad, Anil Yadav, Asha Tyagi, and RS Rautela. Correlates of in-hospital covid-19 deaths: a competing risks survival time analysis of retrospective mortality data. *Disaster Medicine and Public Health Preparedness*, pages 1–27, 2021.

[54] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[55] Louis Gordon and Richard A Olshen. Tree-structured survival analysis. *Cancer Treatment Reports*, 69(10):1065–1069, 1985.

[56] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

[57] Nancy L Guo, Kursad Tosun, and Kimberly Horn. Impact and interactions between smoking and traditional prognostic factors in lung cancer progression. *Lung cancer*, 66(3):386–392, 2009.

[58] Il Do Ha, Minjung Lee, Seungyoung Oh, Jong-Hyeon Jeong, Richard Sylvester, and Youngjo Lee. Variable selection in subdistribution hazard frailty models with competing risks data. *Statistics in Medicine*, 33(26):4590–4604, 2014.

[59] Sebastien Haneuse and Kyu Ha Lee. Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circulation: Cardiovascular Quality and Outcomes*, 9(3):322–331, 2016.

[60] Lin Hao, Juncheol Kim, Sookhee Kwon, and Il Do Ha. Deep learning-based survival analysis for high-dimensional survival data. *Mathematics*, 9(11):1244, 2021.

[61] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[62] Miguel A Hernán and James M Robins. Causal inference, 2010.

[63] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[64] Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in Medicine*, 23(1):77–91, 2004.

[65] Jue Hou, Jelena Bradic, and Ronghui Xu. Inference under fine-gray competing risks model with high-dimensional covariates. *Electronic Journal of Statistics*, 13(2):4449–4507, 2019.

[66] Jin-Jian Hsieh and Yu-Ting Huang. Regression analysis based on conditional likelihood approach under semi-competing risks data. *Lifetime data analysis*, 18(3), 2012.

[67] Chen Hu and Jon Arni Steingrimsson. Personalized risk prediction in clinical oncology research: applications and practical issues using survival trees and random forests. *Journal of Biopharmaceutical Statistics*, 28(2):333–349, 2018.

[68] Liangyuan Hu, Jung-Yi Lin, Keith Sigel, and Minal Kale. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. *Annals of Epidemiology*, 62:36–42, 2021.

[69] Yen-Tsung Huang. Causal mediation of semicompeting risks. *Biometrics*, 77(4):1143–1154, 2021.

[70] Kentaro Inamura and Yuichi Ishikawa. Lung cancer progression and metastasis from the prognostic point of view. *Clinical & experimental metastasis*, 27(6):389–397, 2010.

[71] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[72] Hemant Ishwaran, Udaya B Kogalur, Xi Chen, and Andy J Minn. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1):115–132, 2011.

[73] Hemant Ishwaran and Min Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, 38(4):558–582, 2019.

[74] Ina Jazić, Deborah Schrag, Daniel J Sargent, and Sebastien Haneuse. Beyond composite endpoints analysis: semicompeting risks as an underutilized framework for cancer research. *JNCI: Journal of the National Cancer Institute*, 108(12), 2016.

[75] How Jing and Alexander J Smola. Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 515–524, 2017.

[76] Søren Johansen. An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, pages 165–174, 1983.

[77] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[78] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[79] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *stat*, 1050(2):1–10, 2016.

[80] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.

[81] Eric S Kawaguchi, Jenny I Shen, Gang Li, and Marc A Suchard. A fast and scalable implementation method for competing risks data with the r package fastcmprsk. *arXiv preprint arXiv:1905.07438*, 2019.

[82] SM Keller, MG Vangel, S Adak, H Wagner, JH Schiller, A Herskovic, R Komaki, MC Perry, RS Marks, RB Livingston, et al. The influence of gender on survival and tumor recurrence following adjuvant therapy of completely resected stages ii and iiia non-small cell lung cancer. *Lung cancer*, 37(3):303–309, 2002.

[83] Bernard Koch, Tim Sainburg, Pablo Geraldo, Song Jiang, Yizhou Sun, and Jacob Gates Foster. Deep learning of potential outcomes. *arXiv preprint arXiv:2110.04442*, 2021.

[84] Michael T Koller, Heike Raatz, Ewout W Steyerberg, and Marcel Wolbers. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in Medicine*, 31(11-12):1089–1097, 2012.

[85] Bryan Lau, Stephen R Cole, and Stephen J Gange. Competing risk regression models for epidemiologic data. *American Journal of Epidemiology*, 170(2):244–256, 2009.

[86] Jennifer G Le-Rademacher, Ryan A Peterson, Terry M Therneau, Ben L Sanford, Richard M Stone, and Sumithra J Mandrekar. Application of multi-state models in cancer clinical trials. *Clinical Trials*, 15(5):489–498, 2018.

[87] Michael LeBlanc and John Crowley. Relative risk trees for censored survival data. *Biometrics*, pages 411–425, 1992.

[88] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.

[89] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[90] Jing Li, Ying Zhang, Giorgos Bakoyannis, and Sujuan Gao. On shared gamma-frailty conditional markov model for semicompeting risks data. *Statistics in Medicine*, 39(23):3042–3058, 2020.

[91] Yi Li and Xihong Lin. Covariate measurement errors in frailty models for clustered survival data. *Biometrika*, 87(4):849–866, 2000.

[92] Wenhua Liang, Jun Liu, and Jianxing He. Driving the improvement of lung cancer prognosis. *Cancer Cell*, 38(4):449–451, 2020.

[93] Knut Liestbl, Per Kragh Andersen, and Ulrich Andersen. Survival analysis and neural nets. *Statistics in medicine*, 13(12):1189–1200, 1994.

[94] Paulo JG Lisboa, H Wong, P Harris, and Ric Swindell. A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine*, 28(1):1–25, 2003.

[95] Brent R Logan, Mei-Jie Zhang, and John P Klein. Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics*, 67(1):1–7, 2011.

[96] Dustin M Long and Michael G Hudgens. Sharpening bounds on principal effects with covariates. *Biometrics*, 69(4):812–819, 2013.

[97] Thomas J Lynch, Daphne W Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A Okimoto, Brian W Brannigan, Patricia L Harris, Sara M Haserlat, Jeffrey G Supko, and Frank G Haluska. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004.

[98] Tetsuya Mitsudomi, Kenichi Suda, and Yasushi Yatabe. Surgery for nsclc in the era of personalized medicine. *Nature reviews Clinical oncology*, 10(4):235–244, 2013.

[99] Daniel Nevo and Malka Gorfine. Causal inference for semi-competing risks data. *Biostatistics*, 23(4):1115–1132, 2022.

[100] William S Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

[101] David Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493, 1989.

[102] Annalisa Orenti, Patrizia Boracchi, Giuseppe Marano, Elia Biganzoli, and Federico Ambrogi. A pseudo-values regression model for non-fatal event free survival in the presence of semi-competing risks. *Statistical Methods & Applications*, pages 1–19, 2021.

[103] J Guillermo Paez, Pasi A Janne, Jeffrey C Lee, Sean Tracy, Heidi Greulich, Stacey Gabriel, Paula Herman, Frederic J Kaye, Neal Lindeman, Titus J Boggon, et al. Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500, 2004.

[104] Judea Pearl. Causal inference in statistics: An overview. *Statistical Surveys*, 3:96–146, 2009.

[105] Judea Pearl. Principal stratification–a goal or a tool? *The international journal of biostatistics*, 7(1):1–13, 2011.

[106] Limin Peng and Jason P Fine. Regression modeling of semicompeting risks data. *Biometrics*, 63(1):96–108, 2007.

[107] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[108] Katerina Politi and Roy S Herbst. Lung cancer in the era of precision medicine. *Clinical cancer research*, 21(10):2213–2220, 2015.

[109] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. Fast training of support vector machines for survival analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 243–259. Springer, 2015.

[110] Sanjay Popat, Stephen V. Liu, Nicolas Scheuer, Alind Gupta, Grace G. Hsu, Sreeram V. Ramagopalan, Frank Griesinger, and Vivek Subbiah. Association Between Smoking History and Overall Survival in Patients Receiving Pembrolizumab for First-Line Treatment of Advanced Non–Small Cell Lung Cancer. *JAMA Network Open*, 5(5):e2214046–e2214046, 05 2022.

[111] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114. PMLR, 2016.

[112] Denise Rava. *Survival Analysis and Causal Inference: from Marginal Structural Cox to Additive Hazards Model and beyond*. University of California, San Diego, 2021.

[113] Harrison T Reeder, Junwei Lu, and Sebastien Haneuse. Penalized estimation of frailty-based illness-death models for semi-competing risks. *arXiv preprint arXiv:2202.00618*, 2022.

[114] Lynn AG Ries, D Harkins, M Krapcho, Angela Mariotto, BA Miller, Eric J Feuer, Limin X Clegg, MP Eisner, Marie-Josèphe Horner, Nadia Howlader, et al. Seer cancer statistics review, 1975-2003. *National Cancer Institute*, 2006.

[115] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[116] Federico Rotolo, Catherine Legrand, and Ingrid Van Keilegom. A simulation procedure based on copulas to generate clustered multi-state survival data. *Computer methods and programs in biomedicine*, 109(3):305–312, 2013.

[117] Jacqueline E Rudolph, Catherine R Lesko, and Ashley I Naimi. Causal inference in the face of competing events. *Current Epidemiology Reports*, 7(3):125–131, 2020.

[118] Camille Sabathé, Per K Andersen, Catherine Helmer, Thomas A Gerds, Hélène Jacqmin-Gadda, and Pierre Joly. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical methods in medical research*, 29(3):752–764, 2020.

[119] Stephen Salerno and Yi Li. Deep learning of semi-competing risk data via a new neural expectation-maximization algorithm, 2022.

[120] Stephen Salerno and Yi Li. High-dimensional survival analysis: Methods and applications. *arXiv*, Preprint posted online May 5, 2022. arXiv:2205.02948 [stat.ME]. doi: 10.48550/arXiv.2205.02948.

[121] Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 655–660. IEEE, 2007.

[122] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1):17–48, 2023.

[123] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[124] Madan Somvanshi, Pranjali Chavan, Shital Tambade, and S. V. Shinde. A review of machine learning techniques using decision tree and support vector machine. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–7, 2016.

[125] Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: an r package for machine learning in survival analysis. *Bioinformatics*, 37(17):2789–2791, 2021.

[126] Jon Arni Steingrimsson, Liqun Diao, Annette M Molinaro, and Robert L Strawderman. Doubly robust survival trees. *Statistics in Medicine*, 35(20):3595–3612, 2016.

[127] Jon Arni Steingrimsson, Liqun Diao, and Robert L Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525):370–383, 2019.

[128] Jon Arni Steingrimsson and Samantha Morrison. Deep learning for survival outcomes. *Statistics in medicine*, 39(17):2339–2349, 2020.

[129] Matthew A Steliga and Carolyn M Dresler. Epidemiology of lung cancer: smoking, secondhand smoke, and genetics. *Surgical Oncology Clinics*, 20(4):605–618, 2011.

[130] Ori M Stitelman, Victor De Gruttola, and Mark J van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics*, 8(1), 2012.

[131] Erling Sverdrup. Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Scandinavian Actuarial Journal*, 1965(3-4):184–211, 1965.

[132] Eric J Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Statistics in medicine*, 33(21):3601–3628, 2014.

[133] Terry M Therneau, Patricia M Grambsch, and Thomas R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[134] Donna Tjandra, Yifei He, and Jenna Wiens. A hierarchical approach to multi-event survival analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):591–599, 2021.

[135] Jeng-Sen Tseng, Chun-Ju Chiang, Kun-Chieh Chen, Zhe-Rong Zheng, Tsung-Ying Yang, Wen-Chung Lee, Kuo-Hsuan Hsu, Yen-Hsiang Huang, Tsang-Wu Liu, Jiun-Yi Hsia, et al. Association of smoking with patient characteristics and outcomes in small cell lung carcinoma, 2011-2018. *JAMA network open*, 5(3):e224830–e224830, 2022.

[136] Hidetaka Uramoto and Fumihiro Tanaka. Prediction of recurrence after complete resection in patients with nsclc. *Anticancer research*, 32(9):3953–3960, 2012.

[137] Vanya Van Belle, Kristiaan Pelckmans, JAK Suykens, and Sabine Van Huffel. Support vector machines for survival analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, pages 1–8, 2007.

[138] Vanya Van Belle, Kristiaan Pelckmans, Sabine Van Huffel, and Johan AK Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

[139] Vladimir Vapnik, Isabel Guyon, and Trevor Hastie. Support vector machines. *Mach. Learn*, 20(3):273–297, 1995.

[140] Ashley J Vargas and Curtis C Harris. Biomarker development in the precision medicine era: lung cancer as a case study. *Nature Reviews Cancer*, 16(8):525–537, 2016.

[141] Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

[142] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[143] Gavin Wright, Renee L Manser, Graham Byrnes, David Hart, and Donald A Campbell. Surgery for non-small cell lung cancer: systematic review and meta-analysis of randomised controlled trials. *Thorax*, 61(7):597–603, 2006.

[144] Yujiao Wu, Jie Ma, Xiaoshui Huang, Sai Ho Ling, and Steven Weidong Su. Deepmmsa: A novel multimodal deep learning method for non-small cell lung cancer survival analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1468–1472. IEEE, 2021.

[145] Jinfeng Xu, John D Kalbfleisch, and Beechoo Tai. Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725, 2010.

[146] Yanxun Xu, Daniel Scharfstein, Peter Müller, and Michael Daniels. A bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics*, 23(1):34–49, 2022.

[147] Guolei Yang, Ying Cai, and Chandan K Reddy. Spatio-temporal check-in time prediction with recurrent neural network based survival analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[148] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 406–414. Springer, 2017.

[149] Jessica G Young, Mats J Stensrud, Eric J Tchetgen Tchetgen, and Miguel A Hernán. A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39(8):1199–1236, 2020.

[150] Qianyu Yuan, Tianrun Cai, Chuan Hong, Mulong Du, Bruce E Johnson, Michael Lanuti, Tianxi Cai, and David C Christiani. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. *JAMA Network Open*, 4(7):e2114723–e2114723, 2021.

[151] Cecilia Zappa and Shaker A Mousa. Non-small cell lung cancer: current treatment and future advances. *Translational lung cancer research*, 5(3):288, 2016.

[152] Tamir Zehavi and Daniel Nevo. Matching methods for truncation by death problems. *arXiv preprint arXiv:2110.10186*, 2021.

[153] Junni L Zhang and Donald B Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.

[154] Lili Zhao and Dai Feng. Deep neural networks for survival analysis using pseudo values. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3308–3314, 2020.

[155] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021.

[156] Blaž Zupan, Janez Demšar, Michael W Kattan, J Robert Beck, and Ivan Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75, 2000.