ABSTRACT

Title of Dissertation: CAUSAL SURVIVAL ANALYSIS - MACHINE LEARNING ASSISTED MODELS: STRUCTURAL NESTED ACCELERATED FAILURE TIME MODEL AND THRESHOLD REGRESSION

Yiming Chen
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Mei-Ling Ting Lee
Department of Epidemiology and Biostatistics

Time-varying confounding for intervention complicates causal survival analysis when the data are collected in a longitudinal manner. Traditional survival models that only adjust for time-dependent covariates provide a biased causal conclusion for the intervention effect. Some techniques have been developed to address this challenge. Nevertheless, these existing methods may still lack power, and suffer from computational burden given high dimensional data with a temporally connected nature. The first part of this dissertation focuses on one of the methods that deal with time-varying confounding, the Structural Nested Model and associated G-estimation. Two Neural Networks (GE-SCORE and GE-MIMIC) were proposed to estimate the Structural Nested Accelerated Failure Time Model. The proposed algorithms can provide less biased and individualized intervention causal effect estimation. The second part explored the causal interpretations and applications of the First-Hitting-Time based Threshold Regression Model using a

Wiener process. Moreover, a Neural Network expansion of this specific type of Threshold Regression (TRNN) was explored for the first time.

CAUSAL SURVIVAL ANALYSIS - MACHINE LEARNING ASSISTED MODELS: STRUCTURAL NESTED ACCELERATED FAILURE TIME MODEL AND THRESHOLD REGRESSION

by

Yiming Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Mei-Ling Ting Lee, Chair
Associate Professor Paul Smith, Co-Chair
Assistant Professor Tianzhou Ma
Associate Professor Takumi Saegusa
Professor Guodong Gao, Dean's representative

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible. Especially my husband Wenhao who has supported me emotionally, financially, and unconditionally. He is not only my best friend but also the mentor of my life. Nevertheless, later he clarified his love is still conditioned on I am a "reasonable person". He has shared his Ph.D. experience with me exhaustively. He also motivated me a lot by doing J.D. at George Washington University while working full-time. We inspired each other to become better people, this is the best part I can expect from a relationship. Many of my friends also shared their sorrows and happiness with me. Though we didn't share the research, I am indeed much supported by the hand-holding. I wish the best for them.

Furthermore, I'd like to thank my adviser, Dr. Lee who has supervised the process. She also provided me with precious career opportunities and instructed me on how to do independent thinking. My co-chair Dr. Smith spent numerous time correcting my bad wording. My English has improved significantly. Dr. Ma has provided many good suggestions for the simulations and results presentations. Though he is not my primary advisor and he was super occupied by many other matters, he made time to thought about my problem and discuss it with me. He is a professor who will think of and inspire students. Dr. Saegusa rescued my defense as a last-minute committee member agreed to serve. Though I never met my Dean's Representative Dr. Gao before the defense day,

I can tell from the few hours of interaction he is so professional and intelligent. His encouragements mean a lot to me. Without their advice and dedication, this work cannot be done.

The journey has been bumpy, but, I am glad I am finally here. I have learned much more than "how to do research" from this degree. It's more like a process to strengthen my mind, deepen my thoughts, and smooth my temper. The accomplishment is well deserved. The feeling of completion is truly rewarded. I have expanded my maximal bandwidth to complete this dissertation as well as this degree, I thank myself for not giving up. Before I started this degree, I was worried about the span can be long. I was not a good student of math starting from a young age so the challenge is obvious. Maybe I am more willing to think compared to others. Maybe I made good choices at some important time points. This is a causal problem and I would never know "what if". But what I know is, if I didn't do the degree, time still flows at the same speed. Now when I look back, I know I didn't waste these years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AFT	Accelerated Failure Time
ANN	Artificial Neural Network
BMI	Body Mass Index
CARDIA	Coronary Artery Risk Development in Young Adults
CHD	Coronary Heart Disease
CI	Confidence Interval
CIT	Conditional Independence Test
CGAN	Conditional Generative Adversarial Net
CHD	Coronary Heart Disease
CRN	Counterfactual Recurrent Network
CRT	Conditional Randomization Test
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
EHR	Electronic Health Record
FHT	First-Hitting-Time
GRU	Gated Recurrent Unit
GAN	Generative Adversarial Net
GCIT	Generative Conditional Independence Test
GLM	Generalized Linear Model
HR	Hazard Ratio
HRT	Holdout Randomization Test
ICA	Ignorable Censoring Assumption
IPCW	Inverse Probability of Censoring Weighting
IPTW	Inverse Probability of Treatment Weighting
IPW	Inverse Probability Weighting

LOCF	Last Observation Carried Forward
LSTM	Long Short-Term Memory
MAR	Missing at Random
MCAR	Missing Completely at Random
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
MSM	Marginal Structural Model
NN	Neural Network
PH	Proportional Hazard
RPSFTM	Rank Preserving Structural Failure Time Model
RCGAN	Recurrent Conditional Generative Adversarial Net
RCT	Randomized Controlled Trial
RNN	Recurrent Neural Network
RMSN	Recurrent Marginal Structural Network
RMST	Restrictive Mean Survival Time
SBP	Systolic Blood Pressure
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SNM	Structural Nested Model
SNDM	Structural Nested Distribution Model
SNMM	Structural Nested Mean Model
SNAFTM	Structural Nested Accelerated Failure Time Model
SRA	Sequential Randomization Assumption
TTE	Time-to-event
TMLE	Targeted Maximum Likelihood Estimation
TR	Threshold Regression

Chapter 1: Problem Motivation and Introduction

In longitudinal observational studies, it's not uncommon that the treatment/exposure of interest will influence a time-varying covariate, yet this covariate might influence future treatment/exposure levels. For example, the human immunodeficiency virus (HIV) is a virus that attacks and destroys cells in the immune system. With disease progression, the human body will be increasingly vulnerable to any infections and diseases because too many immune cells are lost due to HIV infection. The HIV viral load density directly decides whether the patient will develop AIDS (acquired immunodeficiency syndrome) or not. Hence, it's critical to have the viral load density under control. By monitoring the viral load at subsequent clinical visits, doctors can assess the disease's progress and prescribe treatments accordingly. In an observational study, the researchers are trying to evaluate the effect of a proposed treatment on prolonging subjects' AIDS-free time or survival. Some evidence have shown that the treatment could lower the HIV viral load density and hence improve subjects' survival. But some unmeasured/unknown confounders, such as undesirable life habits, are also harmful to subjects' health. Consequently, subjects' HIV viral load and hazard of AIDS may grow simultaneously due to these unknown adverse influences. In practice, given the apparent non-response, the practitioners may increase the treatment levels in later clinical visits to control the viral load.

As a result, traditional methods that adjust for time-dependent covariates (viral load in this case) may lead us to the wrong conclusion that the treatment is ineffective. Some methods have been developed to address this time-varying confounding issue. Nevertheless, given the rapidly growing amounts of data collected, previous methods may suffer from model misspecifications or unacceptable computational burdens.

In a broad sense, a problem in survival analysis is the presence of time-varying confounders. The covariate may vary based on previous treatment; it may further influence the treatment regime at future visits. In such a problem, we cannot identify the causality with traditional time-dependent confounder adjustment techniques. Some methods in causal inference (mostly due to Robins and his colleagues) allow us to attack problems like that. Given the difficulties in specifying a *correct* parametric structural model, it may be beneficial to use some nonparametric models such as neural networks, to estimate the causal parameter of interest. In this dissertation, we propose some modern machine learning algorithms to make valid causal inferences in the presence of time-varying confounders.

In Chapter 2, the background of causal inference is given. Specifically, Section 2.1 and 2.2 provide some fundamental background information in causal inference and important required assumptions. Section 2.3 introduces a specific challenge in longitudinal causal inference, the time-dependent confounding issue. Section 2.4 provides necessary notation clarifications for following discussions. Three previous methods dealing with time-varying confounding, including G-formula, Marginal Structural Model (MSM), and Structural Nested Model (SNM), are presented in Section 2.5. These models have the potential to address different types of outcomes. For the survival outcomes, the Accel-

erated Failure Time (AFT) model is considered in this dissertation due to its benefits in quantifying the survival causal effects. The existing Structural Nested Accelerated Failure Time Model (SNAFTM) and the associated estimation methods (G-estimation) are further reviewed in Sections 2.6 and 2.7. Two possible censoring mechanisms and the correct methods to handle these mechanisms are reviewed in Section 2.8 and Section 2.9.

In Chapter 3, concepts of Neural Networks (NN) and literature on related Machine Learning (ML) algorithms, especially those targeted on causal inference, are reviewed. We further discuss the usefulness of deep implementations in estimating structural nested models, specifically for the investigated model - SNAFTM. Then, the major contribution of this dissertation - two ML algorithms for SNAFTM are presented (GE-SCORE and GE-MIMIC). The contributions of the proposed algorithms are the extension of SNAFTM to a high dimension. Extensive simulations are conducted to compare the performance of the existing and proposed methods. The simulations are designed for distinct intervention effects mechanisms.

In Chapter 4, an application of proposed algorithms with real observational health data (Coronary Artery Risk Development in Young Adults Study) is presented. The research question is the smoking causal effect on time to the first CVD event. Smoking is very likely to be confounded by time-varying covariates in an observational longitudinal study; thus methods introduced in Chapter 3 are practical. Estimations from the existing methods and proposed algorithms are compared. The collusion is that the proposed algorithms can provide more clinically profound conclusions.

Chapter 5 presents another survival model, Threshold Regression (TR) which may have similar benefits (collapsibility) as an AFT model in causal survival analysis. The

second part of this dissertation focuses on the First-Hitting-Time based Threshold Regression Model using a Wiener process without considering longitudinal data. A neural network expansion of TR (TRNN) is developed for cases with intervention occurs at the baseline and the confounding does not depend on time. With the assistance of G-computation, the causal estimand of interest (usually the marginal intervention effect) can be easily obtained. Simulations and applications have been supplied to demonstrate the model properties and its possible applications in causal survival analysis.

Chapter 6 provides the conclusion of the dissertation.

Chapter 2: Background on Causal Inference

2.1 Introduction to Causal Inference

Nearly all questions in medical research concern causality. For instance, the researchers would like to investigate whether an intervention can lower patients' blood pressure or whether a medication postpones patients' first heart attack occurrence. To formalize the causal intuition we already possess in daily life, the concepts of *potential outcomes* or *counnterfactual outcomes* are needed. Consider a dichotomous treatment A with two possible values: $A = 0$ or $A = 1$ and a continuous outcome variable Y. A and Y are random variables; and their realizations are denoted by $a; y$. One way to formally define the causal effect of treatment A on outcome Y is:

$$Y^{a=1} - Y^{a=0} \tag{2.1}$$

In plain English, this reads as the difference between the outcome variable that would have been observed under the treatment value $a = 0$, and the outcome variable that would have been observed under the other treatment value $a = 1$. If the difference is not zero, one can conclude that there is a causal effect of treatment on the outcome, or say, A *causes* Y. Similarly, the causal effect can be defined by the ratio of the potential outcomes, or the

logit difference between binary potential outcomes. The causal effects can be explicitly defined according to the different types of treatment/outcome.

However, the format of the causal effect to be detected should be clarified before any investigation. The challenges arise when one recognizes the fact that, for an individual, we can never observe his/her outcome under two different regimes simultaneously. As a result, randomized controlled trials (RCTs) are the gold standard for identifying the causal effect, not the association, between a proposed treatment or a specific exposure and the outcome of interest. If all potential confounders are well controlled by the randomization procedure and study participants are perfectly compliant with the treatment plan, the trial-identified association between the treatment and the outcome can be regarded as causality. However, in many situations, it's unethical or even feasible to conduct an RCT to answer particular research questions. For instance, there should never be any RCTs investigating the adverse effects of smoking on pregnancy. Alternatively, some observational studies collected smoking status and pregnancy information. These data appear promising for assisting researchers in answering questions such as "does maternal smoking affect newborn birth weights?" Furthermore, the results of an RCT may still be suspect if any non-compliance has occurred after the randomization. Another significant shortcoming of an RCT is the high expense and slow turnaround. Indeed, in recent decades, much effort has been expended to make a causal inference with complex observational longitudinal data due to increasing data sizes due to regular (or daily) data collection schema (e.g., electronic health records (EHRs) generated by scheduled physical examinations, data collected by wearable smart devices). Supplied with rich observational data, it's possible for researchers to answer questions regarding the causal relationship between some

exposures (e.g., smoking, drinking, high-fat diet) and health. Of course, there is no free lunch. The prerequisite of valid causal reasoning includes a check of causal assumptions, picking suitable causal models, etc. Once the causal relationship has been identified, more healthy lifestyles can be recommended at the individual level. Note: the terms "exposure", "treatment" and "intervention" will be used synonymously and interchangeably throughout this dissertation.

2.2 Identifiability Assumptions

Causal inference is hard because the counterfactuals required to identify the effects are not directly observable. There are three assumptions required for a valid causal inference: Consistency, Exchangeability and Positivity (Hernán and Robins, 2020; Rosenbaum and Rubin, 1983). As readers will see, RCTs can identify causal effect just because the estimands are obtained when these three assumptions are true. To facilitate the formal definition, we further introduce the random variable L (potentially high dimensional) to denote the covariates often observed in a clinical or an observational study, for example, gender and race of the subjects. Later, the exposure and covariates will be expanded to the time-varying cases and these three assumptions will also be modified accordingly.

2.2.1 Consistency

The values of treatment under comparison correspond to well-defined interventions that, in turn, correspond to the versions of treatment in the data. Consistency ensures that, an individual i with observed treatment A equal to a, has observed outcome Y equal to

his/her counterfactual outcome Y^a:

$$\text{if } A_i = a \text{ then } Y_i^a = Y_i^A = Y_i. \tag{2.2}$$

This assumption links the counterfactuals with the observables. We could safely regard observed outcomes for some subjects as the counterfactual outcomes of others'.

2.2.2 Exchangeability

This assumption is the cornerstone for causal inference. It simply states that the treatment should be assigned independently of the counterfactual outcome.

$$Y^a \perp\!\!\!\perp A \tag{2.3}$$

In marginally (i.e., unconditionally) randomized experiments, the treated and the untreated subjects are exchangeable because the treated had they remained untreated, would have experienced the same average outcome as the untreated did, and vice versa (Hernán and Robins, 2020). This is so because the randomization ensures that the independent predictors of the outcome are equally distributed between the treated and the untreated groups. The counterfactual outcome Y^a, like one's genetic makeup, can be thought of as a fixed characteristic of a person existing before the treatment A was randomly assigned (such as by the flip of a coin). Hence, Y^a encodes what would have been one's outcome if treatment had been applied and thus does not depend on the treatment one actually receives.

This assumption can be generalized to the statement that the conditional probability of receiving any value of treatment depends only on measured covariates L. In other words, the treatment assignment is still decided by the flip of the coin, but the probability that the coin lands head can depend on observed covariates L.

$$Y^a \perp\!\!\!\perp A|L = l \tag{2.4}$$

Hence, this assumption is also called the no unmeasured confounder assumption. The marginal randomization design would ensure the marginal exchangeability and conditional randomization would ensure the conditional exchangeability. The symbol $\perp\!\!\!\perp$ denotes statistical independence. For example $X \perp\!\!\!\perp Y|Z$ means X is conditionally independent of Y given Z (Dawid, 1979).

2.2.3 Positivity

The probability of receiving any value of treatment conditional on L is greater than zero, i.e., positive.

$$\text{If } f_L(l) \neq 0, \text{ then } f_{A|L}(a|l) > 0 \text{ for all a.} \tag{2.5}$$

In an RCT, subjects have positive probabilities of receiving any treatment.

2.3 Time-dependent Confounders

Confounding arises when there is an important influence on the outcome that differs systematically in the comparison of exposed and unexposed groups. It is then possible that

part or all of the apparent observed effect of exposure is due to these differences, in which case the comparison of the exposure groups is said to be confounded (Greenland et al., 1999). A widely noted complication of causal inference with observational longitudinal data is the existence of *time-varying confounding*. The usual approach to estimate the effect of a time-varying treatment on survival is to model the hazard of failure or the time-to-event (TTE) outcome as a function of the past treatment history using a time-dependent Cox proportional hazard (PH) model (Cox, 1972) or model the logarithm of the survival time by a time-dependent accelerated failure time (AFT) model (Wei, 1992). Traditional methods fail to identify the causal effects of the treatment in the presence of any time-varying confounding, whether or not adjusting for past confounder history (Robins, 1986b).

A formal definition of a time-varying confounder is needed before going deeper. A variable is defined as a time-varying confounder for the exposure/treatment if a) it influences subsequent exposure levels; b) it is influenced by previous exposure levels and c) itself is a risk factor for the outcome of interest, either due to a direct causal path flowing from it to the outcome or there is an unmeasured confounder between this covariate and the outcome. For instance, it is well recognized that smoking is a risk factor for cardiovascular disease (CVD) events (Patel et al., 2016). If one is trying to quantify the causal effect of smoking at the time to the subject's first CVD event, one may notice that blood pressure (BP) influences subsequent smoking status - subjects may quit smoking due to increased health awareness after observing an abnormally high BP. On the other hand, smoking can raise BP, and BP itself is also an important risk factor for CVD events. In conclusion, BP is a time-dependent confounder for smoking on subjects'

time to their first CVD events. Below, a simplified causal Directed Acyclic Graph (DAG) (Pearl, 2009a) with a time-varying treatment at two different time points demonstrates the bias of the traditional method graphically.

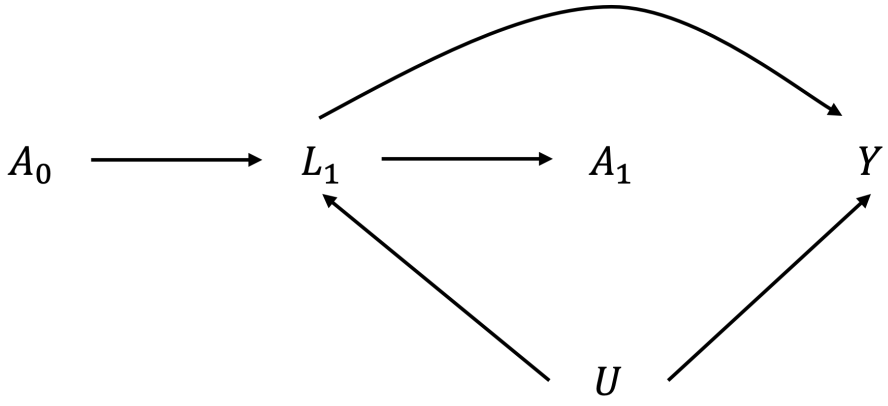


Figure 2.1: A simplified DAG of time-varying smoking status effects on time to first CVD, under the null hypothesis.

The above DAG presents a covariate L at time 1 and treatment at two distinct time points for demonstration purposes. First note that L is a time-dependent mediator for exposure A as it's in the causal pathway from A_0 to Y. U is an unmeasured confounder for L and Y. The DAG represents a null hypothesis that there is no direct causal effect of A on Y, encodes by the absence of direct path runs from A to Y. Once condition on L_1, then a backdoor path $A_0 \rightarrow L_1 \leftarrow U \rightarrow Y$ is opened. Also, the effects flow from $A_0 \rightarrow L_1 \rightarrow Y$ is adjusted away if L_1 has direct effects on Y. If one does not adjust for L_1, the backdoor path $A_1 \leftarrow L_1 \leftarrow U \rightarrow Y$ remains open. The dilemma is whether adjusting for L_1, one cannot identify the unbiased total causal effect of the time-varying A on the outcome Y.

It's important to note that, in practice, one must rely on the disease experts and/or epidemiologists who have the subject matter knowledge to provide the correct form of the

DAG. Testing the validity of the DAG falls out of the scope of this dissertation.

2.4 Data Structure and Notations

This section clarifies the necessary notations that are needed to discuss the unbiased methods to identify the causal treatment effects with time-varying confounding. In an observational study, the data are usually collected longitudinally. Visits are scheduled at discrete time points $t_0, t_1, \cdots, t_K$ with visit number $k = 0, 1, \cdots, K$, where 0 denotes the baseline visit and K denotes the largest possible visit number in the study. Throughout the follow-up, the outcome of interest is monitored. In this dissertation, the outcome will be recorded only once as our research interest is the time to the first event. Let V be the subject's last visit before the event or end of his/her follow-up, whichever comes first, given the fact that some subjects do not experience all visits in a longitudinal study. For one subject i, $i = 1, \cdots, N$, a risk factor could be a time-dependent random process $L_i(t)$, furthermore, the value at a visit k is recorded as $L_{i,k}$, potentially high dimensional. A high-dimensional process $L_i(t)$ can, of course, include time-invariant elements like gender and race. Similarly, the exposure level is a random process $A_i(t)$ that is recorded at each visit by $A_{i,k}$. Another assumption required for the following discussions is that the risk factor/exposure processes jump at and only at the measured times, where process $A(t)$ jumps right after $L(t)$ is observed at visit k. The subject subscript i is suppressed for the following discussion and it should not cause confusion. Vectors are not necessarily in bold fonts.

Let us start with an exposure having binary values such as $A_k = 0$ or 1, which corre-

sponds to exposed or unexposed. The notations $\bar{A}_k = (a_0, \cdots, a_k)$ and $\bar{L}_k = (l_0, \cdots, l_k)$ are used to record the exposure or covariate histories (realizations) up to the visit k. Some examples of treatment regimes are "always treated": $\bar{A} = (1, \cdots, 1)$ and "never treated": $\bar{A} = (0, \cdots, 0)$. A dynamic treatment regime is a treatment given at visit k may depend on covariate histories up to this visit, such as $A_k = g(\bar{l}_k)$. This is particularly common in practice: for example, a treatment for AIDS is only rendered when the HIV viral load is above a certain threshold. For simplicity we assume that the variables L_k and A_k take their values in countable sets, denoted by $\mathcal{L}_k$ and $\mathcal{A}_k$. Lastly, $\mathcal{G}$ denotes the set of all possible treatment regimes, which is countable but may not be finite. We define maps $\bar{g}_k : \mathcal{L}_0 \times \cdots \times \mathcal{L}_k \to \mathcal{A}_0 \times \cdots \times \mathcal{A}_k$ by $\bar{g}_k(\bar{l}_k) = (g_0(l_0), g_1(\bar{l}_1), \ldots, g_k(\bar{l}_k))$.

The outcome of interest is the time from study entry to the first occurrence of a pre-defined endpoint Y. However, the endpoint may not be observable due to censoring. Subjects may be lost to follow-up, or simply do not experience the event before the scheduled study ending date. The resultant observable outcome is $T = \min(Y, C)$, where C is the censoring time. We will postpone the discussion of censoring until Section 2.8. In the spirit of counterfactual outcomes, $Y^{\bar{a}}$ denotes the outcome that would be observed if subjects follow the treatment regime $\bar{A} = \bar{a}$, potentially contrary to fact. The total set of observations are N independent and identically distributed (i.i.d.) observations from the distribution of the random vector $(T, \bar{L}_K, \bar{A}_K)$.

2.5 Existing Methods to Adjust for Time-varying Confounders

The *G-null hypothesis* of interest to be tested with time-varying regimes is

$$Y^{g_1} = Y^{g_2} \text{ for any } g_1, g_2 \in \mathcal{G}. \tag{2.6}$$

There exist three classes of models dealing with time-varying confounding (Daniel et al., 2012), developed mostly by Robins and his colleagues: Marginal Structure Model (MSM) (Robins et al., 2000), G-formula (Robins, 1986a) and Structural Nested Model (SNM) (Robins et al., 1992). Robins and his collaborators (Robins and Hernán, 2008) suggested that these models should be employed instead of traditional time-dependent analysis methods (e.g., a Cox model adjusting for the time-dependent covariate) in any epidemiological studies if the causal estimands are the primary interest.

With enriched history notations, a more comprehensive DAG for the longitudinal data with the outcome observed by end of the study is provided (Figure 2.2). Possible causal paths are plotted with gray arrows and the study hypothesis is denoted by dashed arrows.

The likelihood for observed data O can be partitioned as

$$f_O(o) = f_{Y|\bar{L},\bar{A}}(y, \bar{l}, \bar{a}) \prod_{k=0}^{K} f_{L_k|\bar{L}_{k-1},\bar{A}_{k-1}}(l_k, \bar{l}_{k-1}, \bar{a}_{k-1}) \prod_{k=0}^{K} f_{A_k|\bar{L}_k,\bar{A}_{k-1}}(a_k, \bar{l}_k, \bar{a}_{k-1}) \tag{2.7}$$

These three methods in fact concern different parts of the likelihood. G-computation focuses on estimating the first and second elements, whereas the weights used in MSM are

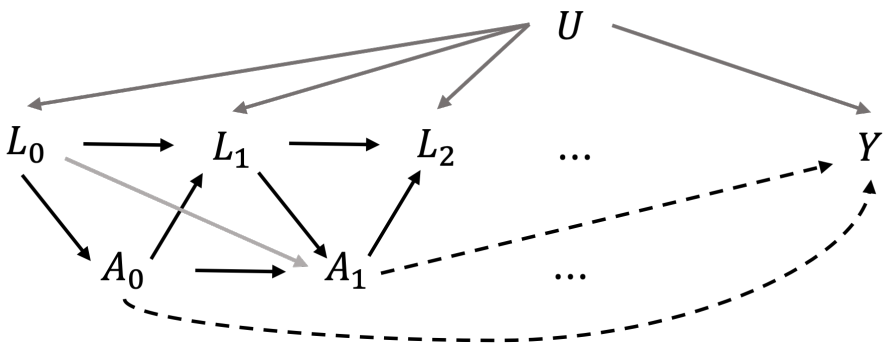


Figure 2.2: A Comprehensive DAG with more Visits and Possible Causal Paths and Study Hypothesis

defined by the last element. The reweighting method used for MSM is also known as Inverse Probability of Treatment Weighting (IPTW). G-formula (also called G-computation/standardization) directly estimates the potential outcome $E(Y^{\bar{a}})$ by simulating the joint distribution of the variables L, A, and Y that would have been observed in a hypothetical study where every subject follows the treatment regime $\bar{A}$. The algorithm is highly computationally demanding when the possible regime set is large. A more difficult situation arises when the treatment level is continuous, such as the injection volume. On the other hand, MSM provides valid causal inference by breaking the connection between the time-depending confounder and the subsequent exposure. More precisely, a pseudo population is created by IPTW and the analytic model is fitted to the pseudo population. Hence, the estimated parameter is the same as the proposed analytic model (e.g., hazard ratio estimated by the Cox PH model) but provides an unbiased causal interpretation. SNM can be viewed as a compromise of these two methods. It is better tailored to deal with the failure of required causal assumptions, does not suffer from the g-null paradox as G-formula does, and is more efficient than the IPTW method with (near) positivity violations. It avoids

the technical difficulties of specifying the law of covariates, the model estimation utilizes the treatment assignment scheme only and has been proved semi-parametrically efficient (van der Laan and Robins, 2003). In general, there are two classes of SNM - the Structure Nested Mean Model (SNMM) and Structure Nested Distribution Model (SNDM) (Vansteelandt and Joffe, 2014).

It has been shown that, these three classes of methods yield identical estimates in some special cases, such as when both time-varying treatment and time-varying confounder are binary (or more generally, categorical) and the number of visits is low (Daniel et al., 2012). The limited number of possible treatment and covariate histories allow the causal parameter of interest to be estimated nonparametrically. However, in more realistic settings with unsaturated models, the parametric models that are needed for each method are different. From the simulation perspective, data can be generated to satisfy all models (Young et al., 2009). A similar data generation scheme as used in (Young et al., 2009) is applied in this dissertation, see Section 3 for more details. Indeed, estimands in these three methods are often different. Their efficiencies as well as implementation difficulties vary across scenarios.

This dissertation focuses on causal inference in survival analysis. For a survival outcome, a specific sub-class model of SNDM - the Structural Nested Accelerated Failure Time Model (SNAFTM) (Robins, 1992, 2005; Robins and Hernán, 2008) has to be used. To clarify, this model is also called the Structural Nested Failure Time Model (SNFTM) in some literature.

2.6 Structural Nested Accelerated Failure Time Model

2.6.1 Structural Nested Accelerated Failure Time Model (SNAFTM)

In survival analysis, one commonly would like to make inferences about the treatment/exposure effects on the time-to-event outcome. Under the counterfactual outcome framework, researchers have attempted to estimate the potential outcomes (time-to-event) directly, that is, the contrary to fact survival times under different exposure regimes. The G-formula is a useful tool for this purpose. Nevertheless, censoring is an intrinsic problem in the survival analysis, which further complicates the direct estimation of counterfactuals. Though the potential outcome is often of research interest, another important class of problems is how to quantify the exposure effect with a few parameters, and whether the effect is statistically significant at the population level. Notably, the SNAFTM equipped with G-estimation serves these research purposes perfectly (Robins, 1992, 2005; Robins and Hernán, 2008). For decades, the Cox Proportional Hazards (PH) model has been criticized for lacking causal meanings(Aalen et al., 2015; Hernán, 2010). The SNAFTM utilizes another common time-to-event model - the AFT model, which circumvents this problem by modeling the relationship between survival time and prognosis factors directly (Cox and Oakes, 1984; Wei, 1992). Though the AFT model itself may not be the most common survival model used in medical research, it is collapsible, which is a desirable property for causal inference. See more information of collapsiblity in Section 5.2.

In an AFT model, the logarithm of the survival time Y is fitted by linear regression

on covariates X such as:

$$\log(\mathbf{Y}) = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.8}$$

If the residual term $\boldsymbol{\varepsilon}$ has a known parametric distribution such as extreme value, normal or logistic distribution, then the AFT model is the usual Weibull, Log-normal, or Log-logistic parametric survival model. A semiparametric AFT model is one with an unspecified residual distribution. The interpretation of the estimated coefficients in the AFT model is straightforward. For example, in the results of an AFT model with mortality as the endpoint and a covariate denoting the treatment group, the covariate coefficient could be interpreted as a certain percentage increase in expected life expectancy under the treatment compared to the control. For instance, a patient could be informed that she/he would be expected to live (say) 20% longer if she/he took the treatment, given the estimated coefficient = $\log(1.2)$. This interpretation can be understood more naturally given the fact that any AFT model can be written as:

$$Y = Y_0 \cdot \exp(\mathbf{X}'\boldsymbol{\beta}), \tag{2.9}$$

where $Y_0 = e^{\boldsymbol{\varepsilon}}$ is the baseline survival time - the survival time without exposure.

Now, we are at a good position to consider using the survival time ratio $\exp(-\psi^*) = Y^{a=1}/Y^{a=0}$ to measure the causal effect of a binary treatment A on each subject's time-to-event outcome with a constant ψ^* for subjects and time points. Throughout this dissertation, ψ^* stands for the true value of ψ, and ψ_0 stands for the estimated value of ψ, under the one-dimensional case. The ψ is potentially high-dimensional because it may contain

a finite vector of parameters. Furthermore, the treatment effect may vary over time when the effect is modified by a time-dependent covariate. The survival time ratio $\exp(-\psi^*)$ is the expansion (or contraction) in survival time attributable to treatment. Hence, this measure can be used to describe a structural AFT model. It is a structural model because the counterfactual outcomes are modeled, rather than observed. A positive value of ψ^* indicates that a subject's survival time when treated or exposed is shorter than the survival time when untreated (i.e., treatment is harmful) whereas a negative value indicates that the survival time when treated is longer than the survival time when untreated (i.e., treatment is beneficial). When $\psi^* = 0$, there is no causal effect of treatment on survival time (Hernán et al., 2005). This echoes the AFT structure in Formula (2.9). One can view $Y^{a=0}$ as the baseline survival time (i.e., the survival time should be without exposure). Often, $Y^{a=0}$ is considered as a genetic marker for an individual, such as hair color.

On top of the structural AFT model, when the treatment/covariates are allowed to vary at each visit $k, k = 0, \cdots, K$. At each visit k, we are able to consider two potential outcomes. The first one is the outcome that would be observed for subject i if he/she followed some treatment regime $\bar{a}$ up to visit k but all treatment is withheld onward (i.e., $Y_i^{a_0,a_1,\cdots,a_k,0,\cdots,0}$); the second is the outcome that would be observed of the same subject following the same regime up to visit $k-1$ and no treatment onward (i.e., $Y_i^{a_0,a_1,\cdots,a_{k-1},0,0,\cdots,0}$). To ease the notation, we use $Y^{\bar{a}_k,\bar{0}}$ to denote $Y^{a_0,a_1,\cdots,a_k,0,\cdots,0}$. Conceptually, there are $K+1$ sub-models in a longitudinal study with K visits. Each model links two potential outcomes defined above. The population in each model is a subset of the one constructed at the previous visit. This is where the "Nested" in SNAFTM comes from.

2.6.2 Blip function

Formally, the *blip function* (Robins, 1997) $\gamma(y, \bar{l}_k, \bar{a}_k)$ can be used to quantify the treatment causal effect on the outcome Y of a final brief "blip" of treatment a_k in the interval $(t_k, t_{k+1}]$.

Definition 2.6.1. A blip function γ is a well-defined function that satisfies the following relationship:

$$P(Y^{\bar{a}_k,\bar{0}} > y|\bar{l}_k, \bar{a}_k, Y > t_k) = P(Y^{\bar{a}_{k-1},\bar{0}} > \gamma_k(y, \bar{l}_k, \bar{a}_k)|\bar{l}_k, \bar{a}_k, Y > t_k). \tag{2.10}$$

Intuitively, the blip function is a transformation *subtracts off* the effect of the treatment on the interval $(t_{k-1}, t_k]$, creating a new outcome that is similar to the potential outcome under no treatment during those same periods. Precisely, this is the "blip-down" function (which Robins refers to as impulse response functions (Robins, 1997)). The "blip-up" function can be defined in a similar manner. In this dissertation, only the blip-down function is used hence we simply call it the blip function. The blip function is a quantile to quantile transformation function and it satisfies the following conditions:

1. $\gamma(y, \bar{l}_k, \bar{a}_k) > t_k$.
2. If $a_k = 0$, $\gamma(y, \bar{l}_k, \bar{a}_k) = y$.
3. $\gamma(y, \bar{l}_k, \bar{a}_k)$ is increasing in y.
4. The derivative of $\gamma(y, \bar{l}_k, \bar{a}_k)$ with respect to y is continuous almost everywhere.

The population follows a SNAFTM or blip model $\gamma^*(y, \bar{l}_k, \bar{a}_k; \psi)$ with respect to L (Robins, 2002) if $\gamma(y, \bar{l}_k, \bar{a}_k) = \gamma^*(y, \bar{l}_k, \bar{a}_k; \psi)$ where $\gamma^*(y, \bar{l}_k, \bar{a}_k; \psi)$ is a known function and ψ is a finite vector of parameters to be estimated taking values in $\mathbb{R}^\nu$. For each value of $\psi \in \mathbb{R}^\nu$, $\gamma^*(y, \bar{l}_k, \bar{a}_k; \psi)$ satisfies four conditions stated above for a blip function. Furthermore, if $\psi = 0$, $\gamma(y, \bar{l}_k, \bar{a}_k, \psi) = y$; $\partial\gamma^*(y, \bar{l}_k, \bar{a}_k; \psi)/\partial\psi'$ and $\partial^2\gamma^*(y, \bar{l}_k, \bar{a}_k; \psi)/\partial\psi'\partial y$ are continuous for all ψ and almost all $y \in (t_k, \infty)$. For the sake of simplicity, we use $\gamma(y, \bar{l}_k, \bar{a}_k; \psi)$ to denote the indexing blip function, which is a function of exposure/covariate histories up to visit k and the parameter to be estimated is ψ. The blip function $\gamma()$ will be enriched to $\gamma_k()$ providing the time-varying nature and nested trial concept. Though the blip function can be applied to any interval of subjects at risk, given $y > t_k$, we consider only the *final* blip, that is, for $y \in (t_k, t_{k+1}]$.

Formally, the Structural Nested Accelerated Failure Time Model (Robins, 2002) can be defined as follows:

Definition 2.6.2. Structural Nested Accelerated Failure Time Model is a model that maps percentiles y of the conditional distribution of the failure time $Y^{\bar{a}_k,\bar{0}}$, given $\bar{L}_k = \bar{l}_k$ and $\bar{A}_k = \bar{a}_k$ into $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi)$ percentiles of the conditional distribution of $Y^{\bar{a}_{k-1},\bar{0}}$, given $\bar{L}_k = \bar{l}_k$ and $\bar{A}_k = \bar{a}_k$, and for subjects who are still at risk (say, alive) at time t_k.

$$S_{Y^{\bar{a}_{k-1},\bar{0}}|\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{a}_k,Y>t_k}\{\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi)\} = S_{Y^{\bar{a}_k,\bar{0}}|\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{a}_k,Y>t_k}(y). \tag{2.11}$$

In the SNAFTM, applying the blip function to the failure time outcome Y removes the effect of a final blip of treatment of magnitude a_k at visit k on the survival outcome. If the blip function is applied to $Y^{\bar{a}_k,\bar{0}}$, it leaves us with $Y^{\bar{a}_{k-1},\bar{0}}$. From the observed

time Y^{a_k}, with recorded exposure/covariates history, one can reach $Y^{\bar{0}}$ - the "treatment-free" counterfactual by applying the blip function recursively until all treatment effects are removed. By using the blip function to link the counterfactual and the observable, the treatment causal effect can be summarized with a few parameters, which further facilitates hypothesis testing.

The simplest example of a blip function in SNAFTM is $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi) = t_k + \int_{t_k}^{y} \exp(\psi a(u))du$. Note, this blip function satisfies conditions specified in the definition and the true ψ^* quantifies the causal survival time ratio between two counterfactual survival outcomes on the log scale, as in the AFT model. If $\psi^* = 0$, there is no causal treatment effect on the outcome. A more complicated blip function can be $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi_1, \psi_2) = t_k + \int_{t_k}^{y} \exp((\psi_1 + \psi_2 l(u))a(u))du$. Now, there is an interaction effect between the covariate and treatment at visit k, and the null treatment effect is encoded as $\psi_1 = \psi_2 = 0$. For instance, it's plausible that the exposure (smoking) effect is larger for a heavy smoker compared to a light smoker. This is explained by allowing the smoking causal effect to be modified by smoking intensity.

The structural AFT model discussed as the example so far is a subclass of structural distribution models - the rank-preserving SNAFTM (RPSNFTM) (Robins, 2002; Robins and Tsiatis, 1991), which coincides with a strong version of the AFT model (Cox and Oakes, 1984). It's called rank-preserving in the sense that given any two subjects i and j, if i fails before j when both followed a particular treatment regime, then subject i will also fail before j if both follow some other treatment regimes. The blip functions shift the potential outcomes themselves rather than their distributions. This class of models is seldom biologically plausible. However, it is easy to communicate and hence, serves the

demonstration purpose well. The untestable rank-preserving assumption can be relaxed (Robins et al., 1992), also, the G-estimation method discussed in Section 2.7 can be implemented if the data follow an AFT model, regardless of whether it's rank-preserving or not (Robins, 2002, 2005; Robins et al., 1992). The blip function in a non-rank-preserving SNAFTM maps the distribution of counterfactual outcomes, not the outcomes themselves, hence does not assume rank preserving.

To close this section, we restate that SNAFTM handles the time-varying confounding bias by never lumping treatments received at different times. SNAFTM allows one to directly model the interaction between treatment at each visit k and past time-dependent covariate history. The SNAFTM sheds more light on answering causal survival questions, especially when time-varying confounders are in the scene.

2.7 G-estimation

The next problem naturally is how to estimate the causal parameter ψ in the SNAFTM and make inferences. The identifying assumptions mentioned in Section 2.2 need to be revised slightly to address this specific type of causal problem (Robins and Hernán, 2008), because now there are time-varying exposures and covariates. The Exchangeability assumption in (2.2.2) can be adapted to the Sequential Randomization Assumption (SRA) (Robins and Hernán, 2008): for all regimes $\bar{a}$, all $\bar{l}_k$ and for $k = 0, \cdots, K$,

$$A_k \perp\!\!\!\perp Y^{\bar{a}_{k-1},\bar{0}} | \bar{L}_k, \bar{A}_{k-1}, Y > t_k. \tag{2.12}$$

This assumption means that the treatment at visit k will be randomly assigned independently of the counterfactual unexposed outcome, but the randomization probabilities may depend on L and A histories prior to treatment at visit k is given. The required independence will only be examined for subjects who are still alive/free of endpoint events at the beginning of each nested trial. Conceptually, each subject contributes a number $(V_i + 1)$ of visits to the final analysis.

The consistency assumption (2.2.1) is updated to say, for any individual,

$$\text{if } \bar{A} = \bar{a} \text{ then } Y^{\bar{A}} = Y^{\bar{a}} = Y. \tag{2.13}$$

The positivity assumption (2.2.3) is updated as to say, for all possible levels of a_k, if

$$f(\bar{L}_k, \bar{A}_{k-1}) > 0, \text{ then } f(a_k|\bar{L}_k, \bar{A}_{k-1}) > 0. \tag{2.14}$$

For dynamic regimes, the positivity assumption stays the same while for exchangeability and consistency, a "strengthened" version is needed. The SRA is updated as to say, for any k, treatment regime $g \in \mathcal{G}$ and $\bar{l}_k$,

$$A_k \perp\!\!\!\perp Y^g | \bar{L}_k, A_{k-1} = g_k(\bar{L}_{k-1}, \bar{A}_{k-2}), Y > t_k \tag{2.15}$$

The strengthened consistency reads, for any individual, any k, and treatment regime $g \in \mathcal{G}, \bar{l}_k$,

$$\text{if } A_k = g_k(\bar{L}_k, \bar{A}_{k-1}), \text{ then } Y^g = Y \text{ and } \bar{L}_K^g = \bar{L}_K, \tag{2.16}$$

where $\bar{L}_k^g$ is the counterfactual L-history through visit k under regime g.

Then, one proper approach to test the G-null hypothesis (2.6) is based on the following theorem (Robins, 1997):

Theorem 1 (G-null Theorem)**.** *The G-null hypothesis (2.6) is true if and only if* $A_k \perp\!\!\!\perp Y|\bar{L}_k, \bar{A}_{k-1}$ *for all* $k = 0, \cdots, K$.

The G-null theorem only refers to the observables; however, it still has implications for the distributions of counterfactual outcomes (Robins, 1997). Combining SRA and the blip function, the G-null theorem indicates that there is no effect of final blip of treatment of size a_k among subjects with each observed history $\bar{l}_k, \bar{a}_k$ [i.e., $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi) = y$] if and only if the G-null hypothesis is true. Thus, the treatment causal effect can be tested via a ν dimensional statistical test of $\psi = 0$.

The simplest blip function $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi) = t_k + \int_{t_k}^{y} \exp(\psi a(u))du$ is used as a running example in following discussion. The first step in G-estimation is to calculate the counterfactual treatment-free outcome with the hypothesized known blip function. If the blip function is correctly specified, the true unexposed counterfactual variable $Y^{\bar{0}}$ can be recovered from the observed quantities $\{Y, \bar{A}_k, \bar{L}_k\}$ by the relationship

$$Y^{\bar{0}} = \int_0^Y \exp\{\psi^* A(u)\}du. \tag{2.17}$$

The counterfactual can be written as a function of ψ. For individual i with a given ψ, we can use function $H(\psi)$ to make the relationship between $Y^{\bar{0}}$ and the parameter of interest

ψ explicit:

$$H_i(\psi) = \int_0^{Y_i} \exp\{\psi A_i(u)\} du. \tag{2.18}$$

Furthermore, consider the nested trial concept. Conditional on $\{\bar{L}_k = \bar{l}_k\}, \bar{A}_{k-1} = \bar{a}_{k-1}$, the survival time up to time t_k becomes a constant and will be always independent of A_k. As a result, it contributes no information to the required conditional independence testing (2.12). Then, $H_i(\psi)$ will only be computed for subjects who are still alive at time t_k with

$$H_i(\psi, k) = t_k + \int_{t_k}^{Y_i} \exp\{\psi A_i(u)\} du. \tag{2.19}$$

This function shows that the counterfactual outcome can be computed by a deterministic function and the mapping from $\{Y_i, \bar{A}_k, \bar{L}_k\}$ to $\{H_i(\psi, k), \bar{A}_k, \bar{L}_k\}$ is one to one. It's also easy to observe that $Y_i^{\bar{0}} = H_i(\psi^*, 0)$. Based on this one-to-one mapping, the observed likelihood $\mathcal{L}\{Y_i, \bar{A}_k, \bar{L}_k\}$ can be written with respect to the counterfactual outcome by variable transformation as:

$$\prod_{i=1}^{N} \left\{ \frac{\partial H_i(\psi)}{\partial Y_i} \right\} f\{H(\psi); \eta\} \prod_{k=0}^{V_i} f\{L_{i,k} | \bar{L}_{i,k-1}, \bar{A}_{i,k-1}, H_i(\psi); \phi\} \prod_{k=0}^{V_i} f(A_{i,k} | \bar{L}_{i,k}, \bar{A}_{i,k-1}; \alpha). \tag{2.20}$$

We focus on specifying the correct parametric form of treatment density function while allowing $f\{H(\psi)\}$ and $f\{L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, H(\psi)\}$ to remain completely unrestricted. The following relationship is true based on the SRA:

$$f\{A_k | \bar{L}_k, \bar{A}_{k-1}, H(\psi^*, k)\} = f\{A_k | \bar{L}_k, \bar{A}_{k-1}\}. \tag{2.21}$$

When the treatment A is a binary variable, a pooled logistic regression has been proved useful in testing the above-stated conditional independence and obtaining a semi-parametric estimation of ψ (Hernán et al., 2005). Once $H(\psi, k)$ is obtained for a specific $\psi = \psi_0$, for all subjects, the term $H(\psi_0, k)$ is added as a covariate to the pooled logistic regression which fits the distribution of A_k. The Markov assumption specified that the treatment probability at visit k is associated only with the most recent treatment/covariate history:

$$\text{logit}(P(A_k = 1|\bar{L}_k, \bar{A}_{k-1})) = \alpha_0 + \alpha_1 L_k + \alpha_2 A_{k-1} + \theta H(\psi_0, k). \tag{2.22}$$

An asymptotic α-level G-test of the hypothesis $\psi_0 = \psi^*$ is performed by using any standard likelihood-based test (e.g., a score test) of the hypothesis $\theta = 0$ based on the likelihood of the pooled logistic model such as in Formula (2.22). These tests are often called the *G-tests*. The 95% confidence interval of ψ_0 is the set of ψ_0 when the hypothesis that $\theta = 0$ cannot be rejected. Of course, a similar logic can be applied when A is no longer binary, but the implementation requires a more sophisticated model specification (Robins et al., 1992). In conclusion, if the density of treatment $f(A_k|\bar{L}_k, \bar{A}_{k-1})$ can be correctly specified, SRA implies that the estimated coefficient (e.g., θ) before the extra term $H(\psi^*, k)$ added to the model should be 0. In other words, the true ψ recovers the latent unexposed counterfactuals, which are conditionally independent of A_k given histories, at each k.

Essentially, estimating the causal parameters in the structural models is equivalent

to solving the estimating equation

$$U(\psi) = \sum_{i=1}^{N} U_i(\psi, \hat{\alpha}) = 0, \tag{2.23}$$

where ψ is the causal parameter of interest (e.g., the causal survival time ratio) and $\hat{\alpha}$ is the consistent estimator of a p-dimensional nuisance parameter α. In the SNAFTM, α denotes the parameters of a parametric distribution that describes the treatment mechanism. For the running example with binary treatment, the explicit G-estimating equation is:

$$U(\psi) = \sum_{i=1}^{N} \sum_{k=0}^{V_i} H_i(\psi, k)\{A_{i,k} - Pr(A_{i,k} = 1 | \bar{L}_{i,k}, \bar{A}_{i,k-1})\} = 0. \tag{2.24}$$

The consistency estimation of ψ by the G-estimation requires the correct specification of the exposure mechanism $f(A_k)$, in addition to the correctness of "known" blip function γ and the SRA. For high-dimensional cases such as $\psi = \{\psi_1, \psi_2, \cdots, \psi_\nu\}$, ν estimating equations can be posed and a χ^2 test statistic can be used to test the conditional independence (Joffe et al., 2011; Robins, 1992).

Last, The SNM and G-estimation were developed for settings in which treatment decisions are being made at discrete times at which all the confounders are observed. Essentially, SNMs (and Robins' other longitudinal causal inference techniques) assume discrete-time underlying data generating processes. There is a fundamental difficulty in the extension to the continuous case because the theory involves conditional distributions that are not uniquely defined for continuous variables. However, the method can be salvaged under certain continuity assumptions (Gill and Robins, 2001). The continuous

data-generating process is not considered in this dissertation. We also refer readers to other continuous case generalizations (Zhang et al., 2011).

2.8 Artificial Censoring

In the previous discussion, the counterfactual survival time $H_i(\psi, k)$ can only be calculated for subjects with observed event times. If the event time is right censored due to scheduled study close out, one may want to apply the same rule but with Y replaced with $T = \min(Y, C)$ which is considered right-censoring in equation (2.17). However, this will introduce bias.

Figure 2.3 aims to provide a simple example that explains the bias and suggested remedy. Assume there are six subjects, three come from the treated group, and the other three are from the untreated group. These subjects are exchangeable in terms of treatment group as the baseline survival times in each group are both $(2, 4, 6)$ (black dots). Further, panels a and b showed that, without administrative censoring, the treatment is beneficial with an acceleration factor equal to 1.5 because the observed survival time for the treated group (red dots) is $(3, 6, 9)$. For the untreated group, the observed survival time is the same as the baseline survival time. Referring back to the G-estimation section, a correct $\psi = -\log(1.5)$ can restore the exchangeability from the observed data (red dots are mapped back to black dots). Panels c and d present the cases when the study is scheduled to end at time 5, the black and red dots still denote the baseline and observed survival time. Now, with the event indicator added, for the untreated group, the observed data are (2,1), (4,1), and (5,0) and for the treated the data are (3,1), (5,0), and (5,0). Even if the

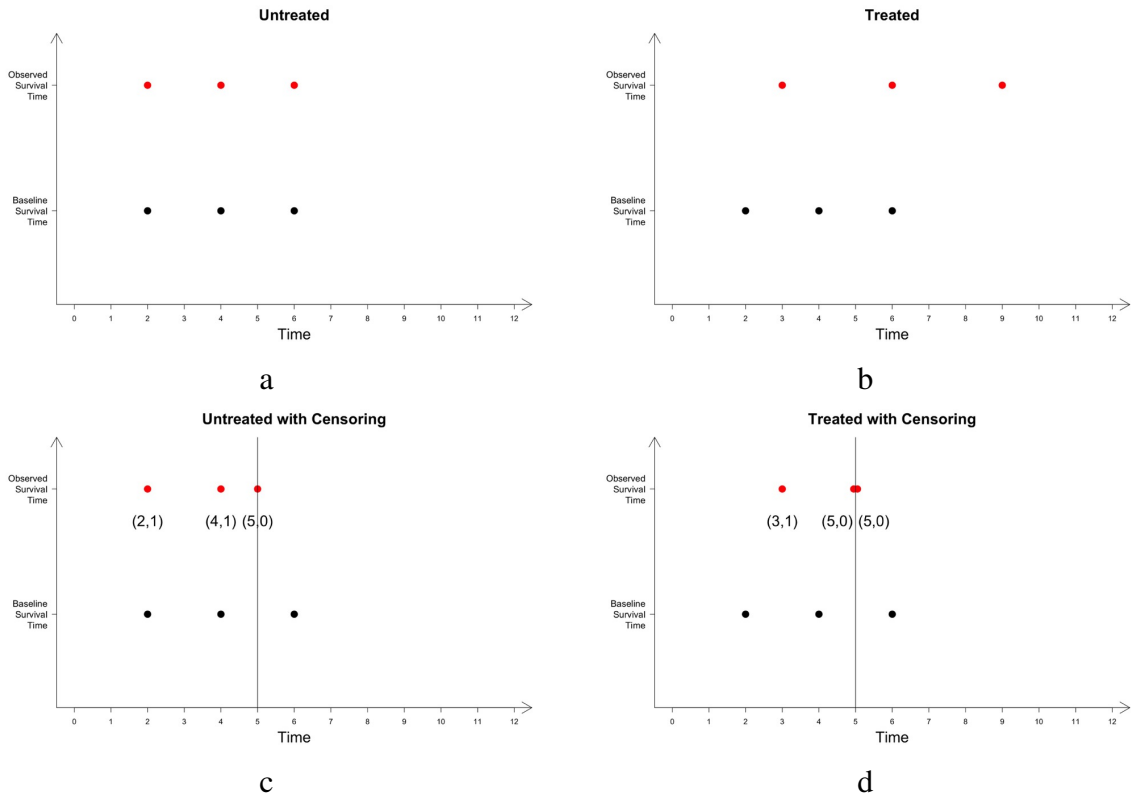


Figure 2.3: Artificial Censoring Example: red dots denote the observed survival times, black dots denote the latent baseline survival times. Baseline survival time are (2, 4, 6) for three subjects in all plots. Treated subjects would have longer survival time (a & b). If censoring at time 5 is further considered, not all events can be observed. Numbers in brackets annotating each red dot in plots c & d are observed event time and event indicator.

true ψ^* is applied to the observed survival time, the exchangeability cannot be recovered as one has reason to believe that people from the treatment group are more likely to be censored and have a smaller baseline survival time. Mathematically speaking, even with $Y_i^0 \perp\!\!\!\perp A_{i,k}$, $C_i \not\perp\!\!\!\perp A_{i,k}$ as the censoring status has a direct relationship with the treatment regime unless the treatment causal effect is 0.

A solution is to use the artificial censoring mechanism which censors some subjects even if they experienced the event. To define artificial censoring, first, replace the event time Y_i in (2.19) with the right censoring time C_i. This information should be available

at the beginning of the study. Then, define the *minimal potential follow-up time* as

$$C_i^*(\psi, k) = \min_{\forall g \in \mathcal{G}} \{t_k + \int_{t_k}^{C_i} \exp\{\psi A(u)\} du\}. \tag{2.25}$$

As the name suggests, $C_i^*(\psi, k)$ denotes the minimal potential follow-up time under all possible treatment regimes at visit k, with respect to a specific ψ. $C_i^*(\psi, k)$ is a fixed function of $\{\bar{L}_k, \bar{A}_{k-1}\}$ but not a function of A_k. Referring to the working example, there are two possible treatment regimes: treated versus untreated. When the ψ is hypothesized as the true value, $-\log(1.5)$, the minimal potential censoring time is obtained when subjects are exposed, which is $5/1.5$. Now, the artificial censored bivariate random variable {time, event indicator}: $\{X_i(\psi, k), \Delta_i(\psi, k)\}$ can be calculated as usual but replacing $C_i(\psi, k)$ with $C_i^*(\psi, k)$. Hence, $X_i(\psi, k) = \min(H_i(\psi, k), C_i^*(\psi, k))$ and $\Delta_i(\psi, k) = \mathbf{1}\{H_i(\psi, k) < C_i^*(\psi, k\}$. In the working example, the constructed artificially censored observations are $\{(2, 1), (5/1.5, 0), (5/1.5, 0)\}$ for the untreated group, and $\{(2, 1), (5/1.5, 0), (5/1.5, 0)\}$ for the treated group. With the original censoring time 5 replaced by the new censoring time $(5/1.5)$, the exchangeability is restored when $\psi = \psi^*$. The newly defined counterfactuals $\{X_i(\psi, k), \Delta_i(\psi, k)\}$ are calculable for all subjects at all visits and they obey SRA when $\psi = \psi^*$. For computational convenience, the artificial event indicator $\Delta_i(\psi)$ will be used in the G-estimation in replace of $H_i(\psi, k)$, to solve for ψ. The consistent and unbiased estimator of ψ can then be obtained.

The SRA holds for any deterministic p-dimensional function $g_k()$ of $H_i(\psi, k)$, and in fact, for any deterministic ν-dimensional function g_k of $\{X_i(\psi, k), \Delta_i(\psi, k)\}$. The choice of g_k should not influence the consistency of the estimator but the efficacy. The

optimal g_k which attains the semiparametric variance bound is provided in (Robins, 1992). However, the optimal estimator is not feasible since it depends on the unknown joint distribution of $\{Y, L(Y), A(Y)\}$. There are some other possible choices that provide double robustness (Mertens and Vansteelandt, 2017; Vock et al., 2013) against misspecification of the exposure mechanism. Nevertheless, the artificial censoring indicator $\Delta_i(\psi, k)$ is much easier to calculate and hence is more popular in practice.

The problems of artificial censoring include the non-smoothness of the resultant estimating equation, potential loss of information, etc (Joffe et al., 2011). As a result, the usual gradient-based algorithms such as Newton-Raphson cannot be applied to G-estimation. Some non-gradient-based algorithms, such as grid searching or Nelder-Mead have to be employed. The target parameter ψ in the simplest blip function is suggested to be estimated by grid search (Hernán et al., 2005), i.e., grid search values on $[-3, 3]$ by a step of 0.01. An oversimplified blip function is very likely to fail to quantify the true causal effect of the exposure. The causal effect may not be linear and there could be complicated interaction effects between covariates and exposure, as well as among their histories. The ranges considered in the grid search algorithm should be wide enough to cover the true value as well as the CIs.This requirement, in turn, renders the algorithm computationally unfeasible for a high-dimensional $\boldsymbol{\psi}$ in a more sophisticated blip function. Some remedies have been proposed to address these issues (Joffe et al., 2011). Yet, these remedies are still computationally intensive, fail to find the solution in high-dimensional cases, and further require a large amount of expert knowledge. These existing obstacles point out the need to develop more stable, user-accessible estimation algorithms for SNAFTM.

2.9 Competing Risk Censoring

Besides the often-encountered right-censoring, subjects may drop out due to other competing risk events. If a subject's loss to follow-up is due to moving out of the area, it's safe to regard them as regular right censored subjects because the censoring scheme is independent of the counterfactual survival outcome. However, if a subject drops out due to illness that makes him or her unable to be compliant with the study, this makes this independent censoring assumption suspicious. To deal with this type of censoring, we further need the Ignorable Censoring Assumption (ICA). ICA states that the observed history $(\bar{L}_k, \bar{A}_{k-1})$ is sufficient to predict censoring at visit k in the sense that censoring carries no residual information about survival under any of the considered exposure regimes, conditional on the covariate history up to visit k (Robins, 1992). Inverse Probability of Censoring Weighting (IPCW) can deal with competing risk censoring in G-estimation under ICA. The weights $W_{i,k}$ are the inverse of an individual's probability of being uncensored until the end of the study, at visit k (Hernán et al., 2005). The competing risk censoring probability at each visit k can be predicted by a model of the censoring mechanism. Using IPCW allows adjusting for confounding by creating a pseudo-population in which the competing risk censoring is independent of the measured confounders. In the pseudo-population, an uncensored subject not only represents himself or herself but also similar but censored subjects. This idea is similar to using weighting for non-response to adjust for missing data (Little and Rubin, 2002).

Specifically, the censoring scheme can be fitted by a logistic (if there is only one censoring mechanism) or multinomial (if there is more than one censoring mechanism)

regression. The censoring model uses all available data to relate the probability of being censored at each measurement to the exposure and covariate history. Then, the estimating equations that define the G-estimator are restricted to subjects who remain uncensored from the competing risk events during follow-up. These subjects are weighted by the reciprocal of the cumulative probability of not being censored until the end of the follow-up. Some doubly robust methods have been proposed to address possible misspecifications of the censoring mechanism (Mertens and Vansteelandt, 2017), for traditional G-estimation.

In any application, the event that qualifies as a competing risk event should be carefully defined. For instance, if the event of interest is the CV death, then deaths due to other causes are competing risk events. Another example is, to examine the causal effect of high blood pressure on CVD events, individuals are competing risk censored when they first report the use of any anti-hypertensive medications (Witteman et al., 1998). Theoretically, other than the scenario of a study defining a pre-fixed treatment regime and using all-cause death as the endpoint, competing risk censoring is possible.

2.10 Conclusion and Other Causal Inference Perspectives

The SNAFTM along with G-estimation has been successfully applied to many studies (Hernán et al., 2005; Robins et al., 1992; Witteman et al., 1998). However, this method is still not popular due to multiple reasons. First, a user has to specify the *correct* form of the blip function, which could be a non-trivial task. Second, there is no ready algorithm to compute the causal parameter. The estimating equation of the causal parameter (especially for the high-dimensional case and when censoring must be taken care of) is

not monotone or differentiable with respect to the parameter to be estimated (Joffe et al., 2011; Vansteelandt and Joffe, 2014). As a result, a non-gradient-based optimizer has to be used. However, the commonly used grid search does not guarantee a unique solution and so the procedure is computationally intensive and inconvenient. This computational challenge further restricts the blip function's possible specifications - only estimation for a low dimensional ψ is feasible. Third, the estimation method heavily relies on a strong but untestable assumption, the sequential randomization assumption. A remedy is to include as many covariates as one can in the analysis to mitigate the risk of assumption violation. Finding a suitable parametric model for the treatment mechanism with rich covariates collected in modern studies could be a prohibitive task in the beginning, and could suffer from the curse of dimensionality. The Markov assumption may permit a comparable simple formula such as (2.22), but it's obviously too restrictive to agree with clinical practice. Nevertheless, if an important confounder is left out during the data collection period of an observational study, the assumption violation is inevitable. Fourth, G-estimation resultant confidence intervals (CIs) can be uselessly wide or even infinite when the SRA assumption is violated. Robins (Robins, 1997) specified the conditions where the CIs are finite and proposed some practical guides. But, the existing practical guides are still vague for day-to-day analysis and require a large amount of side input. The optimal choice of g_k which attains the semiparametric variance bound is not accessible and barely used.

If the null hypothesis of no treatment effect has been rejected and the causal parameter ψ has been estimated, one might wish to estimate the survival distribution $Pr(Y^g > y)$ of the outcome under specific treatment regimes g in a way consistent with the estimator ψ_0. This can be done by estimating the distribution of $Y^{\bar{0}}$ (e.g. by the empirical distribu-

tion of) and the empirical distribution of L_k given $\bar{L}_{k-1}, \bar{A}_{k-1}$ and 0 $(k = 0, \cdots, K)$ for histories $\bar{L}_{k-1}, \bar{A}_{k-1}$ consistent with g.An approximate sample $\tilde{T}_i^g$ from the distribution of T^g could then be generated by using these estimated distributions by the Monte Carlo algorithm (Lok et al., 2004; Robins et al., 1992).

Besides Robins' counterfactual outcome framework, there are other causal inference perspectives. For example, because the counterfactual outcomes are not always observable, it's natural to think it's a missing data problem (Ding and Li, 2018). Some Bayesian methods have been proved useful as well (Oganisian and Roy, 2020; Rubin, 1978). However, these discussions are beyond the scope of this dissertation. The main contribution of this dissertation is extending the SNAFTM/G-estimation to the deep learning field and developing useful machine learning algorithms. Starting from the next chapter, it's assumed that the required causal inference background has been well elaborated and discussions will be mostly on the computational side.

Chapter 3: Neural Network based Algorithms for SNAFTM

3.1 Background

Computers/machines learning from data to perform a task by using algorithms instead of being explicitly programmed is called Machine Learning (ML). In a broad sense, deep learning (LeCun et al., 2015) is a subfield of ML. Deep learning includes algorithms that analyze data with a layered structure called an Artificial Neural Network (ANN) (Goodfellow et al., 2016). ANN's history can be traced back to 1943 when McCulloch and Pitts published the first mathematical model of a neural network (McCulloch and Pitts, 2021). ANN can be viewed as an estimation method that models the relationship between inputs and outputs using layers of connected computational units (neurons), patterned after the biological neural networks of brains. Computational units in the network - neurons - sit between the inputs and outputs. A hidden layer neuron receives its inputs from the neurons in the preceding layer of the network, adds up each input multiplied by the corresponding weight, and passes the result on to neurons in the next layer with an activation function applied. The network structure allows data-driven learning of the appropriate model, in addition to learning the parameters of that model (Farrell et al., 2021). ANNs make up the backbone of deep learning algorithms. This estimation method has become increasingly popular in causal discovery given its capability to capture the complex

dependencies between variables by using a great amount of available data.

Deep learning techniques are useful tools for clinical data analysis in many senses. On the one hand, observational data can be used to train a supervised learning model. However, the model may lack causal meanings if it fails to adjust for the bias introduced by time-varying confounders. Alternatively, given potentially complicated covariates and treatment distributions, along with their temporally connected nature at consecutive time points, the three aforementioned methods in Section 2.5 all suffer from model misspecification. To improve estimation consistency and efficiency, a natural thought is to develop some deep learning algorithms for existing methods that address the time-varying confounding issue. This Chapter starts with reviewing related literature. The proposed algorithms will then be mathematically justified and described in detail.

3.1.1 Recurrent Neural Network

To start with, ANNs such as Recurrent Neural Networks (RNNs) have achieved state-of-the-art performance in processing sequential data or time series data. The most important property of RNN is its ability to *remember* the history. At each time step t, the network takes accumulated information at the prior step - the hidden state h_{t-1} - besides current covariates z_t as input. While traditional deep neural networks assume that inputs and outputs are independent of each other, the feedback loops in RNN allow its output to depend on the prior elements within the sequence. Depending on the input and target output sequence lengths, there are four types of RNN (Liu, 2020), many-to-one, many-to-many, etc. Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is

the most popular RNN variant as it contains a special unit (memory cell) in addition to standard units to maintain information in memory for long periods. Hence, LSTM has some advantages in learning long-term temporal dependencies. RNN and its variants have become impressive benchmarks in solving ordinal or temporal problems, such as language translation, natural language processing, and speech recognition (Sherstinsky, 2020).

Figure 3.1 displays how a RNN takes the input sequence $\{z_t\}_{t=1}^{T}$ and outputs a prediction sequence $\{y_t\}_{t=1}^{T}$. Here, z_t and y_t can be vectors at each time point. The widely

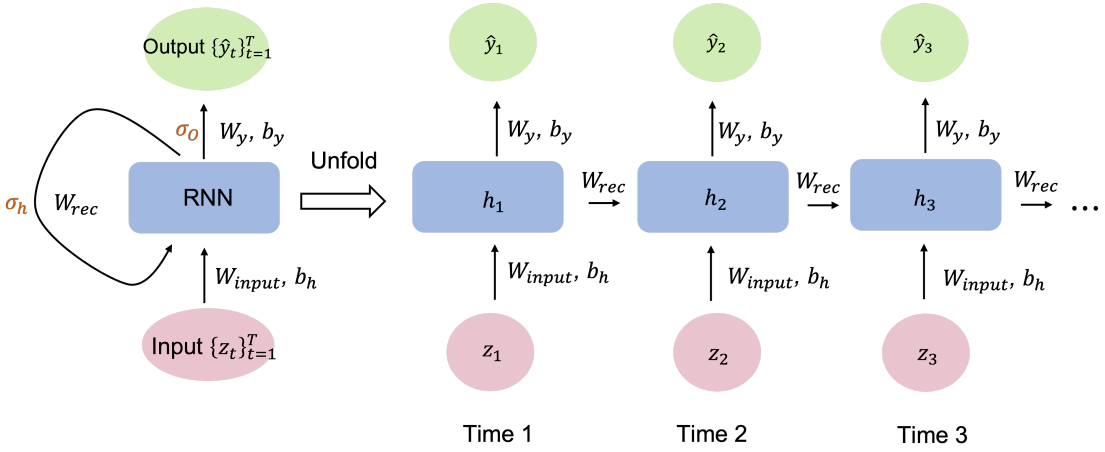


Figure 3.1: An example unfolded many-to-many RNN.

known equation below describes how the simple RNN unit takes the current covariates input z_t and the previously hidden state h_{t-1} when generating the currently hidden state h_t. In this case, the model parameters are the recurrent weight matrix $\boldsymbol{W}_{rec}$, the input weight matrix $\boldsymbol{W}_{input}$ and the bias b_h.

$$h_t = \sigma_h(\boldsymbol{W}_{rec}h_{t-1} + \boldsymbol{W}_{input}z_t + b_h). \tag{3.1}$$

The many-to-many RNN is capable of producing output $\hat{y}_t$ from the current hidden state h_t by further introducing an output weight matrix $\boldsymbol{W}_y$ and the bias associated with output - b_y.

$$\hat{y}_t = \sigma_o(\boldsymbol{W}_y h_t + b_y). \tag{3.2}$$

Element-wise activation functions σ can be chosen accordingly for the generation of the hidden state and the output estimation, as σ_h in (3.1) and σ_o in (3.2). The loss function $\mathcal{L}$ of all time steps is defined based on the loss at every time step as follows:

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T} \mathcal{L}(\hat{y}_t, y_t). \tag{3.3}$$

We refer readers to (Goodfellow et al., 2016) for a complete treatment of RNN.

From the perspective of the gradient information in optimization, popular neural network optimization methods can be divided into three categories: first-order optimization methods, which are represented by the widely used stochastic gradient methods; high-order optimization methods, in which Newton's method is a typical example; and heuristic derivative-free optimization methods, in which the coordinate descent method is a representative. In this dissertation, we focus on the first category, the gradient-based optimization methods. Gradient Descent is generally attributed to Cauchy back to 1847 (Cauchy, 1847). The gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. The algorithm proceeds by taking repeated steps in the opposite direction of the gradient of the objective function at the current step during the iteration. Furthermore, the Stochastic Gradient Descent (SGD)

is preceded by a stochastic approximation, as first described by Robbins and Monro in the 1950s (Robbins and Monro, 1951). The idea of stochastic gradient descent is using one sample randomly to update the gradient per iteration, instead of directly calculating the exact value of the gradient. The stochastic gradient is an unbiased estimate of the real gradient (Robbins and Monro, 1951). The algorithm learning rate determines the step size in each iteration, and thus influences the number of iterations to converge. Besides SGD, there are many other first-order optimization methods such as AdaGrad (Duchi et al., 2011), RMSProp (Tieleman and Hinton, 2015) and Adam (Kingma and Ba, 2015). Specifically, Adam combined benefits from adaptive gradient methods AdaGrad and RMSProp; thus it has gained increasing popularity in recent years. Its advantages include the magnitudes of parameter updates are invariant to rescaling of the gradient, its stepsizes are approximately bounded by the stepsize hyperparameter, it does not require a stationary objective, it works with sparse gradients, and it naturally performs a form of step size annealing (Kingma and Ba, 2015). For a review of optimization methods used in the contemporary machine learning community, see paper (Bottou et al., 2018; Ruder, 2016; Sun et al., 2020).

Clinical data differ from other sequence data as they often include a mixture of static (e.g., race, gender) and dynamic information, which is recorded at each visit (e.g., lab measurements, medications prescribed) (Esteban et al., 2016). During the past decades, RNN along with its variants have been successfully applied to medical longitudinal data analysis and outperformed other methods in future event prediction. Given the fact that the medical records are growing longer (more visits) and wider (more information is collected at each visit), the capability of handling high-dimensional data makes the neu-

ral network one of the most competitive tools in medical data analysis. In the causal inference field, it's also not uncommon that researchers sought help from the ML community (Chernozhukov et al., 2018). The time-varying confounding issue only persists for longitudinal data. Hence, RNN and its variants are popular tools when researchers attempt to solve causal problems. To this end, Li et al. (Li et al., 2020) proposed the G-Net, a deep-learning approach to G-computation. Lim et al. (Lim et al., 2018) introduced the Recurrent Marginal Structural Networks (RMSN) which combines MSMs with RNN. On the basis of Lim et al.'s work, Counterfactual Recurrent Network (CRN) (Bica et al., 2020) was developed to make the counterfactual estimation of future outcomes. CRN constructs treatment invariant representations at each time step to remove the bias from time-varying confounders. The CRN also has adversarial training layers which break the association between subjects' history and treatment assignment. Thus, the network can be reliably used for estimating counterfactual outcomes. The algorithm can also estimate counterfactual outcomes for treatment plans (not just single treatment). The counterfactual outcomes can then be used in answering clinical questions such as when to initiate/terminate the treatment.

3.1.2 ML Assisted AFT

Some deep applications of the AFT model (Martinsson, 2016; Yang et al., 2017) may be helpful because the SNAFTM is inherited from the AFT. Nevertheless, these existing deep AFT algorithms assume specific parametric forms for the survival distributions, for instance, the Log-normal or Weibull distribution. The loss functions are either

defined with respect to a specific likelihood of the data or the mean square error (MSE) between predicted and observed survival outcomes. As shown in Chapter 2, the traditional time-dependent AFT model fails to identify the true treatment causal effect in the presence of time-varying confounders. In contrast, algorithms proposed in this chapter target relaxing the known parametric assumption of the blip function as well as the treatment mechanism, while keeping the inference under the SNAFTM framework. The goal is to provide an unbiased estimation of the intervention's causal effect after controlling for the time-varying confounding issues.

3.1.3 ML Assisted Conditional Independence Testing

The three existing methods discussed in Section 2.5 adjust for time-varying confounding have different estimands and are developed to answer distinct research questions. SNAFTM has advantages in survival causal inference as demonstrated in Section 2.6. Though the hybridization of G-computation/MSM with ML algorithms seems successful, a similar deep implementation of G-estimation has not been discovered. The gap is of interest to be filled in this dissertation.

To embrace deep learning methods into the G-estimation method, we first noted that conditional independence stated in Formula (2.12) is the cornerstone of G-estimation. In this dissertation, the hypotheses to be tested based on (2.12) can be written as: for each visit k, $k = 0, \cdots, K$,

$$H_0 : A_k \perp\!\!\!\perp Y^{\bar{a}_{k-1},0,\cdots,0} | \bar{L}_k, \bar{A}_{k-1} \text{ vs. } H_A : A_k \not\perp\!\!\!\perp Y^{\bar{a}_{k-1},0,\cdots,0} | \bar{L}_k, \bar{A}_{k-1}, \tag{3.4}$$

where $Y^{\bar{a}_{k-1},\bar{0}}$ stands for the exposure-free counterfactuals as before. Further, let $\bar{Z}_k = \{\bar{L}_k, \bar{A}_{k-1}\}$ denote all treatment/covariate histories observed just before A_k is experienced. Then the above formula can be re-written as:

$$H_0 : A_k \perp\!\!\!\perp Y^{\bar{a}_{k-1},\bar{0}} | \bar{Z}_k \text{ vs. } H_A : Y^{\bar{a}_{k-1},\bar{0}} \not\perp\!\!\!\perp A_k | \bar{Z}_k. \tag{3.5}$$

This null hypothesis can be additionally written with respect to conditional distributions such as:

$$H_0 : Pr(A_k | \bar{Z}_k, Y^{\bar{a}_{k-1},\bar{0}}) = Pr(A_k | \bar{Z}_k) \sim q_{H_0}(A_k). \tag{3.6}$$

We observe that the hypothesis testing problem is actually a conditional independence testing problem. This requirement of conditional independence in causal inference naturally motivates us to consider taking advantage of the recent advances in conditional independence tests (CIT) using NNs. Nevertheless, current literature focuses on the CIT and outputs a valid p-value, that is, to perform the CIT of $H_0 : X \perp\!\!\!\perp Y|Z$ versus $H_A : X \not\perp\!\!\!\perp Y|Z$ with the assistance of NNs. To name a few of these CIT algorithms, the Model-X knockoffs (Candès et al., 2018) (Conditional Randomization Test; CRT) algorithm aims at testing $X_j \perp\!\!\!\perp Y|X_{-j}$ by constructing *knockoffs* $\tilde{X}_j$ of the variable X_j. The knockoff samples are drawn from the complete conditional $P(X_j|X_{-j})$, namely, the distribution of the j^{th} feature given all other features. The complete conditional serves as a valid null by preserving the joint dependency structure of $Pr(X)$ and removing any dependency between X_j and Y. Finally, the CRT compares a test statistic $T(\tilde{X}_j; X_{-j}; Y)$ on the null samples to the statistic on the observed sample $T(X_j; X_{-j}; Y)$ and yields a

p-value. On the bases of CRT, the Holdout Randomization Test (HRT) (Tansey et al., 2021) works by first splitting the data into training $\{X; Y\}$ and testing $\{X'; Y'\}$ sets. A predictive model π is fit on the training set and repeatedly evaluated on the testing set. The prediction quality $\pi(X')$ is compared to $\pi(\tilde{X}')$, where $\tilde{X}'$ is a copy of X' that replaces the feature X'_j to be tested with a null sample $\tilde{X}'_j$. A p-value for the hypothesis test is then approximated by repeatedly sampling and comparing predictive performance under the null with performance using the original data. Nevertheless, it's often hard to generate null samples required by these algorithms as there is no access to the true null data generation mechanism. Recent advances in Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) provide some new working directions here. GAN is a two-player game in which a generator model is trained to generate new samples, and a discriminator model tries to classify examples as either real or fake (generated). The generator and the discriminator models are trained together until the discriminator model is fooled about half the time, meaning the generator model has become creditably good at generating plausible samples. GAN allows the production of realistic samples across a range of problem domains, such as realistic photos of objects, scenes, or people. A well-trained GAN is able to produce photos very similar to genuine ones. Even humans cannot tell which are fake. Specifically, the Generative Conditional Independence Test (GCIT) (Bellot and van der Schaar, 2019) is a test that is based on a modification of GAN that generates null samples from distribution under the conditional independence assumption (H_0), while maintaining good power in high dimensional data. The GCIT rejects the conditional independence hypothesis when the observed correlation is at extreme values of its empirical null distribution, i.e., at the 5% tail of the approximated empirical null

distribution.

3.2 Notations

Though the notations needed in this section largely follow the notations specified in Section 2.4. For readers' convenience, we specify them again here. The data are assumed to come from a longitudinal study, with $k = (0, 1, \cdots, K)$ follow-up visits at discrete time points $t_k, k = (0, 1, \cdots, K)$, t_K is the fixed maximum scheduled visit time and $t_0 = 0$ denotes time of the baseline visit. Subjects are usually under follow-up after the last visit for a pre-specified duration and the maximum follow-up time is denoted by t_{K+1}. Due to the limited follow-up time, administrative censoring is expected. The observed event time is hence $T = \min(Y, C)$, where Y is the true survival time and $C = t_{K+1}$.

For individual i, at each visit k, data are collected on a treatment variable A_k, a vector of covariates L_k, an outcome event indicator variable E_k, and a follow-up indicator variable C_k. We assume the following implication for related variables: $L_k \rightarrow A_k \rightarrow \{C_{k+1}, E_{k+1}\}$. As before, $A_k \in \{1, 0\}$ indicates that a subject was exposed or not at visit k. Variable $C_k \in \{1, 0\}$ indicates whether a subject was under follow-up or was lost to follow-up at or before visit k. Variable $E_k \in \{1, 0\}$ indicates whether a subject acquired the event of interest by visit k. Variable E_k has an absorbing state = 1 (i.e., E_k cannot transit from 1 to 0 once $E_k = 1$). Once $E_k = 1$, $T = Y$ with $t_{k-1} < Y < t_k$. Note, E_k will only be observable at visit k if $C_k = 1$. Variable C_k has an absorbing state = 0. At the baseline visit $k = 0$, we define $E_0 = 0, C_0 = 1$ and L_0 as a vector of baseline variables. Further, $A_{-1} = L_{-1} = 0$. As before, let $\bar{Z}_k = \{\bar{L}_k, \bar{A}_{k-1}\}$ denote all histories

observed just prior to A_k. To describe the data needed for RNN, we use Z_k to denote the covariate random variable at each time point (potentially high dimensional), $\{z_k\}_{k=0}^{K}$ is the observed sequences, serve as the input sequence to the RNN.

Still, γ is reserved for the blip function. The blip function is a function of $\bar{Z}_{\gamma,k}$ where $\bar{Z}_{\gamma,k}$ is a potential subset of the history $\bar{Z}_k$. We use $Y^{\bar{a}_{k-1},\bar{0}}$ to denote the counterfactual of interest - the exposure-free outcomes. Equivalently, the counterfactual can be denoted by $H(\hat{\psi}_k, k)$ which makes the relationship between ψ and counterfactual outcomes explicit. The index i for the individual is omitted unless necessary. Vectors are not necessarily bold to ease notations.

3.3 Mathematical Foundations of Proposed Algorithms

To attack the limitations of G-estimation discussed in Section 2.10, we propose to use neural networks to generate the counterfactual outcome sequence of interest and to estimate the causal parameters in the associated SNAFTMs. NNs are also used to improve the efficiency in testing conditional independence. These proposed networks will be optimized with respect to the properly chosen loss functions which capture the conditional correlation between the counterfactual outcome and exposure at each time point, given histories before the exposure A_k. Specifically, this dissertation proposed two deep-learning algorithms for SNAFTM. The procedures of these two algorithms are clearly illustrated in Figure 3.4, 3.5, where readers shall see the connections and differences among them. Both algorithms begin by utilizing an RNN to estimate the sequence of the causal survival ratio, and further the sequence of counterfactual outcomes. Formulas of

this estimation step will be given in this section. These two proposed algorithms differ by the losses employed (discussed further in Section 3.4), in other words, the mechanisms to enforce conditional independence. As shown in Section 3.1.3, G-estimation is closely related to the CIT. We further argue that, by minimizing the loss to zero, the network would estimate the true causal parameter ψ^* in the SNAFTM. The resultant counterfactual outcomes would assure the conditional independence required for valid causal inference. Detailed structures of these algorithms will be given in Sections 3.5, 3.6.

These two proposed algorithms both use an RNN to first estimate the sequence of the causal survival ratio $\hat{\psi}_k$ with $\bar{Z}_{\gamma,k}$, and then the counterfactual outcome indexed by the SNAFTM. Let Z_γ be the set of covariates that are believed to be associated with the acceleration factor used in the blip function, and let $\bar{Z}_{\gamma,k}$ denotes the history of the corresponding covariate. For example, $\bar{Z}_{\gamma,k}$ can be the history of a one-dimensional time-dependent covariate L_k such as smoking density (packs/year) collected at each visit k because we suspect that the smoking intensity directly relates to the compressing rate of cardiovascular event-free time. Elements of Z_γ can also include some static covariates. It's also reasonable to suspect that males and females respond differently to the same exposure, so adding gender to Z_γ can have this biological assumption well absorbed. If no interaction effect between a_k and other prognostic variables is assumed, $\bar{Z}_{\gamma,k}$ can simply be a one-dimensional constant for all time steps. With Z_γ, the network is capable of remembering the history and recursively updating the hidden state at each time step:

$$h_k = \sigma_h(\boldsymbol{W}_{rec} h_{k-1} + \boldsymbol{W}_{input} Z_{\gamma,k} + b_h). \tag{3.7}$$

The hidden state can be further transferred to any desired dimension by a time-distributed layer. Hence we can obtain an output sequence:

$$\hat{\psi}_k = \sigma_\psi(\boldsymbol{W}_\psi h_k + b_\psi). \tag{3.8}$$

By changing the activation function of the time-distributed layer, we can control the output range of prediction. A popular activation function for RNN hidden state σ_h is `tanh`, for σ_ψ is `sigmoid`. In an RNN, though only the current covariate vector $Z_{\gamma,k}$ enters the model at each time step, history $\bar{Z}_{\gamma,k-1}$ has been absorbed into the hidden state h_{k-1} at previous steps. Eventually, the network will receive the complete sequence $\bar{Z}_K$ and the model prediction at each step is based on history. It's important to emphasize again that RNNs are designed to hold past or historic information of sequential data. In this sense, the network has memory. Note, $\hat{\psi}_k$ technically ranges from $-\infty$ to $+\infty$. However, the empirical values observed from the data are usually in a more compact range.

A popular variant of a simple RNN unit is the LSTM unit which contains multiple gates and cells. In total, there are six equations in LSTM. At a single time point, the LSTM unit takes the input Z_t, h_{t-1} and c_{t-1} that are calculated from previous step, the unit output $\hat{y}_t$ and also generates the h_t and c_t for the consumption of the next time step of the LSTM. More specifically, let $\boldsymbol{U}$ denote the weight matrix that connects the inputs to the hidden layer and let $\boldsymbol{W}$ denote the the weight matrix that connects the previous hidden layer and the current hidden layer. We can describe three gates - input gate, forget

gate, and output gate in LSTM by equations:

$$
\begin{aligned}
i_t &= \sigma(\boldsymbol{W}_i h_{t-1} + \boldsymbol{U}_i Z_t + b_i), \\
f_t &= \sigma(\boldsymbol{W}_f h_{t-1} + \boldsymbol{U}_f Z_t + b_f), \\
o_t &= \sigma(\boldsymbol{W}_o h_{t-1} + \boldsymbol{U}_o Z_t + b_o).
\end{aligned}
\tag{3.9}
$$

As the names suggest, the input gate controls the amount of new information that will be saved in the cell state, forget gate decides how much information will be discarded from the cell state, and the output gate decides the final output for this LSTM unit, at time step t. The cell state and candidate cell state follow the equations:

$$
\begin{aligned}
c'_t &= \sigma(\boldsymbol{W}_c h_{t-1} + \boldsymbol{U}_c Z_t + b_c), \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot c'_t.
\end{aligned}
\tag{3.10}
$$

The hidden state of this time step directly relates to the current cell state:

$$
h_t = o_t \cdot \sigma(c_t). \tag{3.11}
$$

Finally, the output $\hat{y}_t$ is a function of the hidden state h_t.

Compared to the over-simplified SNAFTM with a constant acceleration factor, it's more biologically plausible to model $\hat{\psi}_k$ by incorporating individual exposure/covariate histories. The RNN fulfills the goal perfectly. The many-to-many RNN used in the counterfactual estimation process is summarized in Figure 3.2. Clearly, with sequence input of covariates $z_{\gamma,k}$, $\hat{\psi}_k$ is now a function of covariates, hence it should not be treated as a

constant across all visits, for all subjects. The magnitude of current covariates' influence is controlled by the input weights $\boldsymbol{W}_{input}$. The $\hat{\psi}_k$ has accounted for covariates history, as well as exposure history by adjusting the weights $\boldsymbol{W}_{rec}$ before the hidden state h_t. The overall influence of covariates and their histories on the causal parameter $\hat{\psi}_k$ is further controlled by $\boldsymbol{W}_{\psi}$ when the hidden state is decoded to the target estimand. This allows a nonlinear format between all inputs and target output. The output layer which is attached to the time distributed layer performs as the blip function and calculates the counterfactual survival time for each individual if treatment is withheld onward from visit k. To name some blip function forms that can be represented by this RNN, the interaction blip function we saw earlier $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi_1, \psi_2) = t_k + (y - t_k) \exp((\psi_1 + \psi_2 l_k) a_k)$ can be rewritten as $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi_1, \psi_2) = t_k + (y - t_k) \exp(\hat{\psi}_k a_k)$ where $\hat{\psi}_k = \psi_1 + \psi_2 l_k$. We could view the simplest blip function example where $\gamma_k(y, \bar{l}_k, \bar{a}_k; \psi) = t_k + (y - t_k) \exp(\psi a_k)$ as an oversimplified version of (3.8) where $b_\psi = \psi$, $\boldsymbol{W}_\psi$ is a zero matrix and the activation function is the identity transformation. The interaction blip function mentioned above can be decomposed in a similar manner. If the RNN cell only has one unit, the weights $\boldsymbol{W}_{input}$ before all covariate other than l_k will be 0, $b_h = 0$, no histories need to be addressed hence $\boldsymbol{W}_{rec} = 0$, both activation functions σ_h and σ_ψ can be identity function. Last, let $\boldsymbol{W}_\psi = \mathbf{1}$ and $b_\psi = \psi_1$, so that the simplest scenario can be adequately addressed.

Nevertheless, the blip function characterized by this layer

$$\gamma_k(y, \bar{l}_k, \bar{a}_{k-1}; \hat{\psi}_k) = t_k + (y - t_k) \exp(\hat{\psi}_k a_k); t_k < y \leq t_{k+1} \tag{3.12}$$

still satisfies the requirements stated in Section 2.6. This blip function indexes a SNAFTM

in which the sequence of counterfactual survival time sequence can be calculated as

$$H(\hat{\psi}_k, k) = Y^{\bar{a}_{k-1},\bar{0}} = t_k + \int_{t_k}^{Y} \exp\{\hat{\psi}_u a_u\} du \tag{3.13}$$

The counterfactual survival time following the SNAFTM is calculated backward supplied with the estimated sequence $\hat{\psi}_k$ and observed event time. Though the $\hat{\psi}_k$ has a much more

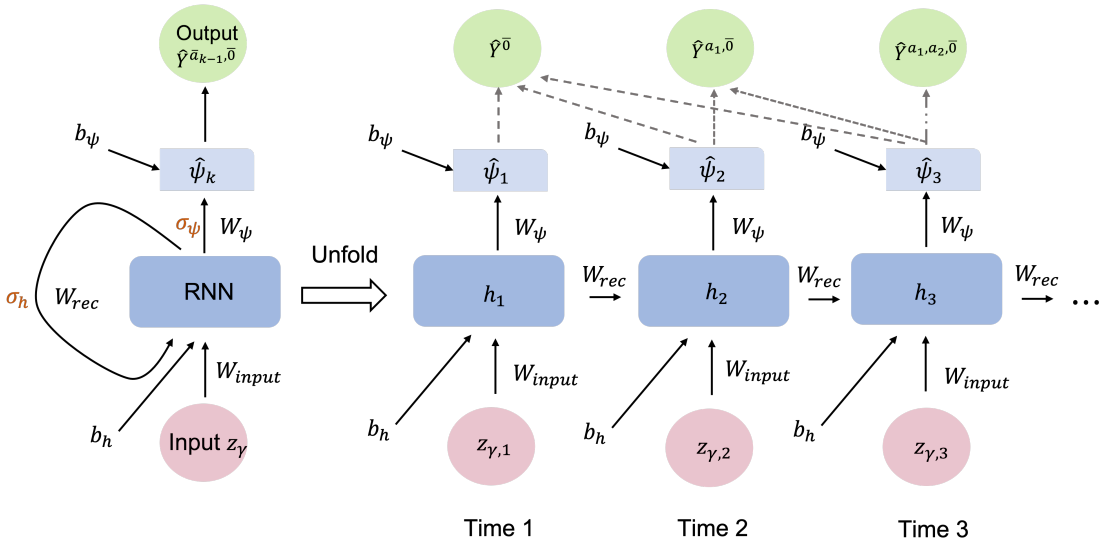


Figure 3.2: The unfold many-to-many RNNs are used to predict the counterfactual outcomes, in both proposed algorithms.

extensive meaning than the previous constant ψ, the interpretation would stay mostly the same. This parameter still characterizes the effect that suspending treatment a_k at visit k is the change to the residual lifetime $t_K - t_k$ with a sequence of factors $\exp(\hat{\psi}_k a_k)$. Note, t_k will be the same for all subjects who are still at risk at visit k, hence, using a simplified counterfactual outcome without the term t_k is sufficient in training the network. This many-to-many RNN outputs a sequence of the remaining counterfactual exposure-free survival times $\{Y^{\bar{a}_{k-1},\bar{0}}\}$. As we have discussed in Section 2.8 that calculating counter-

factual unexposed time with the right censored time creates selection bias, artificial censoring has been incorporated into the output layer of this counterfactual outcome network. The minimal potential follow-up time is defined as:

$$C(\hat{\psi}_k, k) = \min\Big\{t_k + \sum_{u=k}^{K}(t_{u+1} - t_u)\exp(\hat{\psi}_u a_u)\Big\}; k = 0, \ldots, K. \tag{3.14}$$

For any subject with a given potential censoring time (which is known at the beginning of the trial), the artificial censoring time is nothing but the earliest time on the transformed scale at which there can be censoring. Here, the same time scale transformation that applied to event time was applied to the potential censoring time. While training the network, the minimum is taken over all empirical $C(\hat{\psi}_k, k)$ values for subjects in a batch, recursively at each visit k. We follow (Joffe et al., 2011) to use the empirical censoring value instead of the theoretical value, as the gap between these two quantities can be large in practice. We would like to artificially censor fewer subjects thus preserving more information. Another effort made to decrease the artificial censoring rate in practice is to use the largest potential censoring time for all subjects (Robins et al., 1992). Due to staggered entry, subjects tend to have different entering times but when calculating the artificial censoring time, the minimum should be taken with respect to subjects with common original potential censoring time. Replacing the potential censoring time for all subjects with the largest potential censoring time (i.e., t_{K+1}) first reduces the information loss amount and also improves the efficiency by dissociating the variation in counterfactual outcome induced by staggered entering times. Nonetheless, subjects with shorter follow-up times who do not experience a failure event during the observation period are

considered censored by loss to follow-up at the max censoring time.

Once the network produces the counterfactual survival time $H(\hat{\psi}_k, k)$, the artificial censoring indicator and artificial censored time follow as: $X(\hat{\psi}, k) = \min(H(\hat{\psi}, k), C(\hat{\psi}, k))$ and $\Delta(\hat{\psi}, k) = \mathbf{1}\{H(\hat{\psi}, k) < C(\hat{\psi}, k)\}$. As demonstrated in Section 2.7, any function of $\{X(\hat{\psi}, k), \Delta(\hat{\psi}, k)\}$ is independent of the treatment/exposure assigned at time k. Examples of such functions include $g_k(X(\hat{\psi}, k), \Delta(\hat{\psi}, k)) = X(\hat{\psi}, k)$. Hence, the dual output of this network should both be conditionally independent of A_k given exposure/covariates histories $\bar{Z}_k$, following the G-estimation logic. In the previous G-estimation, only the artificial censoring indicator is used in the estimating equation as the representative of the counterfactual outcomes. Nevertheless, the indicator is a step function, which is not preferred when the true parameter space is high as the algorithm often fails to converge to optimal values (Joffe et al., 2011). A suggested remedy is to use a smoothed version of the indicator (Joffe et al., 2011). Specifically,

$$\begin{aligned}\Delta^*(\hat{\psi}, k) &= \Delta(\hat{\psi}, k)\left\{\frac{C(\hat{\psi}, k) - X(\hat{\psi}, k)}{C(\hat{\psi}, k)}\right\} \\ &= \begin{cases} 1 - \dfrac{X(\hat{\psi}, k)}{C(\hat{\psi}, k)}, & \text{if } X(\hat{\psi}, k) < C(\hat{\psi}, k) \\ 0, & \text{if } X(\hat{\psi}, k) = C(\hat{\psi}, k) \end{cases}\end{aligned}$$

During our own experiments, it's discovered that, performances of proposed algorithms are similar using either $X(\hat{\psi}, k)$ or $\Delta^*(\hat{\psi}, k)$. From the formula, these two functions contain the same amount of information about the counterfactual causal parameter. Note that $\Delta^*(\hat{\psi}, k)$ has a range of $[0, 1]$ which seems more friendly to the optimization problem. Nonetheless, it's discovered by us that including both two terms in the loss function in-

creases the chance of convergence without increasing too much computational burden. To improve the performance, the bivariate output $\{X(\hat{\psi}, k), \Delta^*(\hat{\psi}, k)\}$ is used in both proposed algorithms to enforce the conditional independence. In other words, the conditional dependency (as the loss) will be computed between each of $\{X(\hat{\psi}, k), \Delta^*(\hat{\psi}, k)\}$ and another target vector. The final objective function contains the sum of these two pairs of losses.

Like other typical machine learning algorithms, data will first be split into training and testing sets. The network's weights will be updated by SGD during training with respect to the specific loss. The algorithm is monitored and potentially stopped early by validation loss. This can effectively avoid overfitting. Then, the trained network will *estimate* the treatment-free survival outcome with respect to test data. The intermediate outputs of the network -, $\hat{\psi}_k$, identify the potentially time-varying individualized survival causal effect of exposure on expanding/compressing subjects' survival time. Compared to estimating the counterfactual outcomes directly, the $\hat{\psi}_k$ allows a better communication of exposure effects. Thus, the estimation can be used to recommend treatment in the future or to motivate exposure cessation.

The two proposed algorithms have the following notable advantages:

- None of these algorithms requires a known formula for the blip function and the RNN structured blip function still satisfies conditions required for the blip function presented in Section 2.6.

- They incorporate heterogeneity as the individual survival causal effects $\hat{\psi}_{k,i}$ are estimated based on subject-specific treatment/exposure history. In fact, subjects'

histories can be consolidated to the largest extent by using the RNN structure, when exploring the possible blip function formats.

- These algorithms release users from an infeasible grid search on a potentially high-dimensional space. Instead of plugging in different ψ values from a wide range and refitting the exposure model (such as logistic regression for binary exposure or other regressions for non-binary exposure) repeatedly, the network searches for the solution by gradient descent directly.

- These algorithms can handle high-dimensional data.

3.4 Loss Functions for two Proposed Algorithms

Following the concept of nested trial, there are $K+1$ sub-models in a SNAFTM. Subjects who are still alive at each visit k contribute to each sub-model. Under the general setting of RNN, a loss is calculated at each time step (visit), with respect to a specific loss function. The overall loss is calculated by taking the average of these $K+1$ losses. The network is optimized when the overall loss is minimized. RNN lends itself naturally to the nested trial setting. How to estimate the counterfactual outcomes has been explained in the above section, we now discuss the loss function prepared for each proposed algorithm in the following sub-sections.

3.4.1 Loss Function for GE-SCORE - Score of the Partial Likelihood

In the standard G-estimation, the basis is the conditional independence guaranteed by SRA (2.12) under the null hypothesis. In other words, if the exposure model

$f(A_k|\bar{L}_k, \bar{A}_{k-1}; \alpha_0) = f(A_k|\bar{Z}_k; \alpha_0)$ can be correctly specified, the SRA under null hypothesis is guaranteed by (2.21). Here, $\alpha_0 \in \mathbb{R}^p$ is an unknown parameter and, for each $\alpha \in \mathbb{R}^p$, $f(A_k|\bar{Z}_k; \alpha)$ is a density with respect to a measure μ. Moreover, the exposure model will only be fitted for subjects with $Y_i > t_k$ at each k. This condition is omitted from the formula. As in Section 2.7, θ denotes the coefficient of the extra term of counterfactual outcome to be added to the exposure model and ψ is the causal parameter of interest, with dimension ν, potentially $\nu > 1$. We define the likelihood as:

$$\begin{aligned}\mathcal{L}(\alpha, \theta, \psi) &= \prod_{i=1}^{N} \mathcal{L}_i(\alpha, \theta, \psi) \\ &= \prod_{i=1}^{N} \prod_{k=0}^{V_i} \mathcal{L}_{k,i}(\alpha, \theta, \psi) \\ &= \prod_{i=1}^{N} \prod_{k=0}^{V_i} \frac{f(A_{k,i}|\bar{Z}_{k,i}; \alpha) \cdot \exp[\theta Q_{k,i}\{A_{k,i}, H_i(\psi, k)\}]}{\int f(a_k|\bar{Z}_{k,i}; \alpha) \cdot \exp[\theta Q_{k,i}\{A_{k,i}, H_i(\psi, k)\}] d\mu(a_k)}.\end{aligned} \tag{3.15}$$

Here, $\theta \in \mathbb{R}^\nu$ and $Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} = q(A_{k,i}, H_{k,i}(\psi), \bar{Z}_{k,i}) \in \mathbb{R}^\nu$, where $q()$ is a fixed function. For computational simplicity, we set $Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} = [G_{k,i}\{H_i(\psi, k)\}]A_{k,i}$. Additionally, $G_{k,i}\{H_i(\psi, k)\} = g_k(H_i(\psi, k), \bar{Z}_{k,i}) \in \mathbb{R}^\nu$, where $g_k()$ is a fixed function briefly discussed in Section 2.7. For example, if $\nu = 1$, $g_k(H_i(\psi, k), \bar{Z}_{k,i})$ may equal to $H_i(\psi, k)$. Given correct $f(A_k|\bar{Z}_k; \alpha_0)$, SRA and counterfactual outcomes are calculated correctly by (3.13), $\mathcal{L}_{k,i}(\alpha, \theta, \psi^*)$ is a correctly specified model for $f(A_k|\bar{Z}_k, H(\psi^*))$ with true values α_0 and $\theta_0 = 0$ so that $\mathcal{L}_{k,i}(\alpha, \theta, \psi^*)$ is a correctly specified partial likelihood. Let $\tilde{\alpha} = \tilde{\alpha}(\psi)$ maximize $\mathcal{L}(\alpha, 0, \psi)$ and $(\hat{\alpha}, \hat{\theta})$ maximize $\mathcal{L}(\alpha, \theta, \psi^*)$. Define

$$S_{\theta,k,i}(\alpha, \theta, \psi) = \frac{\partial \log \mathcal{L}_{k,i}(\alpha, \theta, \psi)}{\partial \theta}. \tag{3.16}$$

Other score functions can be defined analogously. Solving $S_\theta(\tilde{\alpha}, 0, \psi) = 0$ leads to a consistent estimation of ψ. Taking the partial derivative of $\log \mathcal{L}_{k,i}(\alpha, \theta, \psi)$ with respect to θ directly leads to

$$
\begin{aligned}
&S_{\theta,k,i}(\tilde{\alpha}, 0, \psi) \\
&\quad = \frac{\partial \log \mathcal{L}_{k,i}(\alpha, \theta, \psi)}{\partial \theta} \\
&\quad = \frac{\partial}{\partial \theta} \Bigg\{ \log f(A_{k,i}|\bar{Z}_{k,i}; \alpha) + \theta Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} \\
&\qquad\quad - \log \int f(a_k|\bar{Z}_{k,i}; \alpha) \cdot \exp[\theta Q_{k,i}\{A_{k,i}, H_i(\psi, k)\}] d\mu(a_k) \Bigg\} \\
&\quad = Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} \\
&\qquad\quad - \frac{\int f(A_{k,i}|\bar{Z}_{k,i}; \alpha) \cdot \exp[\theta Q_{k,i}\{A_{k,i}, H_i(\psi, k)\}] \cdot \ Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} d\mu(a_k)}{\int f(A_{k,i}|\bar{Z}_{k,i}; \alpha) \cdot \exp[\theta Q_{k,i}\{A_{k,i}, H_i(\psi, k)\}] d\mu(a_k)} \\
&\quad \overset{\alpha=\tilde{\alpha},\theta=0}{=} Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} - \int f(A_{k,i}|\bar{Z}_{k,i}; \tilde{\alpha}) \cdot Q_{k,i}\{A_{k,i}, H_i(\psi, k)\} d\mu(a_k)
\end{aligned}
\tag{3.17}
$$

When $\nu = 1$ and we set $G_{k,i}\{H_i(\psi, k)\} = H_i(\psi, k)$, the above score function can further be reduced to

$$
\begin{aligned}
S_{\theta,k,i}(\tilde{\alpha}, 0, \psi) &= H_i(\psi, k) \cdot A_{k,i} - \int f(A_{k,i}|\bar{Z}_{k,i}; \tilde{\alpha}) \cdot H_i(\psi, k) \cdot A_{k,i} d\mu(a_k) \\
&= H_i(\psi, k) \cdot [A_{k,i} - \int f(A_{k,i}|\bar{Z}_{k,i}; \tilde{\alpha}) \cdot A_{k,i} d\mu(a_k)] \\
&= H_i(\psi, k) \cdot [A_{k,i} - E(A_{k,i}|\bar{Z}_{k,i})]
\end{aligned}
\tag{3.18}
$$

Last, the loss function for Algorithm I is based on the score function:

$$
S_\theta(\tilde{\alpha}, 0, \psi) = \sum_{i=1}^{N} \sum_{k=0}^{V_i} S_{\theta,k,i}(\tilde{\alpha}, 0, \psi) = 0. \tag{3.19}
$$

This score function has a mean zero. It can be written as $S_{k,i}(\psi)$ to make the relationship between ψ and the score explicit. The $\hat{\psi}$ obtained by solving this estimating equation follows an asymptotically normal distribution with mean = ψ^* and variance that can be consistently estimated from the observed data. Solving the score function for ψ is equivalent to minimizing the empirical linear correlation between the counterfactuals and residuals. We use a correlation coefficient (e.g. Pearson correlation coefficient) between the residual and counterfactual as the loss function. A zero Pearson correlation coefficient means adding the counterfactual cannot provide any new information to the linear model explaining A_k which already contains $\bar{Z}_k$. The algorithm I is hence called G-Estimation based on the Score of the partial likelihood given by an RNN (GE-SCORE).

3.4.2 Loss Function for GE-MIMIC - Difference Between Observed and Empirical Correlation Metric

The unbiasedness of $\hat{\psi}$ in Algorithm I relies on the correct specification of $Pr(A_k|\bar{Z}_k)$, which would not likely be a trivial task, especially when $\nu > 1$. Even though it is advantageous to use rich likelihood models for $Pr(A_k|\bar{Z}_k)$, given histories accumulated with time, the curse of dimensionality becomes inevitable. Using RNN can partially mitigate the high dimensional challenges. Nevertheless, traditional G-estimation and the GE-SCORE use a loss function that depends on the assumption that a model can capture the exposure mechanism sufficiently and that conditional independence is guaranteed by the zero correlation coefficient when the counterfactual is added to the model as an extra term. The conditional independence obtained under this scenario is only linear. It's com-

pletely possible that A_k and $H(k, \psi)$ are nonlinearly dependent (when $\psi_0 \neq \psi^*$) but they seem linearly independent. Under this case, one fails to reject H_0 (3.5) when it's wrong, in other words, commits a type II error. In terms of the causal parameter estimation, the resultant confidence interval will include incorrect ψ values and the estimation efficiency tends to be low. More robust algorithms are thus desirable.

To ensure both linear and nonlinear conditional independence, some advanced machine learning algorithms (see Section 3.1.3) are considered when performing the testing. We first observe that the null hypothesis (conditional independence) is likely to be true when the correlation obtained with observed a_k and $Y^{\bar{a}_{k-1},\bar{0}}$ is close to the same correlation measure obtained under H_0 (3.6). Yet, the distribution of exposure under H_0 is not accessible to researchers. We propose to first generate numbers of data sets that are consistent with H_0, namely, to repeatedly generate $\tilde{a}_k$ conditional on $\bar{z}_k$. Machine learning algorithms such as RNN are powerful tools to fulfill this goal. Then, we define a statistic that captures the dependency between A_k and $Y^{\bar{a}_{k-1},\bar{0}}$. Computing the statistic for each generated copy of $\tilde{a}_k$ and $Y^{\bar{a}_{k-1},\bar{0}}$ will approximate the distribution of the statistics under the null distribution. The true counterfactual should yield a similar statistic with a_k and $\tilde{a}_k$ because the latter is generated under the conditional independence assumption.

Usually, the practitioners have confidence in specifying the null distribution of exposure at each time step, $Pr(A_k|\bar{Z}_k)$, especially when the exposure of interest is low dimensional. If the exposure is discrete as in this dissertation, the target is actually to learn the probability of random exposure variable A taking each possible value. As a result, desired null samples $\tilde{a}_k$ may be simply generated by repeatedly drawing from the possible outcome set with the predicted probability. Deep networks such as RNN

are extremely powerful in predicting the $\mathbb{E}(A_k|\bar{Z}_k)$. In the continuous exposure case, adding noises to the conditional expectation of exposure given histories $\mathbb{E}\ (A_k|\bar{Z}_k)$ (Sen et al., 2018) should approximate the null distribution. The validity of Algorithm II is ensured because under the null hypothesis (3.6), the generated sequence of random triples $(\{\tilde{a}_k\}_m, Y^{\bar{a}_{k-1},\bar{0}}, \bar{Z}_k), m = 1, \cdots, M$ are exchangeable, at each visit k. Consequently, the sequence of summary statistic $\rho_{k,m}$ which is a measurable function of the generated sequences, is also exchangeable. Specifically, the summary statistics ρ projects the generated and observed samples into a real-value scalar, by the mapping $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \to \mathbb{R}$.

A sequence of random variables is said to be exchangeable if its distribution is invariant under permutations. The "representation theorem" for exchangeable sequences of random variables, first stated by de Finetti and extended by Diaconis and Freedman (Diaconis and Freedman, 1980) for finite sequences is useful for Algorithm II. Previous literature has shown that every sequence of conditionally i.i.d. random variables can be considered as a sequence of exchangeable random variables (Heath and Sudderth, 1976). For a specific predicted counterfactual outcome sequence $Y^{\bar{a}_{k-1},\bar{0}}$, by construction, the resulting random sequence of data sets $(\{\tilde{a_k}\}_m, Y^{\bar{a}_{k-1},\bar{0}}, \bar{z}_k), m = 1, \cdots, M$ is exchangeable at each visit k. This is justified by the fact that $\{\tilde{a}_k\}_m$ is generated by sampling from the same Bernoulli or Multinomial distribution, at each m. Therefore, the sequence of statistics $\rho_{k,m}$ is also exchangeable within each k.

Next, we argue that the null samples $\tilde{a}_k$ generated by the regressor part preserve the correlation between a_k and $\bar{z}_k$ but remove all potential correlations between the counterfactual outcome $Y^{\bar{a}_{k-1},\bar{0}}$ and a_k. For the discrete exposure case, a prediction algorithm

with sufficient capacity can estimate the conditional distribution $q_{H_0}(A_k)$ (RHS of the null hypothesis (3.6)) exactly. The null samples generated from the null model preserves the dependence structure $Pr(A_k, \bar{Z}_k)$ but break any dependence between A_k and $Y^{\bar{a}_{k-1},\bar{0}}$. Intuitively, $\tilde{A}_k$ is random variable generated without looking at the counterfactual outcome $Y^{\bar{a}_{k-1},\bar{0}}$ at visit k. Replacing the A_k with $\tilde{A}_k$ will break the causal path from $Y^{\bar{a}_{k-1},\bar{0}}$ to A_k if there is any. Per the law of large numbers, the ρ_k should be close to the mean of $\hat{\rho}_{k,m}$ when M is large (Figure 3.3), under the null hypothesis. The algorithm hence uses the mean of the absolute difference between the mean of $\hat{\rho}_{k,m}$ and observed ρ_k across all $K+1$ visits as the overall loss.

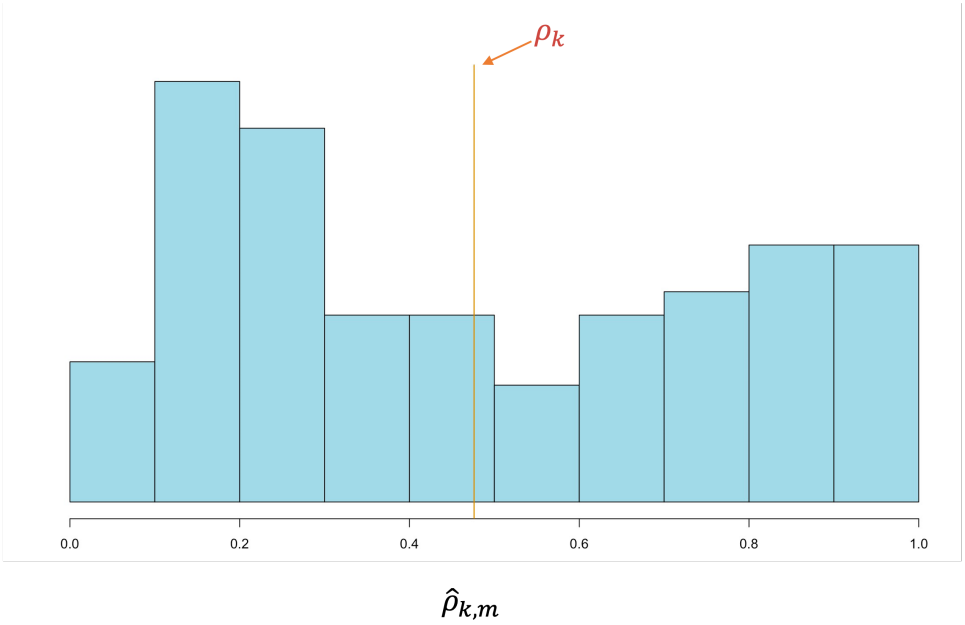


Figure 3.3: A sample empirical distribution of ρ_k and $\hat{\rho}_{k,m}$ under the null hypothesis

Note, the null sample generation scheme in Algorithm II can be viewed as regression-based, although a neural network replaces an ordinary regression in this algorithm. The purpose of the regression is to mimic the null sample distribution while conditioning on a

set of covariates. Algorithm II is hence called GE-MIMIC.

3.5 Procedures of GE-SCORE

Assume there exists a proper law of distribution $Pr(A_k|\bar{Z}_k)$ (potentially characterized by a neural network) which is believed to fit the exposure mechanism sufficiently. Algorithm GE-SCORE proceeds by:

Step 1 Obtain the law of exposure condition on $\bar{Z}_k$ by using a normal regression (e.g., for binary exposure, a logistics regression) or an RNN. Calculate $\hat{a}_{k,i}$ and the residual of estimated exposure for each subject, at each visit as $r_{k,i} = a_{k,i} - \hat{a}_{k,i}$. Here, $r_{k,i}$ follows the general definition of regression residuals.

Step 2 Prepare the data for the counterfactual estimation RNN. This step includes padding all individual sequences $z_{\gamma,i}$ to a length = $K + 1$, adding a covariate for follow-up length based on the visit schedule, etc.

Step 3 Structure the blip function by a many-to-many RNN as described in (3.8). The target output is the estimated sequence of causal survival ratio for each individual on each current time interval: $\{\hat{\psi}_0, \hat{\psi}_1, \cdots, \hat{\psi}_K\}$. Then, the sequence of counterfactual outcomes $Y^{\bar{a}_{k-1},\bar{0}}$ can be calculated by a lambda layer which further incorporates artificial censoring.

Step 4 For each epoch of data, use the loss function $\mathcal{L}_{GE-MIMIC,k} = \rho(r_k, Y^{\bar{a}_{k-1},\bar{0}})$ to update the network weights.

The above steps are repeated until convergence. The algorithm can be viewed as a deep

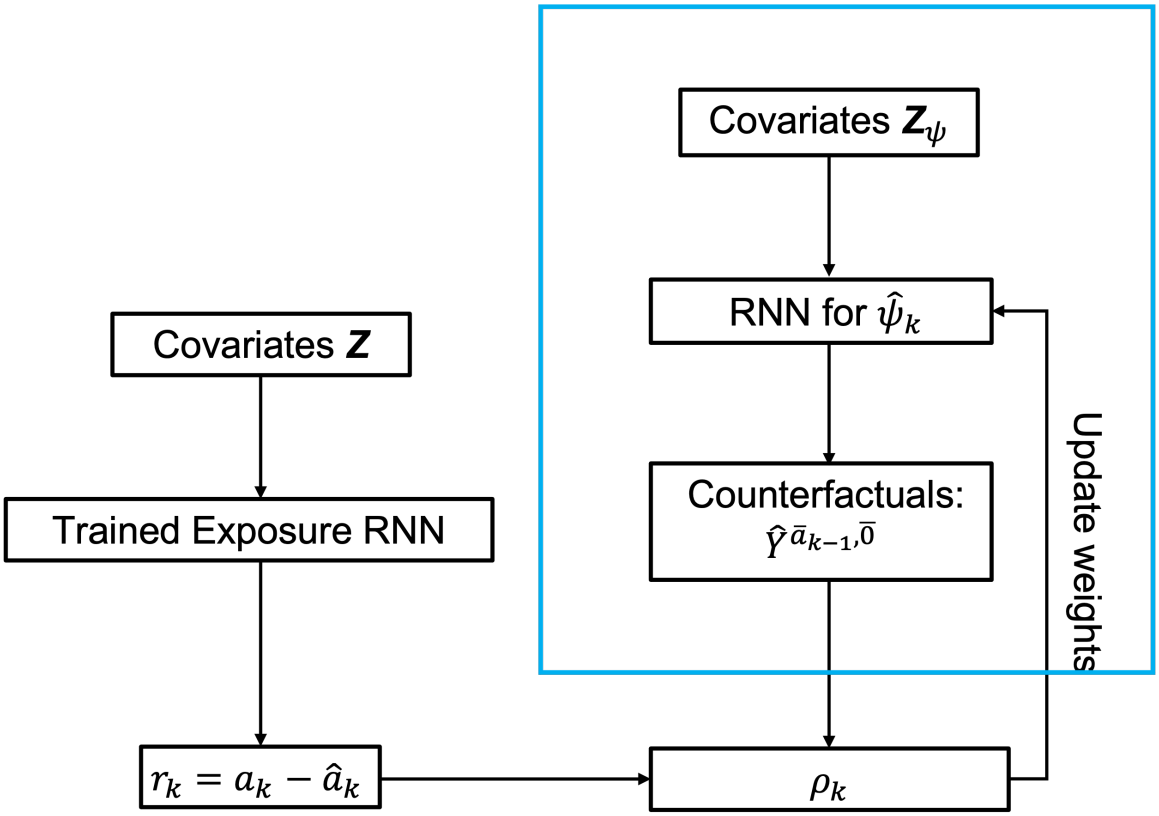


Figure 3.4: Architecture of GE-SCORE

extension of the traditional G-estimation.

3.6 Procedures of GE-MIMIC

To establish the desired exchangeability among the generated sequences of the correlation statistics, GE-MIMIC first approximates the null distribution of exposure - $Pr(A_k|\bar{Z}_k)$ by a neural network. After generating null samples $\tilde{a}_k$ which is consistent with the obtained conditional null distribution, the correlation metric ρ that captures the $\{\tilde{a}_k, Y^{\bar{a}_{k-1},\bar{0}}\}$ dependency is computed at each visit. Note, the $\tilde{a}_k$ and $Y^{\bar{a}_{k-1},\bar{0}}$ are both one-dimensional vectors at each visit k. This process can be repeated M times. At each time we create a new batch of samples with freshly generated null samples $\tilde{a}_k$ but other

information stays the same. When M is sufficiently large, the resultant sequence $\hat{\rho}_{k,m}$ approximates the empirical distribution of correlation coefficients ρ_k between $Y^{\bar{a}_{k-1},\bar{0}}$ and A_k under the null hypothesis. Finally, ρ_k is the correlation coefficient calculated with observed data a_k and the same batch of counterfactual outcomes $Y^{\bar{a}_{k-1},\bar{0}}$ used in $\hat{\rho}_k$ calculation. The GE-MIMIC proceeds as:

Step 1 Train a many-to-many RNN to predict the sequence of exposure conditional expectations $\mathbb{E}(A_k|\bar{Z}_k)$. For a discrete random variable, this network predicts the probability $P(A_k = a_k)$.

Step 2 For each $m, m = 1, \cdots, M$, generate a sequence of noise random variables $\{\varepsilon\}_k$. For example, $\boldsymbol{\varepsilon} \sim MVN(0, \Sigma)$, where Σ controls the magnitude of noises, hence the parameter should be chosen with respect to the data. Calculate $\{\tilde{a}_k\}_m = \mathbb{E}(A_k|\bar{Z}_k) + \{\varepsilon_k\}_m$ for given $\bar{Z}_k$. For binary/multinomial distributed exposure, directly draw realized values of the discrete variable based on $\mathbb{E}(A_k|\bar{Z}_k)$ with no noise added. Repeat the process for all M batches of samples.

Step 3 Prepare the data for the counterfactual RNN. This step includes padding all individual sequences $Z_{\gamma,i}$ to a length = $K + 1$, adding a covariate for follow-up length based on the visit schedule, etc.

Step 4 Structure the blip function by a many-to-many RNN as described in (3.8). The target output is the estimated sequence of causal survival ratio for each individual on each current time interval: $\{\hat{\psi}_0, \hat{\psi}_1, \cdots, \hat{\psi}_K\}$. Then, the sequence of counterfactual outcomes $Y^{\bar{a}_{k-1},\bar{0}}$ can be calculated by a lambda layer which further incorporates artificial censoring.

Step 5 Calculate $\hat{\rho}_{k,m}(\{\tilde{a}_k\}_m, Y^{\bar{a}_{k-1},\bar{0}})$ for each batch of null samples, $m = 1, \cdots, M$. At each visit k, calculate the observed correlation coefficient $\rho_k(a_k, Y^{\bar{a}_{k-1},\bar{0}})$. The loss function $\mathcal{L}_{GE-MIMIC,k} = |\hat{\rho}_{k,m} - \rho_k|$ as defined in Section 3.4.2.

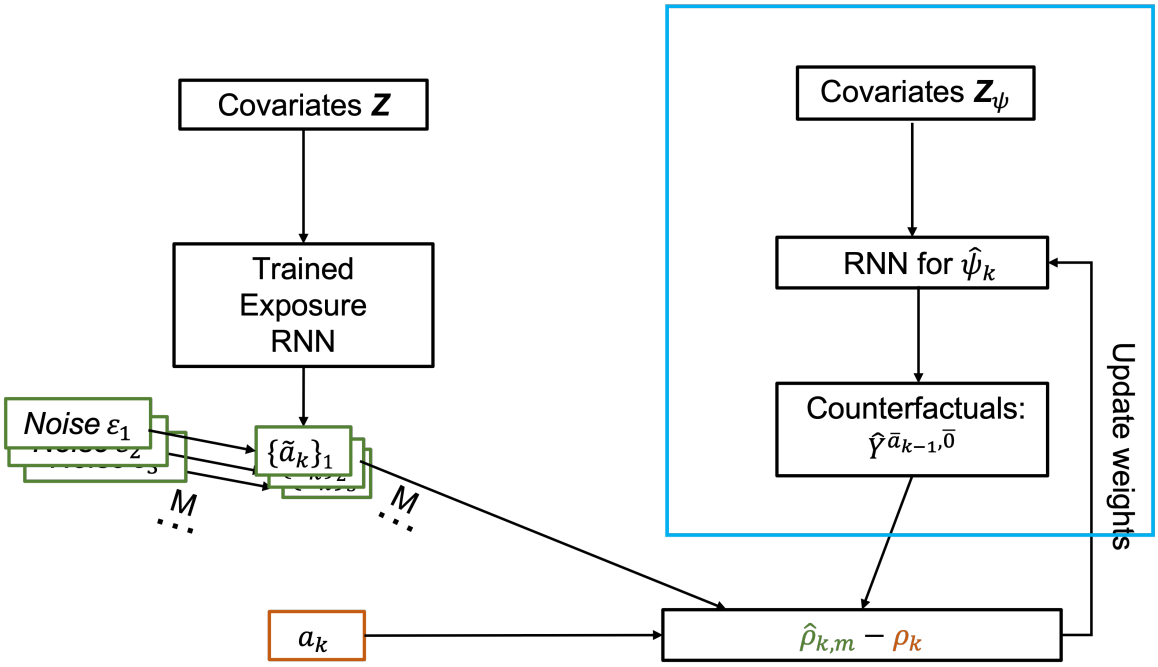


Figure 3.5: Architecture of GE-MIMIC

The estimation efficiency of $\hat{\psi}_k$ will depend on the accuracy of the conditional expectation $\mathbb{E}(A_k|\bar{Z}_k)$. The random variable $\tilde{A}_k$ should be centered approximately at $\mathbb{E}(A_k|\bar{Z}_k)$ if the true conditional distribution is unimodal and noises are chosen appropriately. Essentially, the required assumption is that the distribution of $\tilde{A}_k$ has a large overlap with the true null distribution $q_{H_0}(A_k|\bar{Z}_k)$ (Sen et al., 2018). However, there is no need to add noise to the scope of this dissertation as the generation of null samples based on the Bernoulli distribution has introduced randomnesses to the null samples. Note, above Steps 3 and 4 are the same as Steps 2 and 3 of GE-SCORE. This counterfactual RNN structure will be the same for both proposed algorithms, as shown in the blue boxes in Figures 3.4 and 3.5.

Multiple metrics can be used to assess the conditional dependence required by the GE-MIMIC loss functions. Ideally, the metrics employed should be comprehensive to consider linear as well as non-linear correlation. The Pearson Correlation coefficient measures the linear relationship between two variables. It's defined as:

$$\hat{\rho}_{PC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})}\sqrt{\sum_i (y_i - \bar{y})}}. \tag{3.20}$$

More generally, the distance correlation (DC) (Székely et al., 2007) measures both linear and nonlinear association between two random variables or random vectors. Similar to the Pearson correlation coefficient, the population distance correlation coefficient is zero if and only if the random vectors are independent. Unlike the Pearson correlation, the distance correlation coefficient (Formula 3.21) can only take values in the unit interval $[0, 1]$. With covariance and variance being expanded to distance covariance ($dCov$) and distance variance ($dVar$), the distance correlation coefficient is defined as:

$$\hat{\rho}_{DC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{dCov(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{dVar(\boldsymbol{x}) dVar(\boldsymbol{y})}}. \tag{3.21}$$

Though there are many other correlation/distance metrics, they are not suitable as a neural network loss function. Choice of correlation metrics also depends on their differentiability - the optimization is usually along the direction of the steepest loss decrease which heavily relies on the derivatives. Some popular distance measures such as Kolmogorov-Smirnov (K-S) statistics cannot be implemented in the proposed algorithm as there is no way to minimize such loss with an available optimizer. The correlation

metrics presented here are preserved for one-dimensional random variables in this dissertation. In other words, $(\boldsymbol{x}, \boldsymbol{y})$ are both one-dimensional.

3.7 Competing Risk Censoring

The artificial censoring method deals with administrative censoring in survival analysis. For competing risk censored observations we use IPCW to adjust for the bias as discussed in Section 2.9. ICA (Robins, 1992; Robins et al., 1994) here states that $C_k \perp\!\!\!\perp Y^{\bar{a}_k,\bar{0}} | \bar{L}_{k-1}, \bar{A}_{k-1}, C_{k-1} = 1$. In other words, at each visit k, the observed history $(\bar{L}_{k-1}, \bar{A}_{k-1})$ is sufficient to predict whether subject stay under follow-up until visit k. In other words, the competing risk censoring carries no residual information about survival under any of the considered exposure regimes, conditional on the covariate history up to visit k. The predicted conditional probability $\pi_{k,i}$ of under follow-up at visit k can be fitted by standard regression methods (such as a logistic model):

$$\pi_{k,i} = Pr(C_{k,i} = 1 | \bar{Z}_{fu,k-1,i}, C_{k-1,i} = 1) \tag{3.22}$$

Here, Z_{fu} is the set of random variables that makes the ICA assumption true at each visit k. It usually includes some baseline characteristics, treatment history $\bar{A}_{k-1}$ and time-varying confounders $\bar{L}_{k-1}$. To correct the potential bias introduced by informative censoring in G-estimation, the original samples are weighted by the inverse of the cumulative product of the predicted probability of staying under follow-up $W_{k,i} = 1/\prod_{s=k}^{V_i} \pi_{s,i}$ until subjects' last visit. Subjects who are competing risk censored receive weights of 0. This weighted sampling process asymptotically equals to the weight of the individual

estimating equation term (3.18):

$$S(\psi) = \sum_{i=1}^{N} \sum_{k=0}^{V_i} C_{T,i} S_{k,i}(\psi) W_{k,i}. \tag{3.23}$$

Though the model is now fitted with non-competing risk censored subjects (those with had an observed survival time $T \leq C$ or who were still under follow-up at the end of the study) only, the estimator for ψ is still unbiased (Mertens and Vansteelandt, 2017). The unbiasedness can be proved as below:

$$\begin{aligned}
E\left(\frac{C_{T,i}}{\prod_{s=k}^{V_i} \pi_{s,i}} S_{k,i}(\psi)\right) &= E\left(\prod_{s=k}^{V_i} \frac{I(C_{s,i}=1)}{\pi_{s,i}} S_{k,i}(\psi)\right) \\
&= E\left[E\Big\{\prod_{s=k}^{V_i} \frac{I(C_{s,i}=1)}{\pi_{s,i}} S_{k,i}(\psi)\Big| \bar{Z}_{fu,V_i,i}, C_{V_i-1,i}=1\Big\}\right] \\
&= E\left[\frac{E(C_{V_i,i}=1|\bar{Z}_{fu,V_i-1,i}, C_{V_i-1,i}=1)}{\pi_{s,i}}\right. \\
&\qquad \left. \times E\Big\{\prod_{s=k}^{V_i-1} \frac{I(C_{s,i}=1)}{\pi_{s,i}} S_{k,i}(\psi)\Big| \bar{Z}_{fu,V_i,i}, C_{V_i-1,i}=1\Big\}\right] \\
&= E\left(\prod_{s=k}^{V_i-1} \frac{I(C_{s,i}=1)}{\pi_{s,i}} S_{k,i}(\psi)\right) \\
&= \cdots \\
&= E\{I(C_{k,i}=1) S_{k,i}(\psi)\} \\
&= 0
\end{aligned} \tag{3.24}$$

To make the weights less unstable, the numerator 1 in $W_{k,i}$ can be replaced by the predicted probability as Formula (3.22) but now excluding time-varying confounders from Z_{fu} (Mertens and Vansteelandt, 2017). The unbiasedness of the estimator can be proved

analogously.

There are various ways to estimate the conditional probabilities need by the weights. Popular choices include logistic regression. In this dissertation, we propose to use the state-of-the-art machine learning algorithm (RNN) in predicting sequential data to predict the under follow-up sequence.

3.8 Simulation

To assess the performance of the proposed algorithms, this section includes the simulation schema and results. Different scenarios are basically designed with respect to different blip functions, which are supposed to reflect the biological natures of distinct disease processes. Algorithms described in Section 3.5 and 3.6 are compared with the original G-estimation method with grid search, as well as other methods (e.g., MSM) dealing with time-varying confounding.

The simulation scheme is close to (Young et al., 2009) and it's well-accepted in previous literature that treats time-varying confounding. One first needs to know the visit schedule and the study length to set up the data. An example visit schedule is three visits $k = (0, 1, 2)$, scheduled at year $t_k = (0, 2, 4)$. After the last visit at year 4, subjects are followed for an additional two years, hence the study is scheduled to end at year 6, the maximal follow-up time $t_K = 6$. Different visit schedules with longer/shorter study duration or unequally spaced visits are easy to incorporate. For each individual i in a sample with size N, the following steps are implemented to generate data.

- Step 1:

- Simulate baseline static characteristics if any. For instance, Gender is a binary variable with 1 denoting female such that Gender $\sim Bernoulli(0.5)$.
- Simulate baseline survival time $Y^{\bar{0}} \sim Weibull(a, b)$, where the p.d.f of Weibull is defined as $f(x; a, b) = (\frac{a}{b})(\frac{x}{b})^{a-1} \exp(-(\frac{x}{b})^a)$, $a = 1, b = 5$.
- Define $L_{-1} = A_{-1} = 0$. Then for each visit k,
  - when $k = 0$, simulate $L_0 \sim N(0, 0.5)$. Simulate A_0 by $\text{logit}[Pr(A_0 = 1|L_0; \boldsymbol{\beta})] = \beta_0 + \beta_1 L_0 + \epsilon$ if the exposure model is Logistic, where $\epsilon \sim N(0, 0.01)$, $\boldsymbol{\beta} = (-0.4, 0.5)$. For the non-Logistic exposure model, replace the logit model with random sampling from the nonlinear function set.
  - when $k > 0$, simulate $L_k \sim N(\alpha_1 A_{k-1} + \alpha_2 L_{k-1} + \alpha_3\{Y^{\bar{0}} < 20\}, \epsilon)$, where $\{\alpha_1, \alpha_2, \alpha_3\} = \{1, 0.5, 1\}$; simulate A_k by $\text{logit}[Pr(A_k = 1|L_k, A_{k-1}; \boldsymbol{\beta})] = \beta_0 + \beta_1 L_k + \beta_2 A_{k-1} + \epsilon$, where $\boldsymbol{\beta} = (-0.4, 0.5, 1)$. For the non-Logistic exposure model, replace the logit model with a random sample from the nonlinear function set defined later.

- Step 2: Define the blip function and corresponding true ψ values. Use $Y^{\bar{a}_{k-1},\bar{0}} = t_k + \int_0^Y \exp(\psi A_k)$ as an example; here the true ψ is a constant, unrelated to other covariates.

- Step 3: Simulate the event indicator E_k and the observed survival time T based on the blip function defined in Step 2, $Y^{\bar{0}}$ and exposure/covariates simulated in Step 1. With the example blip function, at each $k, k = (0, \cdots, K)$:

  if $Y^{\bar{0}} > \int_0^{t_{k+1}} \exp(\psi A(u))du$ then $E_k = 0, T = t_{k+1}$;

else if $Y^{\bar{0}} \leq \int_0^{t_{k+1}} \exp(\psi A(u))du$ then $E_k = 1, T = t_k + [Y^{\bar{0}} - \int_0^{t_k} \exp(\psi A(u))du] \cdot \exp(-\psi A_k)$.

- Step 4: Redefine $L_k = 0, A_k = 0$ for all k when $t_k > T$. There are $V_i + 1$ valid records for each subject, given the data is arranged in a long format.

The nonlinear exposure model set we considered include: $\text{logit}[Pr(A_k = 1)] = \beta_0 + \log(|\beta_1 L_k + \beta_2 A_{k-1}|) + N(0, 0.01), \text{logit}[Pr(A_k = 1)] = 10 * \sin(\beta_0 + \beta_1 L_k + \beta_2 A_{k-1} + N(0, 0.01)), Pr(A_k = 1) = |\sin(\beta_0 + \beta_1 L_k + \beta_2 A_{k-1} + N(0, 0.01))|$, with $\boldsymbol{\beta} = (-0.4, 0.5, 1)$ as previously defined.

All models and experiments were implemented in R (R Core Team, 2021) with `keras` (Chollet et al., 2017). This package provides an interface to Keras from R. Keras (Chollet et al., 2015) is a high-level neural networks API of either TensorFlow or Theano. The `kerasR` interfaces with Python 3.8.8 and Tensorflow 2.9.1 (Abadi et al., 2015).

3.9 Hyperparameter Tuning

To name a few hyperparameters that may influence the algorithm performance, one can consider the number of neurons, number of layers, algorithm optimizer, learning rate, etc. The hyperparameters need to be tuned case by case, depending on the data format/dimension, problem/algorithm complexity, etc. We present some considerations with the simulation scenarios as examples.

There are two networks in both proposed algorithms, the exposure prediction network and the counterfactual prediction network. First, GE-SCORE and GE-MIMIC both rely on a faithful model of $Pr(A_k|\bar{Z}_k)$. This is a standard supervised learning problem.

A shallow network is sufficient for our sequence-to-sequence prediction problem given the fact that the targeted output (the exposure sequence $\{A_k\}$) is low dimensional. Moreover, we are not targeting predicting the exposure mechanism with 100% confidence in this problem. A perfect prediction provides no residuals/randomness for the later counterfactual network to learn. If the perfect prediction is observed, the positivity assumption is violated, and a closer examination of the data and reconsideration of the method is required. In conclusion, we choose a one-layer simple RNN with 20 units for the exposure prediction network. Smaller or larger numbers of units have been explored but no significant difference observed. More advanced recurrent units such as LSTM can easily replace the simple RNN unit if the simple RNN is believed to not be sufficient when a much more complicated exposure mechanism is expected. The `sigmoid` activation function is chosen for the output layer as we are predicting the probability of exposure.

The counterfactual outcome network is capable of taking a high dimensional input as it's free of grid search ideas. However, the network complexity relies on domain knowledge. For instance, a doctor may advise on whether there is an effect modifier for the exposure causal effect on the outcome, or whether there are multiple higher order terms, etc. Usually, a comparable simple RNN is adequate if traditional survival analysis methods have provided some promising results in previous research. We use a one-layer RNN or LSTM with one unit for all scenarios presented later. We noted that the network pruning issue is extremely important in the counterfactual outcome network as the causal survival ratio ψ may only relate to a small subset of covariates. In our simulation settings, the input dimension is low. Hence the inclusion of many neurons in the hidden layer does no good. Usually, one can employ dropout layers which randomly set input

units to 0 with a prespecified rate at each step during training. L-1 and L-2 regularizers are examples of other popular regularization methods. These methods apply penalties on layer parameters or layer activity during optimization by summing the weight matrices of the nodes into the loss function that the network optimizes. However, these methods are not particularly useful for our counterfactual outcome network due to the complexity of our loss function. One notable useful penalty term is the variance of $\hat{\psi}_k$ at each k as it directly penalizes the complexity of prediction. By adding this penalty term to the loss function, predicted causal parameters are encouraged to be less variant and thus well regularized. When the penalty term weights large in the total loss, the algorithm is pushed to generate similar even constant output for $\hat{\psi}_k$. This reduces to G-estimation in one dimension. Thus, the choice of weights of the penalty is a delicate question. We start by considering the scale of the original loss function. In GE-SCORE, the correlation coefficient is calculated as the first step. As the objective function is to be minimized, we add an absolute value sign to the coefficient. The range of the loss function is thus [0, 1]. We have considered some transformation functions (e.g., Fisher z transformation) to expand the loss value range. However, Fisher z transformation has an ignorable influence on values close to 0, but this optimization question needs large gradients especially when the coefficient is already close to 0. Other functions that focus on transferring values close to 0 (e.g., $f(x) = \log((2-x)/x)$) are unstable to use, so the optimization process is likely to be very noisy. In GE-MIMIC, the scale of ρ is small (usually between -1 and 1) and the value may not be close to 0, or at least the objective is not to minimize each ρ to 0, hence we first perform the Fisher transformation to expand the ρ range. Nonetheless, the absolute difference between ρ and $\hat{\rho}$ is still expected to be small. In the end, we manually

magnify loss in both algorithms by 100 times to increase the gradients returned by the loss function. As a result, the theoretical range of GE-SCORE loss is now [0, 100] and the theoretical range of GE-MIMIC loss is broader. After some experiments, we found adding the before-mentioned penalty term of $\hat{\psi}_k$ with a factor of 10 to the loss function works well. The loss function for GE-SCORE at visit k is now:

$$\mathcal{L}'_{GE-SCORE,k} = \gamma_1 \cdot \mathcal{L}_{GE-SCORE,k} + \gamma_2 \cdot var(\hat{\psi}_k). \tag{3.25}$$

Similarly, the loss function for GE-MIMIC at visit k is now:

$$\mathcal{L}'_{GE-MIMIC,k} = \gamma_1 \cdot \mathcal{L}_{GE-MIMIC,k} + \gamma_2 \cdot var(\hat{\psi}_k). \tag{3.26}$$

In both loss functions, γ_1 and γ_2 control the contribution of each term to the total loss. Per our previous discussion, $\gamma_1 = 100$ and $\gamma_2 = 10$ in all following simulations and in the applications in Section 4.4.4.

The correlation measurement used in loss functions is calculated per batch of samples, hence the batch size plays an important role in the algorithm optimization. A batch size that is much smaller than the sample size (e.g., 64) is usually employed in other ML algorithms Processing the data in small batches requires less memory, also, networks train faster with mini-batches because the weights are updated after each propagation. Nonetheless, our proposed algorithms fail to succeed with small batch sizes. The smaller the batch size, the less accurate using the batch sample correlation to estimate the whole sample correlation. Then, the less accurate the batch sample gradient. Both proposed

algorithms require a sufficiently large batch size to ensure the batch of samples well represents the correlation coefficient in the whole sample. It's found that a batch size as small as 25% of the sample size works well. In GE-SCORE, we process 25% of the training sample during each propagation. In GE-MIMIC, the original sample has been expanded significantly (i.e. when $M = 50$, the sample is expanded 50 times with different null samples), we use the original training sample size as the batch size. Thus, it takes M batches to process all training samples.

Initialization of weights is another important hyperparameter that can be tuned. The stochastic nature of machine learning algorithms means that any single run generates different predictions, given a different source of randomness used by the algorithm during optimization. Proper initialization of the weights in a neural network is critical for its convergence (Glorot and Bengio, 2010). Besides fixing the initialization schema to Glorot normal initializer (Glorot and Bengio, 2010), we propose to run each algorithm multiple times (e.g., 20 times). Initializing the network with different weights allows the algorithm to start from different places hence providing a larger likelihood of locating the global optima. This is called Multi-start (Martí et al., 2013). In our algorithm, each run the weights will be initialized with random values coming from a truncated normal distribution which centered on 0 with `stddev = sqrt(2(fanin + fanout))`. Here, `fanin` is the number of input units in the weight tensor and `fanout` is the number of output units in the weight tensor. The final prediction is the average of predictions from the three best runs, evaluated by validation loss. Other initialization techniques are explored but the performance is much more sensitive to whether using Multi-start or not, not the initializer itself. The possible reason is that the optimization problem we are facing is non-convex

with a considerable number of local minima. Reporting the mean of multiple runs to improve the robustness as the result no longer relies on a single optimization attempt. To incorporate a wide range of possible counterfactual causal ratio, we use `4*sigmoid` as the activation function for the counterfactual RNN output layer which allows a range of $[0, 4]$ for $\exp(\hat{\psi})$.

Finally, NAdam optimizer (Dozat, 2016) with a learning rate of 0.005 is used for all models. For a real problem that suffers from the curse of dimensionality, tuning becomes more critical. See Chapter 4 for parameter tuning with cross-validation in a real application.

3.10 Scenarios

As the parameter ψ quantifies the causal survival time ratio and is the parameter of interest, simulation scenarios are designed to include a broad range of blip functions. Blip functions defined in simulation Step 2 above change across scenarios and Step 3 needs to be adjusted accordingly based on specific blip formats, while other steps mostly stay the same. Additionally, the exposure mechanism is allowed to be nonlinear to test the robustness of traditional G-estimation and proposed algorithms under potential misspecifications of the exposure model. Given the fact that all deep learning methods need an adequate number of samples to thrive, N is taken as 3000 in all simulations. We didn't explore the capability of algorithms with a small sample size.

The benchmarked algorithm is G-estimation, either with one-dimensional ψ which treats the causal effect universally for all subjects, all time points; or a G-estimation

searching on two-dimensional $\{\psi_1, \psi_2\}$ space. The search range is set to [-1.5,1.5] with a grid size of 0.02. For two-dimensional G-estimation, the search range is decreased to [-1, 1] with a step size of 0.05, for each dimension. Note, it's very important to include the correct set of covariates in high dimensional G-estimation, i.e., the grid search needs to be performed for the covariate which truly has the causal effect. We explored both correct and incorrect G-estimation formula specifications in the simulation. The G-estimation procedure also requires a model for the exposure, logistic regression is chosen to perform this prediction task with $f(A_k) = \alpha_k + \alpha_1 L_k + \alpha_2 A_{k-1}$.

To be more comprehensive, besides the G-estimation method, we also consider the marginal structure model. Although MSM is able to produce the same set of the casual parameters as G-estimation, for instance, a marginal structural AFT model is able to estimate the causal survival ratio as a G-estimation. In general, MSM relies on picking an analysis model correctly to make an unbiased causal inference. Nevertheless, the data generation mechanisms and appropriate analysis models are unknown to practitioners most of the time. In the following simulation, to make the MSM results on par with others, the marginal structural Weibull model is employed to estimate the causal survival ratio. The stabilized weights used in MSM are estimated by R package `ipw` (van der Wal and Geskus, 2011).

The hyperparameters we employed in the proposed algorithms have been discussed in Section 3.9. Finally, it's completely possible that the causal effect of research interest only relates to a set of covariates instead of the whole available covariates set, based on domain knowledge. As a result, those covariates of suspicion were input into the counterfactual network. In the following simulation, as the dimension of covariate is

small (Gender, L are two simulated covariates) and we would like to show the robustness of proposed algorithms, all simulated covariates plus the lagged exposure (i.e., Gender, L_k, A_{k-1}) are feed into the counterfactual network, at each time point.

The censoring rate is another factor that influences the model performance. The right-censoring proportion is controlled by the underlying survival distribution, true causal survival ratio, follow-up length, etc. In practice, the occurrence of the event of interest in an observational study directly relates to the disease prevalence. Multiple scenarios and magnitudes of causal coefficients we considered in the simulation have reflected a wide range of right censoring (e.g., a high censoring scenario with 80-85% right censored, a low censoring scenario with 20-25% right censored).

Next, we provide all simulation results of scenarios considered in order. For each scenario, we explored the positive and negative parameterization of $\boldsymbol{\psi}$. However, whenever there is an effect modificator such as L, the sign of $\hat{\psi}_k$ may not be consistent for all subjects, at each time point because now $\hat{\psi}_k$ is a function of L_k. Nevertheless, a positive $\hat{\psi}_k$ stands for a harmful treatment effect as before. Note the Weibull MSM has the effect estimated in an opposite direction (i.e., positive estimation means beneficial effect), the estimation signs from different methods have been unified before compared to the true value. Absolute bias is the main evaluation metric. Each scenario with certain parameterization is repeated 100 times. Besides the mean, we further report the 25%, 75% quantile, standard error of the absolute bias to present the bias distribution comprehensively.

3.10.1 Scenario I

Scenario I is the simplest case where the causal effect is a constant across all subjects, all time points. In other words, $\hat{\psi}_k = \psi_1 = -0.5$ or 0.5 throughout.

Parameter Exposure Mechanism Percent Censored	Estimation Method	Mean Absolute Bias	25% Absolute Bias	75% Absolute Bias	Bias s.d.
$\psi_1 = -0.5$ Logistic 40.6% Censored	G-est 1-d	0.049	0.020	0.060	0.033
	G-est 2-d (L)	0.111	0.059	0.152	0.073
	G-est 2-d (Gender)	0.020	0.000	0.027	0.032
	Weibull MSM	0.053	0.024	0.078	0.039
	GE-SCORE (simple RNN)	0.078	0.037	0.103	0.055
	GE-SCORE (LSTM)	0.133	0.062	0.174	0.080
	GE-MIMIC (simple RNN)	0.071	0.037	0.102	0.042
	GE-MIMIC (LSTM)	0.129	0.110	0.145	0.029
$\psi_1 = -0.5$ Non-Logistic 38.7% Censored	G-est 1-d	0.052	0.020	0.080	0.048
	G-est 2-d (L)	0.196	0.067	0.232	0.187
	G-est 2-d (Gender)	0.029	0.000	0.050	0.039
	Weibull MSM	0.062	0.024	0.084	0.048
	GE-SCORE (simple RNN)	0.076	0.038	0.104	0.051
	GE-SCORE (LSTM)	0.091	0.048	0.119	0.065
	GE-MIMIC (simple RNN)	0.087	0.037	0.117	0.067
	GE-MIMIC (LSTM)	0.075	0.042	0.102	0.049
$\psi_1 = 0.5$ Logistic 19.4% Censored	G-est 1-d	0.057	0.020	0.080	0.046
	G-est 2-d (L)	0.105	0.050	0.137	0.078
	G-est 2-d (Gender)	0.080	0.050	0.101	0.045
	Weibull MSM	0.040	0.013	0.055	0.031
	GE-SCORE (simple RNN)	0.058	0.034	0.070	0.036
	GE-SCORE (LSTM)	0.059	0.032	0.069	0.044
	GE-MIMIC (simple RNN)	0.052	0.019	0.072	0.039
	GE-MIMIC (LSTM)	0.096	0.077	0.110	0.027

$\psi_1 = 0.5$ Non-Logistic 21.3% Censored	G-est 1-d	0.066	0.020	0.100	0.060
	G-est 2-d (L)	0.194	0.069	0.232	0.198
	G-est 2-d (Gender)	0.093	0.050	0.125	0.064
	Weibull MSM	0.045	0.012	0.063	0.039
	GE-SCORE (simple RNN)	0.068	0.037	0.088	0.041
	GE-SCORE (LSTM)	0.070	0.027	0.093	0.052
	GE-MIMIC (simple RNN)	0.064	0.034	0.087	0.041
	GE-MIMIC (LSTM)	0.067	0.028	0.094	0.046

Table 3.1: Simulation Results for Scenario I.

3.10.2 Scenario II

Scenario II stands for a slightly more complicated case where the exposure causal effect at each time point k is modified by another covariate, such as L_k. Thus, an interaction effect between L_k and A_k is considered. The true causal effect is: $\hat{\psi}_k = \psi_1 + \psi_2 L_k$ for $k = (0, 1, 2)$.

Parameter Exposure Mechanism Percent Censored	Estimation Method	Mean Absolute Bias	25% Absolute Bias	75% Absolute Bias	Bias s.d.
$\psi_1 = 0$, $\psi_2 = -0.3$ Logistic 39.5% Censored	G-est 1-d	0.373	0.352	0.391	0.028
	G-est 2-d (L)	0.084	0.050	0.104	0.058
	G-est 2-d (Gender)	0.302	0.287	0.312	0.021
	Weibull MSM	0.287	0.281	0.293	0.008
	GE-SCORE (Simple RNN)	0.230	0.197	0.263	0.041
	GE-SCORE (LSTM)	0.217	0.181	0.252	0.045
	GE-MIMIC (Simple RNN)	0.203	0.175	0.227	0.040
	GE-MIMIC (LSTM)	0.236	0.204	0.261	0.038

$\psi_1 = 0$, $\psi_2 = -0.3$ Non-Logistic 37.7% Censored	G-est 1-d	0.327	0.285	0.354	0.053
	G-est 2-d (L)	0.224	0.049	0.352	0.288
	G-est 2-d (Gender)	0.285	0.265	0.302	0.035
	Weibull MSM	0.264	0.239	0.281	0.024
	GE-SCORE (Simple RNN)	0.233	0.181	0.280	0.058
	GE-SCORE (LSTM)	0.219	0.182	0.256	0.047
	GE-MIMIC (Simple RNN)	0.224	0.173	0.274	0.057
	GE-MIMIC (LSTM)	0.216	0.162	0.263	0.053
$\psi_1 = 0$, $\psi_2 = 0.3$ Logistic 19.1% Censored	G-est 1-d	0.295	0.282	0.305	0.017
	G-est 2-d (L)	0.138	0.088	0.180	0.074
	G-est 2-d (Gender)	0.290	0.277	0.302	0.017
	Weibull MSM	0.265	0.260	0.269	0.008
	GE-SCORE (Simple RNN)	0.223	0.197	0.241	0.033
	GE-SCORE (LSTM)	0.218	0.195	0.241	0.034
	GE-MIMIC (Simple RNN)	0.204	0.176	0.231	0.034
	GE-MIMIC (LSTM)	0.198	0.175	0.223	0.031
$\psi_1 = 0$, $\psi_2 = 0.3$ Non-Logistic 21.2% Censored	G-est 1-d	0.288	0.272	0.303	0.026
	G-est 2-d (L)	0.206	0.100	0.293	0.149
	G-est 2-d (Gender)	0.277	0.262	0.292	0.021
	Weibull MSM	0.249	0.237	0.260	0.016
	GE-SCORE (Simple RNN)	0.222	0.186	0.258	0.044
	GE-SCORE (LSTM)	0.226	0.183	0.260	0.048
	GE-MIMIC (Simple RNN)	0.201	0.171	0.241	0.046
	GE-MIMIC (LSTM)	0.201	0.168	0.246	0.047

Table 3.2: Simulation Results for Scenario II.

3.10.3 Scenario III

In scenario III, the current exposure causal effect further relates to exposure and covariates histories. The corresponding true causal effect at visit $k = 0$ is $\hat{\psi}_0 = \psi_1 + \psi_2 Gender + \psi_3 L_1 + \psi_4 Gender \times L_1$. This reflects that the causal exposure effect is modified by the baseline characteristic gender, L_1, as well as an interaction between these two. For $k > 0$, $\hat{\psi}_k = \psi_1 + \psi_2 Gender + \psi_3 \sum_{t=1}^{k} (\frac{1}{t} \bar{L}_{k-t} \bar{A}_{k-t}) + \psi_3 L_k + \psi_4 Gender \times L_k$. Besides all terms includes at time point 0, there is an extra term $L_k A_k$ summed over

previous histories. This term accumulates with time if one has been continuously exposed. On the other hand, this term reduces to zero if one was not exposed throughout. The $\frac{1}{t}$ factor controls the magnitude of histories. Conversely, previous exposure (modified by covariates) further back in time has a smaller, if any, influence on the current exposure causal effect. This scenario becomes plausible when the current exposure effect relates to exposure history, as well as an effect modificator in the present. For instance, it's convincing that the current smoking causal effect differs between a smoker who has been smoking for the past 10 years and another smoker who just started last month.

Parameter Exposure Mechanism Percent Censored	Estimation Method	Mean Absolute Bias	25% Absolute Bias	75% Absolute Bias	Bias s.d.
$\psi_1 = 0$, $\psi_2 = -0.2$, $\psi_3 = -0.2$, $\psi_4 = -0.1$ Logistic 40.7% Censored	G-est 1-d	0.365	0.344	0.383	0.029
	G-est 2-d (L)	0.176	0.150	0.170	0.055
	G-est 2-d (Gender)	0.285	0.272	0.296	0.019
	Weibull MSM	0.298	0.291	0.306	0.011
	GE-SCORE (simple RNN)	0.263	0.235	0.296	0.038
	GE-SCORE (LSTM)	0.252	0.224	0.284	0.038
	GE-MIMIC (simple RNN)	0.243	0.214	0.262	0.035
	GE-MIMIC (LSTM)	0.237	0.211	0.255	0.033
$\psi_1 = 0$, $\psi_2 = -0.2$, $\psi_3 = -0.2$, $\psi_4 = -0.1$ Non-Logistic 38.7% Censored	G-est 1-d	0.345	0.309	0.374	0.049
	G-est 2-d (L)	0.305	0.147	0.398	0.267
	G-est 2-d (Gender)	0.269	0.238	0.292	0.041
	Weibull MSM	0.276	0.255	0.295	0.025
	GE-SCORE (simple RNN)	0.258	0.220	0.301	0.048
	GE-SCORE (LSTM)	0.250	0.211	0.296	0.051
	GE-MIMIC (simple RNN)	0.251	0.218	0.287	0.041
	GE-MIMIC (LSTM)	0.243	0.208	0.280	0.044

$\psi_1 = 0$, $\psi_2 = 0.2$, $\psi_3 = 0.2$, $\psi_4 = 0.1$ Logistic 18.1% Censored	G-est 1-d	0.286	0.272	0.298	0.017
	G-est 2-d (L)	0.182	0.152	0.201	0.039
	G-est 2-d (Gender)	0.261	0.248	0.272	0.018
	Weibull MSM	0.262	0.254	0.269	0.011
	GE-SCORE (simple RNN)	0.233	0.215	0.253	0.028
	GE-SCORE (LSTM)	0.229	0.209	0.246	0.029
	GE-MIMIC (simple RNN)	0.216	0.187	0.248	0.036
	GE-MIMIC (LSTM)	0.213	0.191	0.231	0.026
$\psi_1 = 0$, $\psi_2 = 0.2$, $\psi_3 = 0.2$, $\psi_4 = 0.1$ Non-Logistic 20.3% Censored	G-est 1-d	0.284	0.272	0.296	0.019
	G-est 2-d (L)	0.214	0.149	0.246	0.089
	G-est 2-d (Gender)	0.252	0.234	0.269	0.023
	Weibull MSM	0.250	0.242	0.258	0.013
	GE-SCORE (simple RNN)	0.236	0.208	0.263	0.037
	GE-SCORE (LSTM)	0.233	0.197	0.261	0.042
	GE-MIMIC (simple RNN)	0.215	0.195	0.235	0.026
	GE-MIMIC (LSTM)	0.218	0.187	0.249	0.036

Table 3.3: Simulation Results for Scenario III.

3.10.4 Scenario IV

In scenario IV, the true causal effect is a nonlinear function of Gender, L, and an interaction between them. We first assume exposure is modified by L and gender, as well as an interaction between them, similar to scenario III. Then, a nonlinear transformation is applied using the function:

$$\hat{\psi}_k = f(x_t) = -\log\left[\frac{3}{1+\exp(-x_t^2)}\right]. \tag{3.27}$$

In which the x_t is calculated per time point. $x_0 = \psi_1 + \psi_2 Gender + \psi_3 L_1 + \psi_4 Gender \times L_1$, when $k > 0$, $x_k = \psi_1 + \psi_2 Gender + \psi_3 \sum_{t=1}^{k}(\frac{1}{t}\bar{L}_{k-t}\bar{A}_{k-t}) + \psi_3 L_k + \psi_4 Gender \times L_k$. Finally, we let $\psi_1 = 0, \psi_2 = 0, \psi_3 = 0.4, \psi_4 = 0.2$. The resultant nonlinear function

(3.27) indicates that the exposure effect increases whenever the x diverges from 0, but the effect caps at a certain level (Figure 3.6). Note the function is symmetric about 0, thus the sign of ψ_3, ψ_4 does not influence the sign of $\hat{\psi}$, the exposure is consistently beneficial for the outcome (because the sign of $\hat{\psi}$ is negative). However, we can test a harmful exposure effect case by using a nonlinear transform such as $f(x) = \log\left[3/(1+\exp(-x^2))\right]$. We also considered linear and nonlinear exposure mechanisms for this scenario.

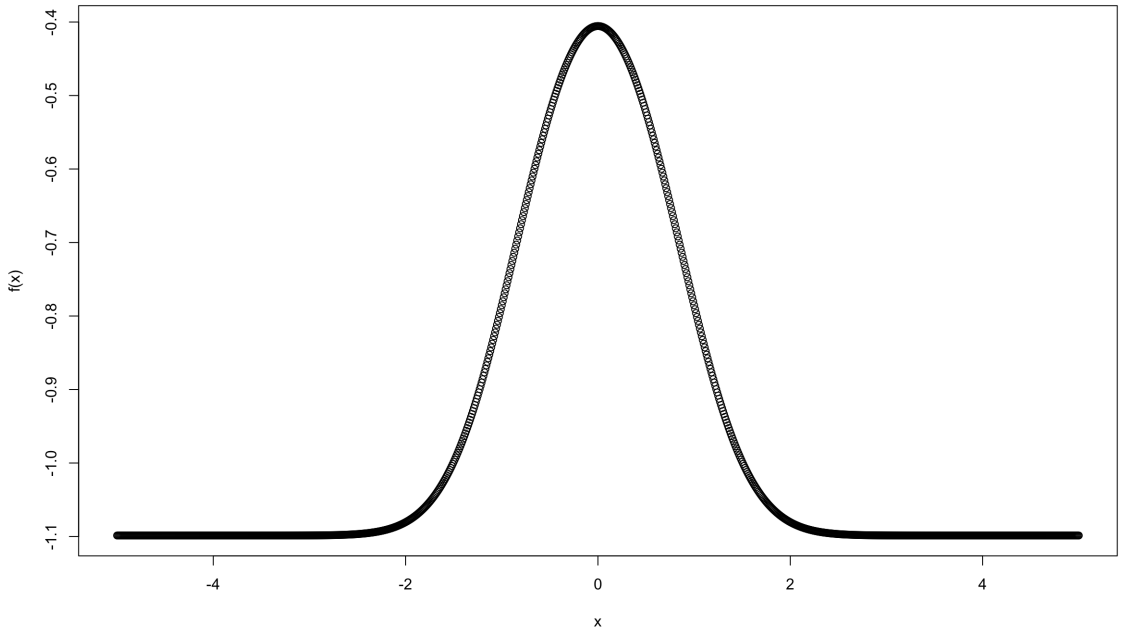


Figure 3.6: Nonlinear transformation function Formula (3.27) applied in S4.

Parameter Exposure Mechanism Percent Censored	Estimation Method	Mean Absolute Bias	25% Absolute Bias	75% Absolute Bias	Bias s.d.

Beneficial effect Logistic 44.0% Censored	G-est 1-d	0.233	0.214	0.253	0.026
	G-est 2-d (L)	0.297	0.248	0.344	0.064
	G-est 2-d (Gender)	0.197	0.178	0.217	0.025
	Weibull MSM	0.184	0.172	0.193	0.017
	GE-SCORE (simple RNN)	0.168	0.145	0.187	0.033
	GE-SCORE (LSTM)	0.157	0.130	0.178	0.035
	GE-MIMIC (simple RNN)	0.159	0.133	0.181	0.036
	GE-MIMIC (LSTM)	0.148	0.124	0.171	0.034
Beneficial effect Non-Logistic 41.5% Censored	G-est 1-d	0.219	0.191	0.241	0.037
	G-est 2-d (L)	0.362	0.278	0.394	0.139
	G-est 2-d (Gender)	0.193	0.166	0.216	0.037
	Weibull MSM	0.171	0.158	0.184	0.023
	GE-SCORE (simple RNN)	0.155	0.127	0.176	0.037
	GE-SCORE (LSTM)	0.153	0.124	0.170	0.044
	GE-MIMIC (simple RNN)	0.168	0.135	0.187	0.050
	GE-MIMIC (LSTM)	0.152	0.126	0.172	0.037
Harmful effect Logistic 15.6% Censored	G-est 1-d	0.156	0.139	0.171	0.021
	G-est 2-d (L)	0.265	0.247	0.275	0.026
	G-est 2-d (Gender)	0.161	0.145	0.174	0.021
	Weibull MSM	0.143	0.131	0.150	0.016
	GE-SCORE (simple RNN)	0.131	0.112	0.144	0.025
	GE-SCORE (LSTM)	0.127	0.110	0.139	0.028
	GE-MIMIC (simple RNN)	0.125	0.108	0.134	0.023
	GE-MIMIC (LSTM)	0.120	0.097	0.136	0.027
Harmful effect Non-Logistic 18.2% Censored	G-est 1-d	0.157	0.139	0.173	0.028
	G-est 2-d (L)	0.374	0.326	0.415	0.081
	G-est 2-d (Gender)	0.164	0.144	0.182	0.027
	Weibull MSM	0.142	0.125	0.156	0.024
	GE-SCORE (simple RNN)	0.131	0.113	0.144	0.028
	GE-SCORE (LSTM)	0.130	0.109	0.146	0.034
	GE-MIMIC (simple RNN)	0.122	0.107	0.140	0.030
	GE-MIMIC (LSTM)	0.118	0.098	0.134	0.030

Table 3.4: Simulation Results for Scenario IV.

3.11 Discussion

G-estimation has benefits over other methods which adjust for time-varying confounding. MSM cannot adjust for any potential effect modifiers; instead, it's predicting the same acceleration rate for all subjects. Under our simulation schema, Weibull MSM should (and indeed does) yield the same estimation as G-estimation 1-d under Scenario I (Young et al., 2009). Though MSM performed very well in Scenario I, this is a result of a correct analysis model specification - such as the Weibull model we choose for the MSM is consistent with the simulation. When examining the MSM results for Scenario II, it's clear that MSM performed much worse compared to Scenario I, as well as the one-dimensional G-estimation. This is because these two methods both failed to consider the effect modifier - the covariate L. One of the significant benefits of MSM is that it can be estimated by standard off-the-shelf software. Nevertheless, the efficiency of MSM is usually the worst among the three methods discussed in Section 2.5 (Daniel et al., 2012) due to the possibly unstable weights. The efficiency of G-estimation typically lies between G-computation and MSM. Often, G-estimation is preferred over G-computation as users are free from identifying and learning the distribution of high dimensional covariates. Moreover, G-computation focuses on predicting the potential outcomes under different regimes directly. There is no set of parameters that quantify the intervention causal effects such as the $\hat{\psi}$ estimated by G-estimation in this dissertation.

The one-dimensional G-estimation is easy to conduct and it's always estimating the mean of the causal survival ratio. Consequently, G-estimation on 1-dim has a small and robust standard error. Consistency of high dimensional G-estimation first depends on the

form of the examined blip function. If we picked the correct blip function (e.g., in Scenario II, estimating ψ_1, ψ_2 with L as the effect modifier), the resultant estimation is close to the true value. Nonetheless, the test statistics used to pick the estimator is not close to 0 under most simulations. Furthermore, a two-dimensional grid search is still manageable. For higher dimensional G-estimation, a non-derivative based optimization algorithm such as the simplex algorithm has to be used but its convergence is always questionable (Joffe et al., 2011). The high dimensional G-estimation becomes significantly unstable when the exposure mechanism is not logistic, that is when the estimation method suffers from misspecification of the exposure model. The 2-d G-estimation with L as the effect modifier has the largest variation of bias when the exposure mechanism model is wrong. This is mainly due to the fact that the exposure mechanism only relates to previous exposure history and current covariate L. Thus, the type I error rate of 2-dimensional G-estimation with L exploded and significant inflation of bias was observed, for all scenarios. On the other hand, one-dimensional G-estimation is less affected, although the standard deviation of bias increased, the mean bias stayed almost the same. Overall, the bias of the G-estimation method increases with the complexity of the blip function. Especially when we move to the non-linear blip function case in Scenario IV, the G-estimation is usually unsuccessful as it's barely possible to write out the correct blip function format.

GE-SCORE is straightforward, especially in the binary exposure case. The exposure prediction network actually reduces to a basic logistic regression if there is no hidden layer. GE-MIMIC is more general in the sense that conditional independence is not enforced by a zero correlation coefficient, but by the similarity of correlation between observed samples and null samples. It is more robust in the sense that the non-linear de-

pendence can be captured if a more comprehensive metric is employed. The conditional independence will be assessed with higher confidence.

In terms of computational burden, both algorithms require fitting an exposure model which does not raise too many challenges providing available mature algorithms for this task. Nevertheless, GE-MIMIC takes extra time to generate null samples for the following steps. The more batches (M) of samples we required, the more processing time the algorithm needed. A reasonable choice such as $M = 50$ is acceptable. A serious computation challenge raises when the loss function is designed to capture the nonlinear dependence, as its computation requires projections into higher dimensional spaces.

Different correlation metrics proposed in Section 3.6 have been explored for GE-MIMIC. However, there seem no benefits of employing a more complicated correlation measure such as DC. Some experiments have been conducted, but with significantly increased running time, the performance barely improved. The first possible reason is comparing to PC, DC is much hard to compute as well as its derivative, creating a burden for optimization. Second, given the fact that we are not testing conditional independence as in other applications, the power is a major concern. If a nonlinear dependence cannot be captured, the type II error will highly likely be inflated. In this work, a metric that captures the linear dependency between two samples suffices for our purpose.

In the simulation, we set $\gamma_1 = 100$ and $\gamma_2 = 10$ based on the simulated sample size and a general expectation of blip function format (the simplest as a constant or complicated as a non-linear format). The penalty term may not be necessary if the constant effect case is not possible at all. We have also considered the simple RNN unit and LSTM unit for the counterfactual network. The choice of the recurrent unit directly relates to the

complexity of the blip function it can handle. From the simulation, LSTM performs better if the simulated blip functions are more complicated such as in Scenarios III and IV. On the other hand, it could introduce too much variance when the true blip function is a constant. Using other more advanced recurrent units such as GRU is completely possible and easy. The neural network structure is subject to tuning. Adopting a more data-adaptive neural network can bring extra benefits.

G-estimation depends crucially on the untestable assumption of no unmeasured confounders. In particular, it relies on having all variables determining exposure both observed and included in one model. Therefore. it is a primary goal of epidemiologists conducting an observational study to collect data on a sufficient number of covariates to make the no unmeasured confounders assumption at least approximately true (Robins, 1997). If the factors related to exposure are not measured, then there may still be bias in the G-estimation (Tilling et al., 2012). Thus, all factors that determine the exposure mechanism need to be well collected and well measured for G-estimation to be successful. On the other hand, the proposed algorithms are more robust.

3.12 Limitations and Future Work

An important limitation of the proposed algorithms we would like to point out first is that the correlation coefficient only reflects a trend (either linear or nonlinear) between two random variables. This is the intrinsic property of the correlation coefficient which cannot be solved by switching to more comprehensive correlation measures such as DC. The intermediate output of the proposed network - the counterfactual survival time and

the causal estimand $\hat{\psi}_k$ cannot be uniquely identified while the proposed loss functions indeed ensure zero correlations in GE-SCORE. Under the null hypothesis, random noise can be applied to examined sequences without influencing the correlation coefficient between them. In other words, more than one sequence can yield near zero correlation coefficients, provided another sequence. By the consistency assumption, the counterfactual unexposed time is the observed time if the subjects stay unexposed. Thus, the algorithms rely on those unexposed samples to succeed. The existence of these null samples is ensured by the positivity assumption. The overall conclusion is, without a pre-specified blip function format, learning the causal estimand $\hat{\psi}_k$ can be challenging, especially in the high-dimensional case. Adding the penalty term of the prediction variation partially mitigates the problem since predictions with large variances are discouraged. From the simulation, we indeed observed a significant boost in estimation performance. Nonetheless, the G-estimation method is semiparametric in the sense we left the counterfactual survival distribution as well as the law of covariates unspecified while focusing on predicting the causal survival ratio. If we are willing to make more parametric assumptions, the loss can be expanded to incorporate likelihood and more explicit estimations would be expected.

Following the specification of the baseline survival time, another interesting research question is the direct prediction of the counterfactual survival time. To do that, besides the G-estimated causal survival time ratio, we further need to specify the law of covariates such as $f(L_k|\bar{L}_{k-1}, \bar{A}_{k-1}, H, C)$ and $f(h|C)$ (Robins et al., 1992), where H is random variable for the true baseline survival time and C is the censoring random variable. It would be difficult if not impossible to specify these parametric models as H is

not observable. Yet, neural networks remain useful for performing the task as they can be viewed as universal approximators for any functions (Hornik et al., 1989).

We have also considered using a trained RCGAN to generate the null samples as required in GE-MIMIC. The proposed algorithm can be named GE-RCGAN. The generative network should be trained with all covariates that are believed related to the exposure mechanism, before implementing any following steps. GANs have recently stood out as a powerful approach for learning and generating random samples from a complex, high-dimensional data distribution. The RCGAN still follows the architecture of a regular GAN, but with both the generator and the discriminator being replaced by RNNs which can generate sequences of real-value data subject to some conditional inputs. We argue that providing the same set of inputs during training motivates the algorithm to learn the conditional distribution. This method is anticipated to be less time efficient as it requires a separate training step of the RCGAN.

GE-RCGAN is a very computationally intensive algorithm. The performance of RCGAN directly relates to the quality and amount of available data. The adversarial net is performing an unsupervised learning task. The global minimum of the loss can only be obtained provided with enough data capacity. Given the fact that we centered our interest on binary or (more general, discrete) exposure, the GE-MIMIC seems sufficient. Noises that are to be added to the null sample generation process for GE-MIMIC are critical though we didn't explore that in this dissertation. The goal is to introduce some randomness to the null samples; hence the empirical distribution (not only the mean) can be approximated appropriately. For more complicated exposure mechanisms such as continuous ones, considering a more complicated null sample generation mechanism could

bring some benefits. Nevertheless, in previous literature, the regression-based algorithm has a better performance than the RCGAN counterpart (Sen et al., 2018).

Unfortunately, in settings with numerous covariates, knowledge of a complete causal diagram, including the causal relationships among all the possible covariates themselves is often unavailable (VanderWeele, 2019). Identification of the correct DAG is mostly based on the relevant domain knowledge. There are graph-based ML algorithms available for causal discovery (Vowels et al., 2022). The identification of the causal DAG falls out of the scope of this dissertation.

3.13 Conclusion

The SNAFTM is an effective method to control time-varying confounding which is often encountered in observational studies. The intervention causal effect would be biased if one fails to adjust for the time-varying confounding effects. G-estimation is the suggested method to estimate a SNAFTM. However, this method is much under popular due to the computation challenges and implementation difficulties. Two proposed ML algorithms - GE-SCORE and GE-MIMIC were proven to be efficient in estimating a SNAFTM. These algorithms have benefits such as the capability to handle high dimensional data, the non-requirement of parametric blip function formats, and robustness to different intervention mechanisms. The intermediate output of the network has clinical meanings. It can be interpreted as quantifying amount, the event free time is expanded or compressed due to the intervention. By using simulated and real observational study data, we showed that these algorithms are flexible enough to provide clinically meaning-

ful estimation, under a wide range of scenarios.

Chapter 4: Application

4.1 Research Question

The research question is the causal effect of smoking (A) on the time to the first occurrence of any cardiovascular disease (CVD) events (Y). Even though an RCT is the gold standard for identifying a certain exposure effect, it is unethical to conduct an RCT to investigate the smoking effect on health. Nevertheless, some epidemiological research has indicated that smoking substantially increases the risk of coronary heart disease (CHD) (Ockene and Miller, 1997). Smoking also acts synergistically with other risk factors (Anderson et al., 1991). Furthermore, smokers are at an increased risk for peripheral vascular disease, cancer, chronic lung disease, and many other chronic diseases. A more recent meta-analysis analyzed 141 cohort studies from 1946 to 2015 on smoking and heart disease. The conclusion is, smoking one cigarette a day raised a male's risk of heart disease by 48% on average over a non-smoker, while smoking 20 cigarettes a day doubled the risk. The risk increases more significantly for women: having one daily cigarette increased the heart disease risk by 57% while smoking 20 cigarettes a day raised the risk by 2.8 times (Hackshaw et al., 2018). However, numerous prospective investigations have demonstrated a rapid decrease in risk with smoking cessation for CHD mortality and stroke (Burch, 1974; Ockene et al., 1990; Wolf, 1988). In the physician's

office, patients are often conscious of their health and most receptive to risk factor intervention, providing an important opportunity for smoking cessation intervention (Ockene, 1987). If the individualized causal relationship between smoking and heart disease could be identified, healthier lifestyles can be recommended at the individual level and bring immediate benefits.

The common challenge of using longitudinal observational data to quantify the causal effect of smoking on the time to CVD event is the time-varying confounding issue, as discussed in previous chapters. There are many potential time-varying confounders for this research problem. For instance, diabetic status can be a time-varying confounder for smoking as it satisfies the three conditions specified in Section 2.3. Simply speaking, diabetes itself is a risk factor for a cardiovascular event. Diabetic status is influenced by previous smoking status. Furthermore, a subject may quit smoking exactly due to positive diabetic status. The same logic applies to other potential time-varying confounders such as blood pressure (BP). See Figure 4.1 for a simplified DAG of this research problem. The black dashed arrows running from A_k to Y denote the hypothesis (whether there is a causal effect of smoking on time to the first CVD event or not) that we are interested in, the grey dashed arrow running from L_k to the Y denote the effects of the covariates on time to the first CVD event. The time-varying confounding issue could also be introduced by the unmeasured confounder U of L_k and Y. The causal effect of smoking (A) on the outcome Y can only be identified when these time-varying confounders (L_k) are controlled appropriately. It's possible for researchers to use appropriate methods to adjust for time-varying confounding. To do that, we first recognized that the details of the exposure (e.g., whether smoking or not, number of cigarettes smoked daily), time-varying

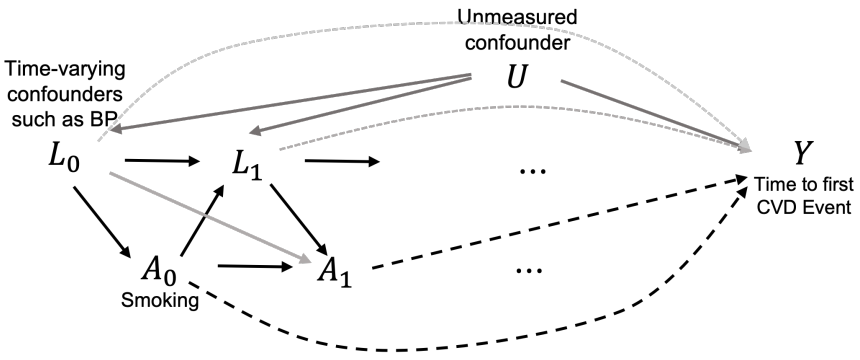


Figure 4.1: A simplified DAG of time-varying smoking status causal effects on time to first CVD event.

confounders at each visit and outcome of interest should be available to the investigator. In Section 4.2, we would introduce an observational study that fits this research purpose perfectly. Thus, the SNAFTM would be an excellent choice to quantify the exposure effect on the time-to-event outcome. However, given the implementation difficulties of the existing G-estimation method, the proposed algorithms - GE-SCORE and GE-MIMIC are expected to shed more light on this research question. Other existing methods were also explored and the results of different methods were compared.

4.2 CARDIA

Coronary Artery Risk Development in Young Adults (CARDIA) Study (Friedman et al., 1988; Loria et al., 2007) examined the development and determinants of clinical and subclinical cardiovascular disease and their risk factors. This study is a population-based observational study of 5,115 participants aged $18-30$ years recruited during $1985-1986$, from four centers in the United States (Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA). The sample was designed to achieve approximately balanced

subgroups of race, gender, education (high school or less, and more than high school), and age ($18-24$ and $25-30$ years old). After the enrollment, follow-up in-person examinations were conducted in $1987-1988$ (Year 2), $1990-1991$ (Year 5), $1992-1993$ (Year 7), $1995-1996$ (Year 10), $2000-2001$ (Year 15), $2005-2006$ (Year 20), $2010-2011$ (Year 25), and $2015-2016$ (Year 30), with retention of 90.5%, 85.6%, 80.6%, 78.3%, 73.6%, 72.0%, 72.2% and 71.0% of living participants, respectively. Other than these in-person visits which collected extensive measures at a subject level in person, there was also yearly contact via letter or phone. The study examination is still ongoing for the remaining participants (the Year 35 examination was conducted from June 2020 to July 2021, but data has not been released). Because this study started to follow subjects during their young adulthood, it has provided evidence that supports intervening during young adulthood to prevent CVD mortality in later decades (Liu et al., 2012). There has also been some research looking into the smoking effect on blood pressure (Luehrs et al., 2021) and the cardiac function (Gidding et al., 1995), using CARDIA data. Nevertheless, none of these studies used the time-to-event outcome, nor considered the causal effect of smoking.

Table 4.1 provides the baseline table for selected characteristics, of the CARDIA study. Note, the sample size for the table is 5,113 instead of 5,115, because two subjects do not have baseline visit data. This study was fairly balanced in terms of gender and race. More than 30% of subjects are current smokers at the baseline. Thus our exposure of interest is not rare in this population.

At each clinical visit, in addition to standard time-varying measurements such as blood pressure, anthropometry, blood lipids, smoking behavior, physical activity, diet,

Characteristics	**Category**	**Count (%)** N=5,113
Sex	Male	2328 (45.5%)
	Female	2785 (54.5%)
Race/Ethnicity	Hispanic	14 (0.3%)
	Black not Hispanic	2642 (51.7%)
	White not Hispanic	2455 (48.0%)
	Don't know	2 (0.0%)
Age at Enrollment	<= 18	254 (5.0%)
	19 - 25	2464 (48.2%)
	26 - 29	1925 (37.6%)
	>= 30	470 (9.2%)
Marital Status	Married	1135 (22.2%)
	Marriage-like Relationship	363 (7.1%)
	Separated/Divorced/Widowed/ Never Married	3593 (70.3%)
	Unknown/Missing	22 (0.4%)
Working Status	Working Full/Part-time	3485 (68.2%)
	Unemployed	505 (9.9%)
	In School	671 (13.1%)
	Keeping House or other Activity	435 (8.5%)
	Unknown/Missing	17 (0.3%)
Education	High School or Lower	2055 (40.2%)
	College	2604 (50.9%)
	Graduate School or Higher	445 (8.7%)
	Unknown/Missing	9 (0.62%)
Smoking Status	Non-Smoker	2891 (56.5%)
	Current Smoker	1566 (30.6%)
	Former Smoker	642 (12.6%)
	Unknown/Missing	14 (0.3%)
Last Physical Examination	Less than one year	1624 (31.7%)
	1 - 2 years	1212 (23.7%)
	2 - 4 years	1108 (21.7%)
	5 or more years	912 (17.8%)
	Never	139 (2.7%)
	Unknown/Missing	118 (2.3%)

Table 4.1: Baseline Characteristics Table before data cleaning and manipulation, CARDIA

pulmonary function, and many psychological factors, the CARDIA study has included other measurements (in subsets or the full cohort) to obtain unique information on other aspects of risk factor development and early morbidity. These have included but are not limited to graded exercise treadmill testing; echocardiography, particularly for measurement of left ventricular mass; cardiovascular reactivity; serum cotinine and sympathetic nervous system activity, etc.

The CARDIA study has high-quality data giving a satisfying retention rate and great compliance of subjects. A large amount of data and rich covariates were collected for individuals during the long range of follow-up (30-35 years by 2021). Per the study schedule, a subject can have at most 9 visits. Currently, there are 39,680 subject-visits records available for our analysis before data cleaning; on average, each subject contributed 7 visits. The study targeted collecting risk factors starting from comparable young ages and following the change while subjects aged. On the one hand, a large amount of data should provide researchers with better opportunities to gain more insights into cardiovascular disease progress. Alternatively, this also poses a burden on conventional statistical analysis methods. Moreover, given repeated measures of the same risk factors across examinations, traditional statistical models may lose power if the inter-correlated nature among covariates and exposure is not incorporated.

Last, the endpoint of this study includes death, CVD events, etc. Endpoints-related events are being adjudicated and are released to the public regularly. There is an Endpoints Surveillance and Adjudication Subcommittee (ESAS) to define endpoints of interest to the study and to determine what information related to these endpoints should be collected. The committee also identifies the criteria for an endpoint and adjudicates

whether a particular episode of care met the established criteria. There are two types of endpoints in the CARDIA study, the mortality endpoints include all cases of death occurring in CARDIA participants; the morbidity endpoint includes an incident event in a CARDIA participant resulting in a minimum 24-hour hospitalization for events such as Myocardial infarction (MI), coronary revascularization, non-MI acute coronary syndrome, congestive heart failure (CHF) and stroke, etc. Event identification occurs during exams and annual follow-ups of cohort members. The outcome event of our research interest - any CVD event (fatal or non-fatal) - has been observed for 355 (6.94%) subjects as of Dec 31, 2021. Hence, most subjects are right-censored. Due to staggered entry, subjects had slightly different on-study times. We use the reported time-to-censor or time-to-event as the event time to be analyzed because the actual calendar enrollment time/event times have been masked from public use. The maximal follow-up time reported is 425.5 months as of Dec 31, 2021, the mean is 408.6 months with (Q1, Q3) = $(405.7, 411.3)$. Given the long follow-up range, the staggered entry during the first $1-2$ years and the slightly different individual visit schedules were ignored to ease the computation. As a result, we use a standard visit schedule at months (0, 24, 60, 84, 120, 180, 240, 300, 360) for all subjects and use 425 months as the end of study time.

4.3 Data Cleaning and Missing Data

In any real observational study, it's fairly common that subjects skip visits due to schedule conflicts or simply because they don't want to attend the study too often. Nevertheless, the covariates/exposure levels were recorded only at the clinical visits. To

fill in missing observations, the most reasonable assumption is that everything remains the same until the next time when the data are updated. This is known as Last Observation Carried Forward (LOCF) (Lachin, 2015). If an observation is missing from the baseline, then there is no non-missing value to carry forward. In this case, we would sort the data backwards and then repeat the LOCF. Thus, a variable will only be missing if there is no value collected at all during the study course. The corresponding subject would be removed from the analysis if the model required complete records.

All statistical models are prone to extreme outliers and an unbalanced distribution of categorical data. The above-mentioned LOCF cannot take care of this. In practice, a model usually fails to converge if there are a negligible number of observations in a stratum, such as a few subjects with missing or unknown responses. A minimal amount of imputation was employed to fill in the missing records or remove these observed extreme unbalanced levels. For instance, for numeric variables, we replaced missing with the population mean: a smoker with smoking intensity missing will be replaced with the population-level mean smoking intensity. Another example is, only a few subjects with drinker status are unknown, we hard-coded those to "not a drinker", to facilitate the model convergence. To reduce the influence of potential numeric outliers, we categorized some numeric variables (such as education years) and treated them as categorical variables in the following analysis. These data-cleaning techniques were considered only for variables that were included in the models. Lastly, because we focus on the time to the subject's first CVD event, once the event is observed (fatal or non-fatal), later records are discarded from the analysis.

4.4 Results

4.4.1 IPCW for competing risk censoring

First note that there is a substantial proportion of subjects who have not experienced the well-defined endpoint in the CARDIA study as of Dec 31, 2021. As proposed in Section 3.7, we used the inverse probability of censoring weighting to deal with competing risk events (Howe et al., 2011). In CARDIA, a non-CV death event should be treated as a competing risk event for a CVD event if the subject died before experiencing any CVD event. Though the overall retention rate is surprisingly good for an observational study that lasts for more than 30 years, there are some subjects who never came back for scheduled study visits due to various (unknown) reasons other than death. We treated these subjects as lost to follow-up if there is no more information that can be located for them after a certain visit. Though the lost to follow-up reason may have been ignorable, such as moving out from the study area, to be conservative we treated lost to follow-up as a competing risk event. Other possible outcomes of subjects after each visit (until the next visit) include experiencing the event of interest (we simply don't consider what happened afterward once the first event of interest has been observed) and still under follow-up. Table 4.2 provides some basic information on the time-varying exposure (smoking) and subject disposition, by visit year.

In the following analyses, subjects with these two types of outcome events (death and lost to follow-up) will receive weights 0. Subjects who were under follow-up or experienced the CVD event will be reweighted to stand for those subjects with weights

Visit Year	Smoking A=1	Outcome by Next Visit			
		Stay under follow-up	CVD event	Non-CV death	Lost-to-follow-up
00 (n=5111)	1564 (30.6%)	4987 (97.6%)	0 (0%)	6 (0.1%)	118 (2.3%)
02 (n=4987)	1519 (30.5%)	4828 (96.8%)	2 (0.0%)	27 (0.5%)	130 (2.6%)
05 (n=4828)	1436 (29.7%)	4716 (97.7%)	5 (0.1%)	12 (0.2%)	95 (2.0%)
07 (n=4716)	1347 (28.6%)	4576 (97.0%)	8 (0.2%)	31 (0.7%)	101 (2.1%)
10 (n=4576)	1243 (27.2%)	4355 (95.2%)	40 (0.9%)	37 (0.8%)	144 (3.1%)
15 (n=4355)	1078 (24.8%)	4093 (94.0%)	55 (1.3%)	48 (1.1%)	159 (3.7%)
20 (n=4093)	921 (22.5%)	3756 (91.8%)	91 (2.2%)	56 (1.4%)	190 (4.6%)
25 (n=3756)	708 (18.8%)	3249 (86.5%)	98 (2.6%)	73 (1.9%)	336 (8.9%)
30 (n=3249)	475 (14.6%)	3106 (95.6%)	56 (1.7%)	87 (2.7%)	0 (0%)

Table 4.2: Time-dependent Exposure and Visit Outcome by Visit Year, CARDIA

0. Let $C_{k,i} = 1$ denote the event that subject i is still under follow-up at visit k. The unstabilized weights $W^c_{k,i}$ were calculated by:

$$W^c_{k,i} = \prod_{s=0}^{k} \frac{1}{P(C_{s,i} = 1 | \bar{C}_{s-1,i} = 1, \bar{A}_{s-1,i}, \bar{L}_{s-1,i}, C_{s-1,i} = 1)}. \tag{4.1}$$

Unstable weights can occur when few values of non-right-censored are found in some strata as defined by the conditioned part of $W^c_{k,i}$. The calculated weights can further be stabilized by using a different numerator than 1. We replace the numerator of above W^c with the probability that a subject remains uncensored at visit k conditional on treatment history (Robins et al., 2000) and subject-specific static characteristics L_{0i}. In this study, we considered statics characteristics: gender, age at enrollment, and race. Note in the denominator, the covariate vector L has included the static part. The stabilized weights $SW^c_{k,i}$ are calculated by:

$$SW^c_{k,i} = \prod_{s=0}^{k} \frac{P(C_{s,i} = 1 | \bar{A}_{s-1,i}, L_{0i}, C_{s-1,i} = 1)}{P(C_{s,i} = 1 | \bar{C}_{s-1,i} = 1, \bar{A}_{s-1,i}, \bar{L}_{s-1,i}, C_{s-1,i} = 1)}. \tag{4.2}$$

In both Formulas (4.1) and (4.2), the denominator is the probability of a subject remaining uncensored at visit k given treatment and time-dependent covariate histories. The conditional non-censored probability at each visit was computed first and then multiplied cumulatively to obtain the final up to visit k non-censored weights. In CARDIA, 377 subjects died from reasons other than CVD events, and 1,273 subjects were lost to follow-up before experiencing any CVD event. Due to the small number of visits with these competing risk events, we combined these two into one category inclusive of all competing risk

events. In this dissertation, we used many-to-many RNNs to model the subject-specific probability of remaining under follow-up at each visit. RNNs with different inputs were used for the denominator and numerator (stabilized weights only). Though no extreme weights were observed (Figure 4.2) before stabilization, the stabilized weights are always advocated. Furthermore, it's a common practice to truncate upper/lower 1% of stabilized weight values. The distribution of weights was not truncated in this application given the comparable centralized distribution of stabilized weights (Figure 4.3).

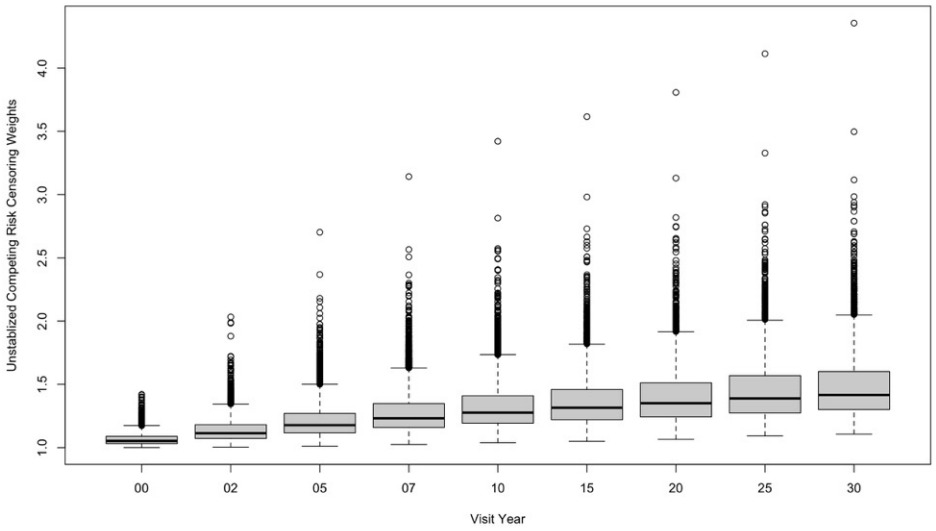


Figure 4.2: Boxplots of unstabilized inverse probability of censoring weights by the visit in CARDIA.

The IPCW used in G-estimation was computed similarly but with small modifications of multiplication order. Because subjects at risk at visit $k+1$ are a subset of subjects at risk at visit k. Thus, the unstabilized weights $W^g_{k,i}$ are now calculated as the following:

$$W^g_{k,i} = \prod_{s=k}^{Y_i} \frac{1}{P(C_{s,i} = 1 | \bar{C}_{s-1,i} = 1, \bar{A}_{s-1,i}, \bar{L}_{s-1,i}, C_{s-1,i} = 1)}. \tag{4.3}$$

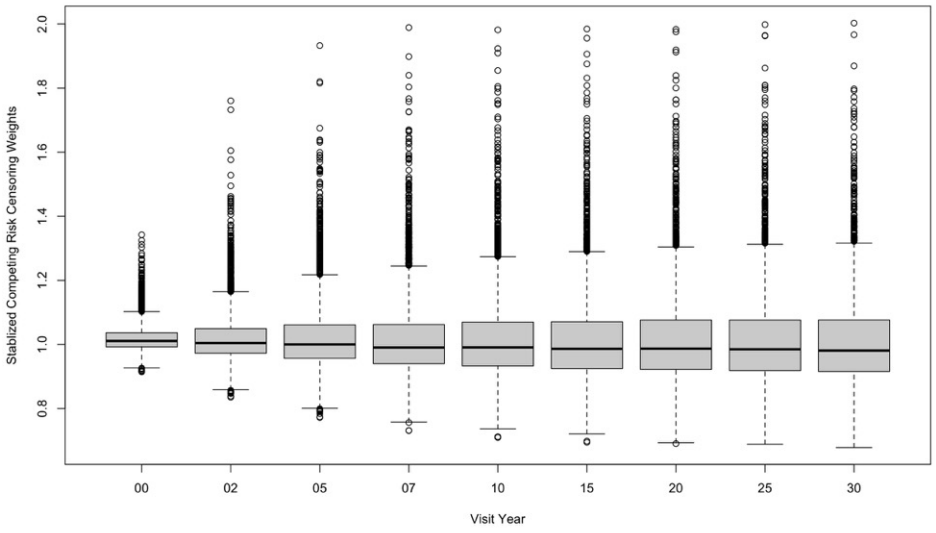


Figure 4.3: Boxplots of stabilized inverse probability of censoring weights by the visit in CARDIA.

The stabilized version of the inverse competing risk censoring weights $SW_{k,i}^{g}$ is (Mertens and Vansteelandt, 2017):

$$SW_{k,i}^{g} = \prod_{s=k}^{Y_i} \frac{P(C_{s,i} = 1 | \bar{A}_{s-1,i}, L_{0i}, C_{s-1,i} = 1)}{P(C_{s,i} = 1 | \bar{C}_{s-1,i} = 1, \bar{A}_{s-1,i}, \bar{L}_{s-1,i}, C_{s-1,i} = 1)}. \tag{4.4}$$

Here, individual conditional stay under follow-up probabilities used in the numerator and denominator were all calculated using many-to-many RNNs. In the G-estimation, the artificial censoring indicator Δ for non-competing risk censored subjects was inversely weighted by these predicted probabilities. This is equivalent to inversely weighting the estimating equation (Formula 3.18). In the proposed GE-SCORE and GR-MIMIC, non-competing risk censored records were inversely weighted while calculating the objective function.

4.4.2 Traditional Survival Analysis Methods

Some traditional statistical analysis methods were explored at the beginning of the analysis. We considered the Cox PH model with baseline characteristics only, the Cox PH model with time-dependent covariates, the Weibull PH model with baseline characteristics only, and the Weibull PH model with time-dependent covariates. We chose the PH model to facilitate the comparison of hazard ratios among all presented models. However, the PH assumption is a strong assumption and HR itself lacks causal meanings as we pointed out in previous sections.

All traditional models controlled for gender (female vs male), race (white vs non-white), age at enrollment (younger than 18, $19-25$, $25-29$, and older than 30), marital status (married or living as married, divorced or separated or widowed, never married), education level (less than high school or GED, some college or college, college above), employment status (full-time/part-time/unemployed), BMI, SBP, self-reported diabetic/hypertension/high cholesterol, exercise intensity score, whether the subject is a drinker or not and family history of heart disease (whether father/mother had a heart attack and whether father/mother had a stroke). After necessary data cleaning, the measurements of these covariates recorded at the first examination are used in all baseline models (N=5,111). In time-dependent models, time-dependent versions of these covariates are used (e.g., the SBP variable was switched to the time-dependent one because it was measured at each clinical visit; also subjects may update their marital status during each visit thus this can also be enriched to the time-dependent version). All these models also require complete records. In total, there were 39,671 records available for the time-

dependent models. All these models identified a passive effect of smoking on time to the first CVD event. In other words, smokers have higher hazards of experiencing CVD events compared to non-smokers. The model fitting results are displayed in Table 4.3. The Cox PH and Weibull PH model yield identical estimations that the hazard of smokers experiencing the event is about 2.27 times that of non-smokers. Both time-dependent models identified more significant smoking effects on time to the first CVD event compared to the baseline models. Nevertheless, the hazard ratios presented here are highly likely to have been biased from a causal perspective.

Models	Hazard Ratio Estimation	Confidence Interval
Cox PH	2.27	(1.80, 2.89)
Time-dependent Cox PH	2.42	(1.88, 3.11)
Weibull PH	2.27	(1.82, 2.84)
Time-dependent Weibull PH	2.50	(2.05, 3.05)

Table 4.3: Estimated hazard ratio of current smokers vs non-smokers on time to first CVD along with 95% CI, CARDIA.

4.4.3 Using G-estimation and MSM to adjust time-varying confounders

To address the time-varying confounding issue as demonstrated in DAG 4.1, the methods discussed in Section 2.5 are necessary. We applied the traditional G-estimation method (one dimensional) as well as the marginal structure model to the CARDIA data, trying to quantify the smoking causal effect on time to the first CVD event while having the time-varying confounding controlled.

4.4.3.1 G-estimation

The G-estimation fitted a SNAFTM:

$$S_{Y^{\bar{a}_{k-1},\bar{0}}|\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{a}_k,Y>t_k}\{\gamma_k(y,\bar{l}_k,\bar{a}_k;\psi)\} = S_{Y^{\bar{a}_k,\bar{0}}|\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{a}_k,Y>t_k}(y). \tag{4.5}$$

with the blip function

$$\gamma_k(y,\bar{l}_k,\bar{a}_k;\psi) = t_k + (y-t_k)\exp(\psi a_k). \tag{4.6}$$

Applying the blip function to the time-to-event outcome Y is equivalent to removing the causal effect of the final blip of smoking at visit k on the outcome. Here we didn't consider any effect modificator for the causal effect. The smoking causal effect is assumed as a constant for all subjects, across time points.

Section 2.7 has presented the procedure of a G-estimation. First, the smoking exposure model was fitted by Generalized Estimating Equations (GEE) with R package `geeglm`. Covariates considered in the exposure model include time-independent variables: gender (female vs male), race (white vs non-white), age at enrollment (younger than 18, 19-25, 25-29, and older than 30); time-varying variables: BMI, education level (less than high school or GED, some college or college, college above), working status (employed full-time/part-time/unemployed), marital status (married or living as married/divorced or separated or widowed/never married), self-reported diabetic/hypertension/high cholesterol, SBP, depression score measured by CES-D scoring system, exercise intensity

score, and alcohol consumption status (drinker or not). Then, the proposed SNAFTM recovers the counterfactual unexposed time if we provide a hypothesized ψ. By using grid search on range $[-1, 1]$ by a step size of 0.01, the estimated causal survival time ratio is 0.57 with a 95% confidence interval $(0.47, 0.66)$ (Figure 4.4).

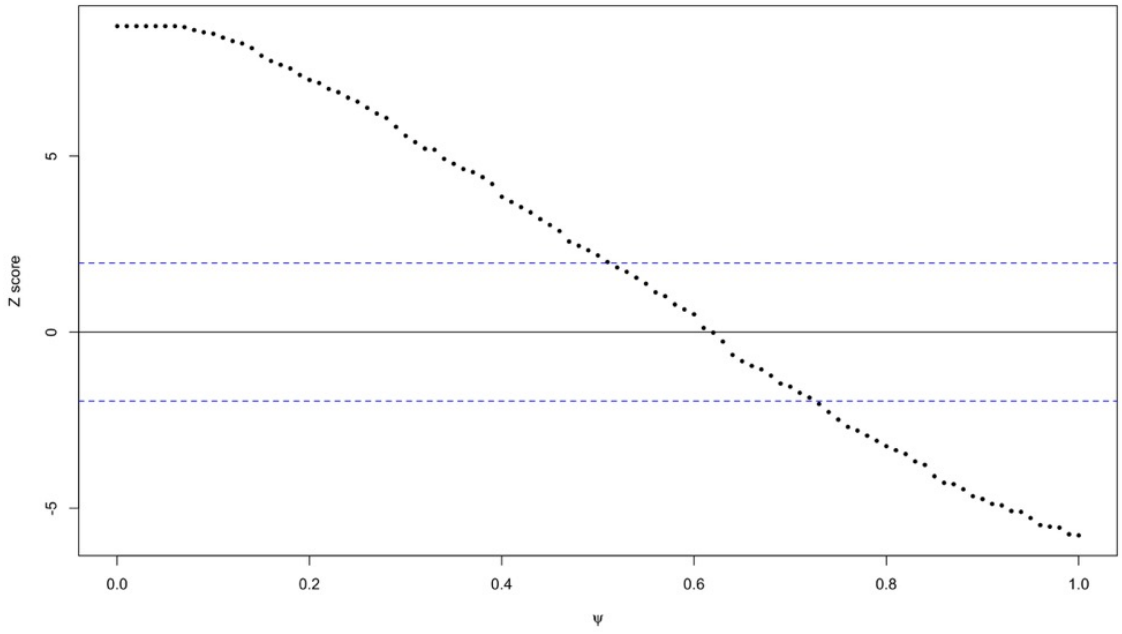


Figure 4.4: G-estimation Result for CARDIA. X-axis is ψ values. Black dots are the Z-score of the counterfactual survival time term in the exposure logistic model. The black solid reference line is where the Z-score = 0 and the blue dashed lines are the interval of $(-1.96, 1.96)$.

The estimation result suggested that smoking as an exposure compresses the CVD-free time by a factor of $\exp(-0.61) = 0.54$ times as of the nonsmoker's CVD event-free time. The G-estimated confidence interval for ψ is $(0.52, 0.72)$.

For all practical purposes, hazards can be thought of as incidence rates, and thus the HR can be roughly interpreted as the incidence rate ratio (Hernán, 2010). The hazard/rate ratio is more commonly used in epidemiology. The causal rate ratio is defined as the

mortality rate if all subjects were exposed from the baseline onward divided by the rate when all subjects were unexposed throughout (Hernán et al., 2005; Witteman et al., 1998). To this end, the acceleration factor $\exp(-\psi)$ can be related to the rate ratio if we further assume that the counterfactual unexposed survival time $Y^{\bar{0}}$ follows a Weibull distribution. That is, we would like to estimate the shape parameter a from a Weibull distribution with survival function $P(Y^{\bar{0}} > y) = \exp((y/b)^a$. The estimated causal rate ratio can be characterized as $\exp(a\hat{\psi})$. For right-censored subjects, we took $C(\hat{\psi}, 0)$ calculated by Formula (3.14) as their time of censoring by the end of follow-up, under ICA. Subjects who were censored by competing risks were excluded from the analysis, but the IPCWs were used to weigh the estimation procedure. The Weibull distribution seems to be a very reasonable distribution for the counterfactual survival time (Figure 4.5). Yet, the fitted curve and the KM curve start to diverge near the end due to the shrinking at-risk set. The shape parameter we obtained is 2.25 with a confidence interval $(2.06, 2.46)$. A shape parameter greater than 1 indicates that the hazard increases with time, which is coincident with our expectation of smoking effect on time to the first CVD event. The corresponding causal rate ratio is $\exp(2.25 * 0.61) = 3.95$ with a confidence interval $(3.22, 5.05)$. Compared with those hazard ratios obtained from traditional models (Table 4.3), G-estimation produced a significantly larger estimated causal effect in terms of the causal rate ratio.

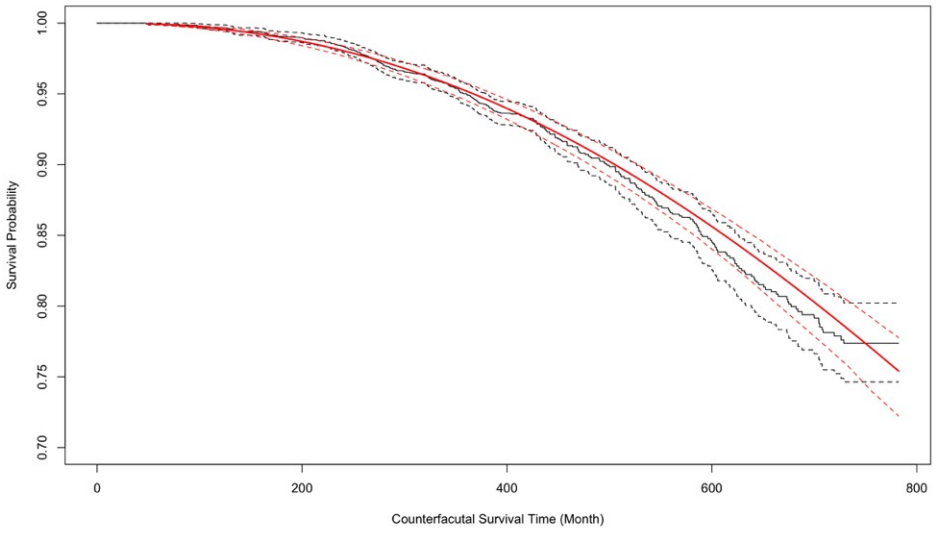


Figure 4.5: Kaplan-Meier curve (black solid line) and 95% CI (black dash lines) of counterfactual survival time U_i, along with fitted Weibull distribution (red solid line) and 95% CI (red dash lines), CARDIA.

4.4.3.2 Marginal Structural Models

Now we consider the Marginal Structural Models. Similar to the competing risk censoring weights Formulas (4.1 and 4.2), the unstabilized/stabilized inverse treatment weights (Robins et al., 2000) at each visit k are computed as:

$$W_{k,i}^{t} = \prod_{s=0}^{k} \frac{1}{P(A_{s,i}|\bar{A}_{s-1,i}, \bar{L}_{s,i}, C_{s-1,i} = 1)}, \tag{4.7}$$

and

$$SW_{k,i}^{t} = \prod_{s=0}^{k} \frac{P(A_{s,i}|\bar{A}_{s-1,i}, L_{0i}, C_{s-1,i} = 1)}{P(A_{s,i}|\bar{A}_{s-1,i}, \bar{L}_{s,i}, C_{s-1,i} = 1)}. \tag{4.8}$$

The exposure mechanism was also fitted by a many-to-many RNN, the same as for the follow-up outcome. For the numerator and denominator in the above weights formula,

RNNs with different inputs were employed. In the numerator, besides exposure history, we considered statics characteristics: gender, age at enrollment, and race. RNNs have been proven competitive in fitting the censoring probabilities and estimating corresponding weights (Lim et al., 2018). Below are boxplots of estimated unstabilized treatment weights (Figure 4.8) and stabilized weights (Figure 4.7) by visit year. From the plots,

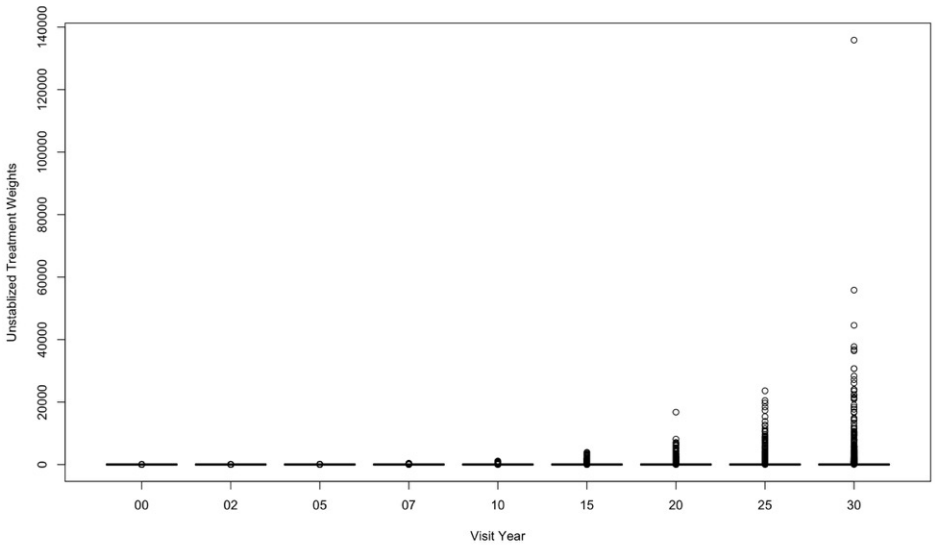


Figure 4.6: Boxplots of unstabilized inverse probability of treatment weights vs. year of visit in CARDIA.

the stabilization is quite critical for the treatment weights because it successfully removed extreme values, and the distribution of weights is much more regularized. The range of the stabilized weights is $(0.25, 6.75)$. If truncate the lower 1% and upper 99%, the range shrink to $(0.56, 2.15)$. Because no extreme weights were observed, we didn't truncate the stabilized weights for treatment in the following computation. The overall weights for the MSM are the product of stabilized inverse competing risk censoring probability weights and stabilized inverse treatment probability weights: $SW_{k,i} = SW^c_{k,i} \times SW^t_{k,i}$. The final

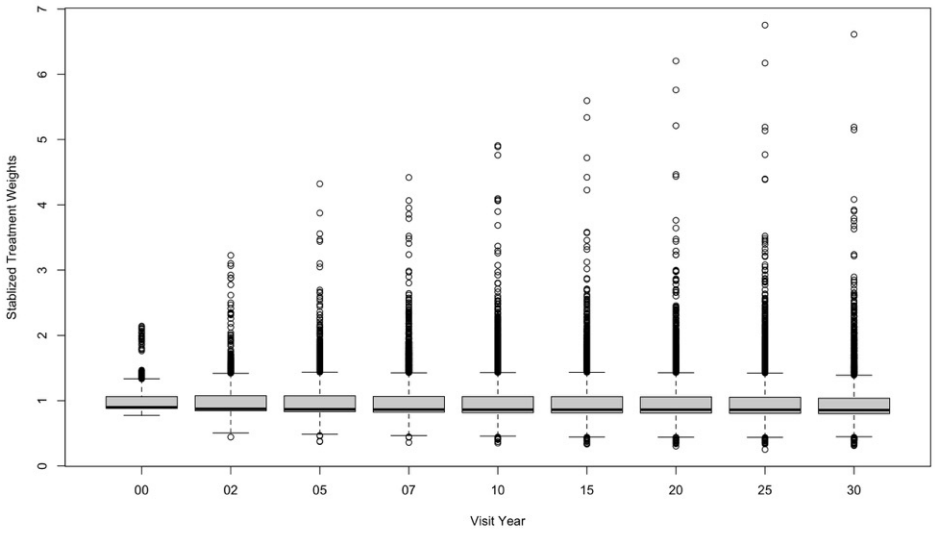


Figure 4.7: Boxplots of stabilized inverse probability of treatment weights vs. year of visit in CARDIA.

weights range from 0.25 to 6.68, with median = 0.90 and mean = 0.99. A necessary condition for the correct model specification is that the mean weight is 1 (Hernán and Robins, 2006), ideally at each time point rather than just overall. The boxplot of log weights and the fact that the mean is close to 1 suggested that the condition is approximately true. The variation of weights grew with follow-up. This is expected as we have fewer subjects remaining under investigation as the study progresses. A marginal Weibull model is fitted with the weighted population while adjusting for sex, age at enrollment, and race, as we adjusted for the numerator part in the inverse. The smoking causal effect on compressing time to the first CVD event estimated by this Weibull MSM is $0.66\ (0.61, 0.72)$. This can be interpreted as smoking compresses the CVD event-free time by 0.66 folds as of the nonsmoker's CVD event-free time. Furthermore, the corresponding causal hazard ratio is $2.86\ (2.32, 3.52)$. This causal hazard ratio is larger than the traditional Weibull model,

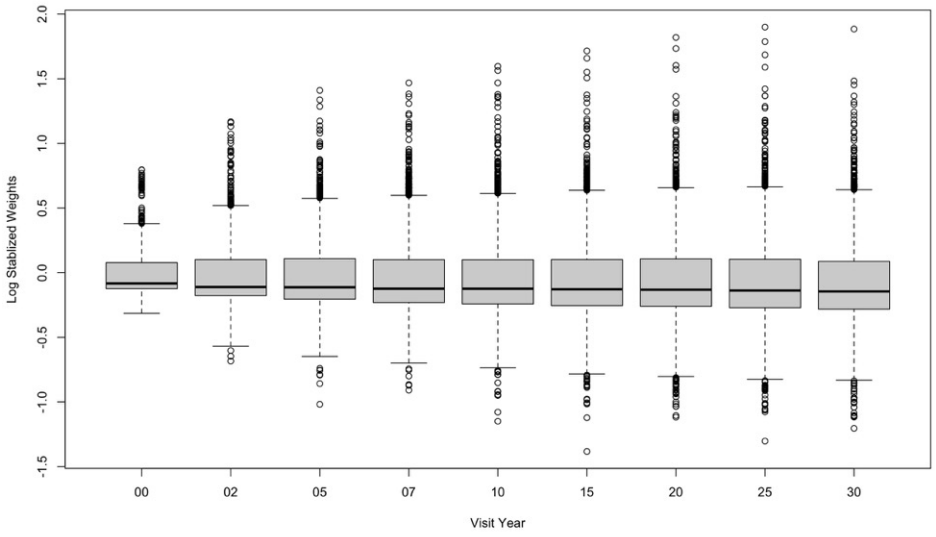


Figure 4.8: Boxplots of log stabilized inverse probability weights by the visit in CARDIA.

which did not adjust for any time-varying confounders (Table 4.3). However, it's much smaller than the G-estimated effect.

With the stabilized weights we can also fit a Cox MSM. Instead of the Weibull model, we fit a Cox PH model using the weighted population. The smoking causal hazard ratio on time to the first CVD event estimated by this Cox MSM is $2.77\ (2.16, 3.55)$. The smoking causal effect detected by this Cox MSM is also larger than the one obtained from the time-dependent Cox model. However, the confidence interval was much wider than that for the Weibull MSM. The Weibull model fits the data better compared to the Cox PH model due to the violation of the PH assumption. Nonetheless, both MSMs explored here suggested that the time-varying confounding issues need to be addressed if one would like to detect a smoking causal effect on the time-to-first CVD event, using the CARDIA study data. Besides G-estimation, the MSM method also succeeded in removing this bias

to some extent.

4.4.4 GE-SCORE and GE-MIMIC

The first step in both GE-SCORE and GE-MIMIC is fitting an exposure model. This has been done previously when fitting inverse treatment probability weights required by the MSM. The many-to-many exposure RNN achieved an accuracy of 90.05% for the smoking status prediction using CARDIA data. GE-SCORE will further calculate the residuals from the exposure model while GE-MIMIC will start to generate null samples based on a Bernoulli distribution. The counterfactual outcome network we employed for both algorithms considered six potential effect modifiers: one-step leaped smoking history (i.e., smoking status at the previous visit), sex (male vs female), alcohol drinking status (drinker vs non-drinker), smoking intensity (number of cigarettes smoked per day), BMI and SBP. As suggested in Section 3.9, we employed the multi-start technique and the final estimation is the average of the three best runs. Last, because from simulation results, LSTM works better than RNN under more complex cases, these two algorithms used the LSTM unit to fit the CARDIA data. As in the simulation, data was split into training (70%), validating (20%) and the prediction is made with respect to 10% test data.

In Table 4.4, we compared the acceleration factors estimated from the baseline Weibull model, time-dependent Weibull, Weibull MSM, G-estimation, GE-SCORE, and GE-MIMIC. As the Cox PH model has been criticized for lacking causal meanings for past decades (Aalen et al., 2015; Hernán, 2010), the acceleration rate is preferred over the hazard ratio in terms of causal interpretation as we repeatedly emphasized in previous

sections. We still compared hazard ratios estimated by different models in Section 4.4.2 and 4.4.3. Yet, the transformation from the acceleration rate to hazard ratio is not only unnecessary but also always introduces another layer of bias in G-estimation, due to the fact that we further need to assume the counterfactual unexposed survival time is Weibull distributed. Note, G-estimation and the two proposed algorithms provided acceleration rates in an opposite direction compared to the regular AFT models (i.e., estimation with a positive sign means a harmful effect in G-estimation, GE-SCORE, and GE-MIMIC), so we have manually unified the signs of estimation for comparison purposes. For the two proposed algorithms, we obtained individualized acceleration factors thus we reported estimation mean and range instead of point estimation and associated confidence intervals in Table 4.4. Technically, results from the proposed algorithms came from more complicated models, which considered several potential effect modifiers. The estimations were individualized after taking distinct characteristics and the time factor into account.

From the table, rows 1-2 are traditional Weibull survival models without considering the time-varying confounding bias. These models identified the significant negative effects of smoking on time-to-first CVD events. Row 3-4 are models that have corrected for the time-varying confounding issue. G-estimation and MSM Weibull are only equivalent under restrictive parametric assumptions (Young et al., 2009). More generally, models from rows 1 - 3 assumed the data underlying survival distribution is Weibull while models from rows 4 and below are semiparametric AFT models which only assumed the acceleration nature but not the baseline survival distribution.

One of the significant benefits of GE-SCORE and GE-MIMIC is their ability to provide individual predictions. Furthermore, they can provide insights into heterogeneity

Models	ψ Estimation	Confidence Interval
Baseline Weibull	-0.34	(-0.44, -0.24)
Time-dependent Weibull	-0.45	(-0.56, -0.34)
Weibull MSM	-0.42	(-0.50, -0.33)
G-estimation	-0.61	(-0.72, -0.52)
GE-SCORE*		
Overall	-0.47	(-0.49, -0.43)
Year 00	-0.48	(-0.49, -0.48)
Year 02	-0.47	(-0.49, -0.45)
Year 05	-0.47	(-0.49, -0.44)
Year 07	-0.47	(-0.49, -0.44)
Year 10	-0.47	(-0.49, -0.44)
Year 15	-0.47	(-0.49, -0.43)
Year 20	-0.47	(-0.49, -0.43)
Year 25	-0.47	(-0.49, -0.43)
Year 30	-0.47	(-0.49, -0.43)
GE-MIMIC*		
Overall	-0.53	(-0.58, -0.49)
Year 00	-0.54	(-0.58, -0.52)
Year 02	-0.53	(-0.58, -0.50)
Year 05	-0.53	(-0.58, -0.50)
Year 07	-0.53	(-0.58, -0.49)
Year 10	-0.53	(-0.58, -0.49)
Year 15	-0.53	(-0.58, -0.49)
Year 20	-0.53	(-0.58, -0.49)
Year 25	-0.53	(-0.58, -0.49)
Year 30	-0.53	(-0.58, -0.49)

*The estimation mean and range were reported.

Table 4.4: Estimated acceleration factor of smoking on time to first CVD event, CARDIA.

among subgroups. Table 4.4 showed that smoking effects do not vary significantly across follow-up years (aging) although the baseline smoking effect seems the most significant. Overall, GE-MIMIC provides a more significant effect estimation along with a slightly wider range. Note that the range of estimation is directly related to the penalty term we add for the variation of $\hat{\psi}_k$. We have explored that, if $\gamma_2 = 0$ in Formulas (3.25) and (3.26), the range of estimation would be expanded substantially.

Now we take a closer look at the subgroup smoking effect. Unlike MSM and 1-

dimensional G-estimation, which provided a constant estimation for all participants, GE-SCORE and GE-MIMIC provided a more comprehensive estimation. Though we have considered six causal effect modifiers for both algorithms, we focus on assessing the group-wise difference for categorical variables: gender, drinker status, and smoking status at the previous visit. To facilitate our understandings of predictions disparity, we first provided a tabulation (Table 4.5) of these characteristics, by visit year.

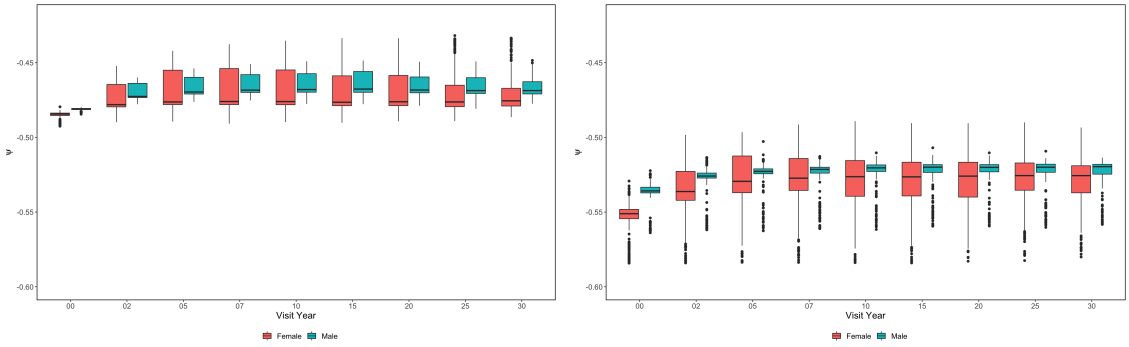
Visit Year	Smoking Status n (%)	Gender		Drinker		Previous Smoker	
		Male	Female	No	Yes	No	Yes
00 (n=5111)	Smoker 1564 (30.6%)	698	818	103	1461	1564	NA
	Non-smoker 3547 (69.4%)	1487	1967	591	2956	3547	NA
	Smoker Proportion	32.1%	29.4%	14.8%	33.1%	30.6%	NA
02 (n=4987)	Smoker 1519 (30.5%)	733	786	116	1403	217	1302
	Non-smoker 3468 (69.5%)	1535	1933	625	2843	3262	206
	Smoker Proportion	32.3%	28.9%	15.7%	33.0%	6.2%	86.3%
05 (n=4828)	Smoker 1436 (29.7%)	698	738	136	1300	178	1258
	Non-smoker 3392 (70.3%)	1487	1905	661	2731	3188	204
	Smoker Proportion	31.9%	27.9%	17.1%	32.2%	5.3%	86.0%
07 (n=4716)	Smoker 1347 (28.6%)	635	694	146	1201	117	1230
	Non-smoker 3369 (71.4%)	1476	1893	700	2669	3209	160
	Smoker Proportion	30.1%	26.8%	17.3%	31.0%	3.5%	88.5%
10 (n=4576)	Smoker 1243 (27.2%)	615	628	152	1091	118	1125
	Non-smoker 3333 (72.8%)	1434	1899	757	2576	3160	173

	Smoker Proportion	30.1%	26.8%	16.7%	29.8%	3.6%	86.7%
15 (n=4355)	Smoker 1078 (24.8%)	530	548	141	937	115	963
	Non-smoker 3277 (75.2%)	1408	1869	742	2535	3080	197
	Smoker Proportion	27.3%	22.7%	16.0%	27.0%	3.6%	83.0%
20 (n=4093)	Smoker 921 (22.5%)	451	470	142	779	88	833
	Non-smoker 3172 (77.5%)	1345	1827	710	2462	3022	150
	Smoker Proportion	25.1%	20.5%	16.7%	24.0%	2.8%	84.7%
25 (n=3756)	Smoker 708 (18.8%)	341	367	126	582	70	638
	Non-smoker 3048 (81.2%)	1277	1771	674	2374	2876	172
	Smoker Proportion	21.1%	17.2%	15.6%	19.7%	2.4%	78.8%
30 (n=3249)	Smoker 475 (14.6%)	228	247	93	382	58	417
	Non-smoker 2774 (85.4%)	1145	1629	626	2148	2629	145
	Smoker Proportion	16.6%	13.2%	12.9%	15.1%	2.2%	74.2%

Table 4.5: Subgroup characteristics distribution for smokers and non-smokers, CARDIA.

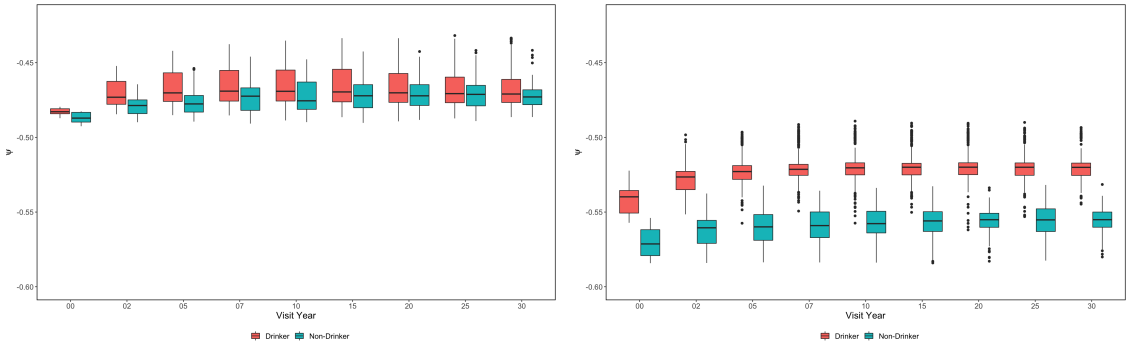
Figures 4.9, 4.10, 4.11 display the estimated smoking effects at each visit year, by gender, drinker status, and smoking status at the previous visit, respectively. The box plots allow us to compare the distributions of the estimated $\hat{\psi}$, across different subgroups.

Interestingly, we observed that smoking has a more substantial harmful effect on females compared with males. The gender disparity of the smoking effect has been noticed



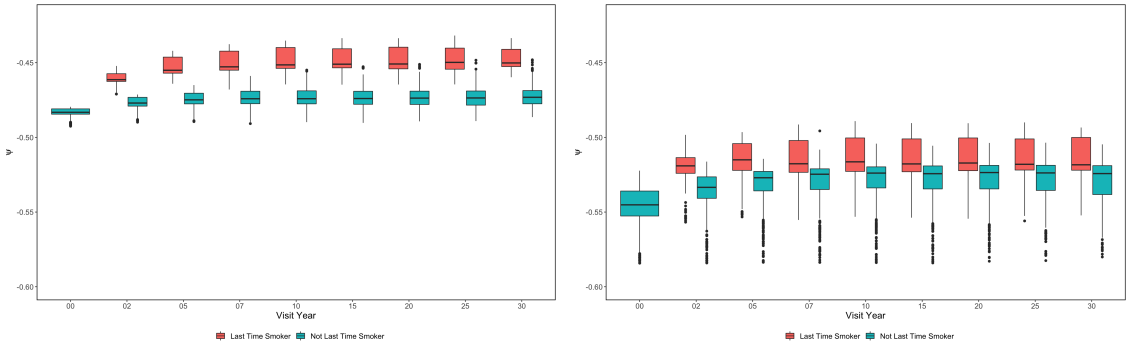
(a) GE-SCORE (b) GE-MIMIC

Figure 4.9: Algorithm estimated $\hat{\psi}_k$ by Gender Group vs the Visit Year, CARDIA



(a) GE-SCORE (b) GE-MIMIC

Figure 4.10: Algorithm estimated $\hat{\psi}_k$ by Drinker Status vs the Visit Year, CARDIA



(a) GE-SCORE (b) GE-MIMIC

Figure 4.11: Algorithm estimated $\hat{\psi}_k$ by Smoking Status at Previous Visit vs the Visit Year, CARDIA

by previous researchers (Hackshaw et al., 2018). Also, the smoking causal effect has a larger variation for females. A smaller proportion of smokers and a smaller proportion of subjects who experience the event of interest in the female group possibly explained the wider spread of the estimation, compared to the male group. Furthermore, the female group may have larger heterogeneity among some other important risk factors for CVD event, which can also increase the variation of the estimation. GE-SCORE didn't identify a significant difference of smoking causal effect between drinkers and non-drinkers. However, GE-MIMIC suggested that smoking does fewer negative effects on drinkers. One possible explanation is that in CARDIA population, proportion of drinker is much higher than non-drinkers. In this application, we simply classify subjects as a drinker and not a drinker without controlling for the drinking intensity. The causality between drinking and CVD events is still under investigation. When considering drinking intensity, a U-shaped drinking effect on CVD event is most plausible per current research (Pletcher et al., 2005). When considering the smoking status at the previous visit, a new smoker or a relapsed smoker suffers more from current smoking. The conclusion that can be delivered to the public is never start smoking, or having stopped, don't start again. Similar plots can be made for smoking intensity, BMI, and SBP if we categorize these continuous variables. Yet, the layout would mostly depend on the subjective categorization, thus those are omitted from presentation. There is also an overall time trend of the estimated effect, The baseline smoking effect seems more noticeable. The spread of the estimation is the smallest at baseline, due to the largest available data size among all visits. Furthermore, the proportion of smokers in this population is also decreasing steadily along the follow-up.

These two proposed algorithms provided conclusions consistent in direction, not only in the sign of the smoking effect (harmful effect) but also consistent in each subgroup. Though GE-MIMIC provided a more significant estimation and the associated range was wider. As with other ML algorithms, the estimations come with randomness. Taking an average of multiple runs can improve robustness as discussed in Section 3.9 but the estimation result may still change with a different initiation seed. For duplication purposes, a random seed was saved for both algorithms and we are only interpreting results generated with this seed. Different seeds have been explored, similar conclusions can be drawn, though the estimation may change. For a future application, the model fitting process can be repeated multiple times without specifying a seed.

4.5 Discussion

CARDIA data were used to quantify the smoking effect on time to the subject's first CVD event. The baseline model estimated a less significant smoking effect compared to time-dependent models, due to the long-term and short-term effects examined by these two sets of models (Xie et al., 2017). The time-dependent model seems very efficient but they still lack causal meanings due to the time-varying confounding issue. For baseline and time-dependent models, we considered the Cox PH and Weibull. Nonetheless, the Weibull model is a more rational choice for this research question because using CARDIA data, the estimated Weibull shape parameter is significantly greater than 1. This indicated that the hazard is not a constant but increases over time, thus suggesting a violation of the PH assumption.

When we moved to methods that correct for time-varying confounding, we considered a SNAFTM that assumes a constant smoking effect for all participants and MSMs with Weibull and Cox PH survival models. The G-estimation explored for CARDIA only searched on 1-dimensional space because of the computational burden and implementation inconvenience. From the simulation, we observed that the 1-dimensional G-estimation is coverage to the mean of the true effect if the exposure effect varies across groups. The stability of estimation is also prone to extreme values. MSM is also estimating a constant smoking effect for all individuals, but with assumptions of underlying survival distribution. In Section 4.4.3.1, we further added the Weibull distribution assumption to the counterfactual unexposed survival time simply for the purpose of comparing the hazard ratio among all methods. The causal rate ratios obtained from G-estimation and MSMs were comparable. Yet, the former estimated a much more significant smoking causal effect. One possible explanation is that G-estimation is a semiparametric method in the sense that we don't need to specify the distribution of the counterfactual unexposed time. The Weibull counterfactual unexposed survival time assumption may have been too strong. In conclusion, when compared to the traditional survival models used in Section 4.4.2, both SNAFTM and MSMs seem to correct the time-varying confounding bias to some extent.

The GE-SCORE and GE-MIMIC are still estimating the SNAFTM but with an enriched acceleration factor. These algorithms returned reasonable estimations, and they provided practical user profiles for the exposure causal effect of interest. The average smoking effects detected by GE-SCORE and GE-MIMIC lie between G-estimation and Weibull MSM. If high-dimensional G-estimation is used for the same task, at least the

blip function needs to include six main effects which have been included in the counterfactual outcome network of GE-SCORE and GE-MIMIC. Grid searching in this high-dimensional space is not feasible and we won't expect the result to be stable. From the algorithm estimation, some meaningful conclusions can be drawn. For instance, the algorithm suggested that females are more vulnerable to smoking compared to males, which coincides with previous research (Hackshaw et al., 2018). Plus, baseline smoking appears to have the most significant effect on time-to-first CVD events and new/relapsed smokers suffer more. Findings from the proposed algorithms provided important information about smoking as a risk factor for CVD events, and thus are valuable for public health. Provided personalized estimation, we expect the communication between health providers and patients to be more effective.

4.6 Limitations

A strong assumption when quantifying the exposure effects is that the exposure level (smoking or no-smoking) only changes at the measurement time after all time-dependent covariates have been measured. This is barely true in any observational studies. The fact is, subjects may start or stop smoking at any time during the study course, but the investigator is only updated with the smoking status during each clinical visit. Nevertheless, given the long follow-up time and subjects' adherence to smoking, we simply ignored the probability that subjects change their smoking status between two visits. Unlike a treatment to be assigned after related clinical measurements are collected, the mechanism of smoking initiation/cessation is complicated. Some important covariates

cannot be collected in a timely manner in an observational study (such as employment status), or they were not collected at all (e.g., the stress scores were not measured in CARDIA though we suspect a high stress level is an important initiative of smoking). The many-to-many RNN achieved more than 90% accuracy in predicting smoking status and this provides some confidence in the exposure model though. All the methods we used for this application involved IPCW so the quality of censoring weights is also important. In CARDIA, given that a small proportion of person-visits has a competing risk censored outcome (Table 4.2), the competing risk censoring mechanism may not be well addressed. Nonetheless, with the state-of-the-art RNN algorithm has been implemented for the exposure/competing risk prediction, we expect the information to be extracted to the maximal amount.

In CARDIA, the visit intervals are not equally spaced. Generally speaking, subjects were examined more often during the first 10-year of the study (four times during the first 10 years with intervals of 2, 3, 2, and 3 years), and later visits were scheduled every 5 years. When fitting the inverse probability weights, ignoring different visit intervals assumes that the effect of the covariates at visit 1 on visit 2 (2 years ago) is the same as the effect of the covariates at visit 2 on visit 3 (3 years ago), which may not be biologically reasonable. Considering the durations between visits can improve the calculation of the inverse probability weights. By ignoring visits at year 02 and year 07, we can manually create equally spaced data (5 years). But abandoning existing data is also a potential pitfall, especially since subjects' adherence is higher at the beginning of the study.

All predictors included in the competing risk censoring weights estimation, treatment weights estimation, in the traditional survival models, and in the GE-SCORE, GE-

MIMIC were all chosen empirically. Covariates collected at each visit also changed slightly over the study course. Though the proposed algorithms are able to handle high-dimensional data, it was not realistic nor necessary to clean and include all covariates collected in the study. For missing data, previous non-missing values are simply carried forward without any imputation. There are many other practical methods that deal with missing values besides the LOCF method but are not explored here. In future applications, different imputation methods may be more reasonable, based on the study settings. A minimal amount of influence of different imputation methods on the proposed algorithms is expected. We also recognized the fact that data collection errors are inevitable in any observational study. Self-reported data, such as smoking status and smoking intensity, are particularly prone to bias due to dishonesty or introspective inability. However, as we have pointed out, there is no way to conduct a clinical trial to investigate the smoking effect. It's also impossible to have investigators observe subjects in day-to-day life. How to best interpret observational data is a critical and challenging research problem. The estimation results could be improved with data of a higher quality.

4.7 Conclusion

We used CARDIA study data to quantify the causal effect of smoking on time to the participant's first CVD event. The causal effect measure we employed is consistent with an AFT model in which the effect is quantified as an acceleration factor of the event free time. To correct the time-varying confounding issue, we explored existing methods: SNAFTM with G-estimation and MSM with COX PH and Weibull survival models. The

proposed algorithms (GE-SCORE and GE-MIMIC) have also been implemented successfully. Compared to the previous G-estimation method for the SNAFTM, both GE-SCORE and GE-MIMIC provided meaningful and individualized estimations of the acceleration factor. Furthermore, the proposed algorithms delivered more insights into the heterogeneities of smoking's causal effect. For example, both algorithms detected a gender disparity in that smoking caused more deceleration of females' first CVD event free time, compared to males.

Chapter 5: Causal Inference in Threshold Regression

5.1 Introduction

This dissertation focuses on causal inference with survival outcomes. Previous sections (e.g., Section 2.6) briefly touched on the fact that the Cox PH model is not favored in causal survival analysis though it's the most prevalent survival model in medical research. The Cox PH model not only suffers from the difficulties of causal interpretation but also fails to achieve collapsibility, which will be defined in Section 5.2. Due to these limitations, other models that are more aligned with causal inference have been proposed. For example, an AFT model builds a linear link between the logarithm of the survival time and predictors by modeling the acceleration rate. The collapsibility of the AFT model is one of the major reasons that it is advocated for causal survival analysis. In previous chapters of this dissertation, we have explored SNAFTM, which roots in the AFT model, as a valid method to estimate exposure causal effect when any time-varying confounder is present. Specifically, we have shown in Chapter 3 that this model maintains its advantages with modern ML methods incorporated.

In addition, the First-Hitting-Time (FHT) based Threshold Regression (abbreviated as TR) (Lee and Whitmore, 2006, 2010) is a well-accepted method to analyze covariate effects on time-to-event endpoints while allowing for clinically meaningful interpreta-

tions. Unlike the first part of the dissertation, we do not consider longitudinal studies in this chapter. It will be shown later that, TR Models using a Wiener process with linear predictors are collapsible and hence lends the model naturally to causal survival inference. In particular, we would like to investigate the causal meaning of TR, as well as how to make a marginal causal conclusion from TR with the assistance of G-computation. The causal assumptions required for this Chapter have been identified in Section 2.2. This chapter is organized as follows: we first present the definitions of confounding and collapsibility. Then we introduce TR as a competitive survival model and provide some desirable properties that support its application in causal survival analysis. How to implement G-computation to estimate the marginal intervention causal effect has also been introduced. Furthermore, we consider a neural network expansion of the TR model: TRNN. Finally, simulations and a real application are presented.

5.2 Confounding and Collapsibility

5.2.1 Time Independent Confounding

The definition of time-independent confounding should be straightforward as the time-varying confounder concept presented in previous chapters has prepared readers for this. The research question is still the identification of the causal effect of the intervention on the time to an individual's first event of interest. In this chapter, we only consider studies where the intervention was introduced and covariates were measured at baseline, without additional information being gathered before the occurrence of the first event. Notations are mostly consistent with those specified in Section 2.4 except that all time-

dependent random variables were simplified to the time-independent version. The intervention is a dichotomous variable A that takes values 0 or 1. The outcome Y is recorded once either during the study (the event happened) or at the end of the study (right censored), and Z is a time-independent confounder (potentially). Note, in previous chapters, we used L to denote a confounder.

The previous chapters of this dissertation deal with the time-varying confounding issue (see Section 2.3 for definition) for longitudinal studies. In this chapter we consider cases with time-independent confounding measured only at baseline. Although less complicated, it remains critical to consider these confounders during study design as well as data analysis (VanderWeele, 2019). When a covariate Z has significant correlations with both the intervention A and the outcome Y, we say Z is a confounder for the causal effect of A on Y because Z is responsible for (at least part of) the appeared relationship between these two. However, the true causal effect of A on Y cannot be identified without a fair adjustment of the confounder Z. Another interesting complication of causal inference is "effect modification" which indicates that the effect of one variable on another varies across strata of a third (VanderWeele and Robins, 2007). Effect modification falls under the broader idea of "interaction" which is more familiar to statisticians. Traditional regression approaches are able to control confounding and deal with effect modification by producing conditional estimates, i.e., by adding potential confounders to the regression models. Similar conditional estimation methods are prevalent in regression-based survival data analysis.

The marginal effect provides a unified and simple way to quantify the causal relationship between the intervention and outcome. Thus, estimating the marginal inter-

vention causal effect is often of research interest (Snowden et al., 2011). G-computation (also called standardization) is an approach that allows ones to obtain marginal intervention effect estimation though the population contains different covariates groups and a set of confounders is measured at each observation (Sjölander, 2016). This method allows investigators to deduce the causality from observational studies as from perfectly randomized controlled trials (Robins, 1986b).

5.2.2 Collapsibility for TR with Linear Predictors

In this chapter, besides the confounding issue, we would also like to inspect collapsibility. It has been emphasized in the literature (Didelez and Stensrud, 2022; Greenland, 1996; Greenland and Pearl, 2011) that confounding and collapsibility are separate issues. Usually, the target population for a new drug includes patients with diverse prognostic baseline factors. Nevertheless, a randomized controlled trial can be used to estimate the intervention effect even if the primary analysis does not consider any baseline covariates as confounders because measured and unmeasured covariates will on average be balanced between intervention groups (U.S. Food and Drug Administration, 2021). Although randomization eliminates the bias introduced by confounding efficiently, collapsibility is a model property that cannot be guaranteed by study design.

Using the triplet $\{A, Z, Y\}$ defined above, we consider the case where Z is a categorical variable. A measure function of association between A and Y, $\psi(A, Y)$, is strictly collapsible across Z if it is constant across the strata (subtables) and this constant value equals the value obtained from the marginal table (Greenland et al., 1999). Generally

speaking, using $\psi_Z(A, Y|Z = z)$ to denote conditional association, the association measure between A and Y is collapsible across Z if ψ is a weighted average of ψ_Z for all $z \in \mathcal{Z}$ (Greenland and Pearl, 2011). In particular, one may expect the estimated causal effect of one covariate (e.g., A) on the outcome (e.g., Y) does not depend upon the inclusion (condition on) or exclusion (not condition on) of another covariate (e.g., Z), especially when these two covariates (A and Z) are uncorrelated. However, the marginal regression parameter for a covariate in the popular Cox PH model does not equal the conditional parameter when other covariates are included in the model (Burgess, 2015; Ford et al., 1995). The noncollapsiblity of the Cox PH model further complicates its application in causal reasoning, on top of the fact that HR has been criticized for decades for lacking causal meanings (Aalen et al., 2015; Hernán, 2010). Though collapsibility can be assured under the Aalen additive hazard model (Martinussen and Vansteelandt, 2013), the causal inference community has been considering using other non-hazard-related models due to the interpretability issue of HR. Not surprisingly, less model restrictive measures, such as differences between suitably adjusted and standardized survival curves or Restricted Mean Survival Time (RMST) (Didelez and Stensrud, 2022; Royston and Parmar, 2013; Uno et al., 2014), have drawn considerable attention. Some recent literature advocates the usage of RMST over a hazard ratio as the target causal measure (Stensrud et al., 2018).

5.3 Causal Inference of TR with Linear Predictors

The TR model conceptualizes a subject's latent health as a random process that drifts down towards a disease threshold (which is usually set at zero) (Lee and Whitmore,

2006, 2010). For a Wiener process where the health begins at $y_0 > 0$ and the boundary is the zero level, the distribution for the first-hitting-time (time-to-event) outcome T is given by an Inverse Gaussian (IG) distribution with p.d.f.

$$f(t|\mu, y_0) = \frac{y_0}{\sqrt{2\pi t^3}} \exp[-\frac{(y_0 + \mu t)^2}{2t}], \tag{5.1}$$

and survival function

$$S(t|\mu, y_0) = 1 - F(t|\mu, y_0) = \Phi\left(\frac{\mu t + y_0}{\sqrt{t}}\right) - \exp(-2y_0\mu)\Phi\left(\frac{\mu t - y_0}{\sqrt{t}}\right), \tag{5.2}$$

where $y_0 > 0, -\infty < \mu < \infty$ and Φ denotes the c.d.f. of the standard normal distribution. In most situations, one expects the deterioration rate μ to be negative so that the process will hit a disease threshold. If the deterioration rate is positive such that $\mu > 0$, the Wiener process may not hit the boundary threshold and the first-hitting-time T is not certain to be finite. This case can be used to model applications with cure-rate where the corresponding p.d.f. is improper with $P(T = \infty) = 1 - \exp(-2y_0\mu)$.

Two parameters can be estimated from the underlying latent health process, namely, the initial health status parameter y_0, which describes the process starting point, and the deterioration rate parameter μ, which describes the rate at which the process drifts towards the threshold. To begin with, we considered TR with linear predictors in Sections 5.2 to 5.5. Nonlinear predictors are considered in Section 5.6 using machine learning methods. These two parameters can be linked with observed regression covariate vectors

$(1, x_1, x_2, \cdots)$ and $(1, z_1, z_2, \cdots)$. For linear predictors:

$$\mu = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots = \boldsymbol{x}\boldsymbol{\alpha}, \tag{5.3}$$

$$\log(y_0) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots = \boldsymbol{z}\boldsymbol{\beta}, \tag{5.4}$$

Note, the initial health status y_0 is set to be larger than 0. Thus a logarithmic function is used to link the parameter y_0 with covariates in Formula (5.4). It's legitimate to use the same covariate vectors for both μ and y_0, or to consider completely different sets of risk factors for these two linear predictors. In other words, covariates in $\mathbf{x}$ and $\mathbf{z}$ can completely differ, coincide, or partially agree. The most plausible function format should come from expert knowledge of scientific research. We often want to consider the intervention A for μ. By adding covariate A to the Formula (5.3), the intervention effect is linked with the deterioration rate. The effect size is estimated by the coefficient of covariate A. Once the function formats are determined by the investigator, the maximum likelihood estimation method is used to estimate coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In most chronic diseases (such as CVD, or kidney disease) studies, the intervention aims at slowing the health deterioration rates. As a result, we focus on investigating the causal effect of the intervention on the deterioration rate μ in this chapter. A simple causal DAG (Figure 5.1) that includes a binary intervention A and a potential confounder Z demonstrates the causal problem under investigation. The corresponding linear prediction function in Formula (5.3) includes predictors A and Z: $\mu = \alpha_0 + \alpha_1 A + \alpha_2 Z$. The coefficient α_1 thus encodes the intervention effect on the deterioration rate μ. C is a

random variable denoting censoring. It is independent of other covariates shown in the DAG.

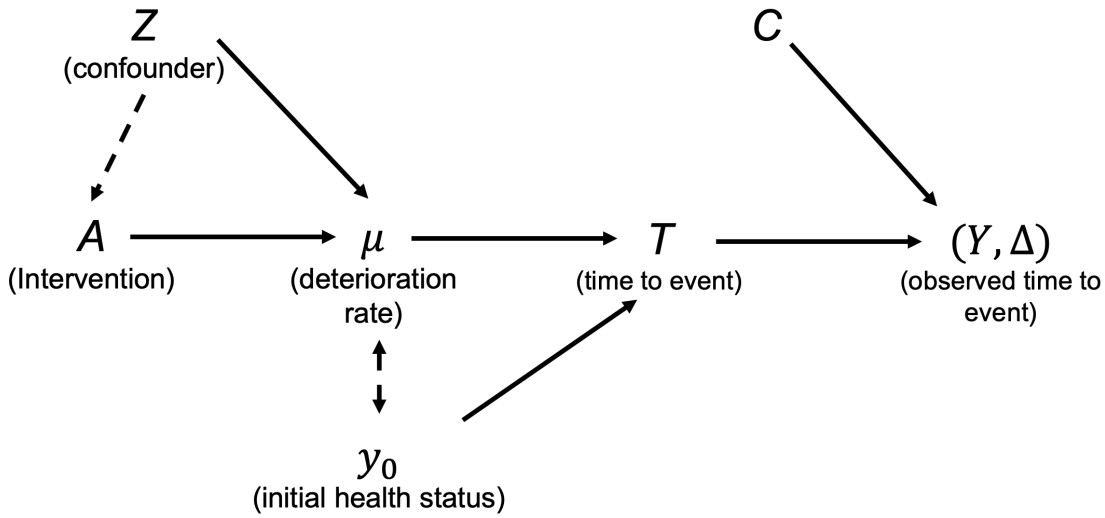


Figure 5.1: DAG of covariates effects on deterioration rate μ in TR.

Now, we follow the definition of marginal unadjusted effects and marginal (causal) effects in (Martinussen and Vansteelandt, 2013). Let ψ be a prediction function. First, we define $\psi(\mu|a) = E_Z(\mu|\text{do}(A = a), Z)$, as a prediction function of the deterioration rate μ. Simply speaking, the $do-$operator (Pearl, 2009b) sets a random variable, such as A, uniformly to a level, such as $A = 1$. The resultant prediction value $\psi = E_Z(\mu|\text{do}(A = a), Z)$ is the population value that would be realized if the random variable A is set to a specific level. The causal estimand of interest is the marginal causal effect $\mathcal{T}$:

$$\begin{aligned}\mathcal{T} &= E_Z(\mu|\text{do}(A = 1), Z) - E_Z(\mu|\text{do}(A = 0), Z) \\ &= \alpha_0 + \alpha_1 + \alpha_2 E(Z) - (\alpha_0 + \alpha_2 E(Z)) \\ &= \alpha_1\end{aligned} \tag{5.5}$$

The Formula (5.5) indicates the pleasant collapsibility of the deterioration rate in TR. The counterfactual deterioration rate is calculated by first setting A to a fixed (potentially contrary to fact) level and then taking the expectation over Z. The marginal intervention causal effect is the difference between two counterfactual outcomes and thus is given by α_1.

Next, the marginal unadjusted intervention effect $\mathcal{T}_m$ of μ can be defined based on a new prediction function $\psi(\mu|a) = E_Z(\mu|A = a, Z)$ without the $do-$operator used in Formula (5.5). It can be calculated by estimating the marginal association using observed data:

$$\begin{aligned} \mathcal{T}_m &= E_Z(\mu|A = 1, Z) - E_Z(\mu|A = 0, Z) \\ &= \alpha_1 + \alpha_2(E_Z(Z|A = 1) - E_Z(Z|A = 0)) \end{aligned} \tag{5.6}$$

which will be a consistent estimator of the marginal causal effect $\mathcal{T}$ if $\alpha_2 = 0$ (i.e., Z is not a significant predictor for μ, given A) or $E_Z(Z|A = 1) - E_Z(Z|A = 0) = 0$ (i.e., A and Z are independent).

Since the MLEs of coefficients of μ and y_0 are dynamically correlated, a valid causal inference of the deterioration rate in TR further requires that the initial health status y_0 is correctly adjusted. Figure 5.1 includes a dashed arrow between μ and y_0. This dashed arrow does not indicate a causal pathway between these two components. Instead, we can view y_0 as a confounder for μ and T. To estimate the causal effect of A on μ and subsequently on T (which is the only causal pathway we are interested in), the backdoor path $\mu \leftarrow y_0 \rightarrow T$ needs to be blocked. This can be ensured either

by: 1) carefully recruiting subjects so that the participants do not dramatically vary in terms of the initial health status, or 2) correctly adjusting the initial health status during analysis. The former approach casts significant challenges because the initial health status is a latent variable that cannot be observed. The assumption may be true approximately if investigators carefully recruit subjects with similar demographics. Consequently, the generalization of the study conclusion becomes unfeasible. However, the complete DAG of a complex disease process is not always available (VanderWeele, 2019). Thus including the correct covariates is not an easy task. To select the correct set of covariates to be adjusted (Vansteelandt et al., 2012), we also need to consider the clinical meaning of each parameter in TR. For example, we recognized that a post-baseline effect should not be adjusted in the initial health status y_0. Nonetheless, including unrelated predictors for μ won't influence the consistency and efficiency of the causal estimator due to the collapsibility. As a result, it is encouraged to consider comprehensive predictor sets, for both parameters.

In particular, the coefficient estimation does not systematically vary with the exclusion of other uncorrelated variables in TR (Hellier et al., 2020), due to its collapsibility. On the other hand, the Cox model does not have this benefit, i.e., the coefficient estimation before the intervention would be biased for the causal effects even when an uncorrelated covariate is omitted from the model. This has been well elaborated in (Crowther et al., 2022; Martinussen and Vansteelandt, 2013; Sjölander, 2016). Further, the interpretation of coefficients in the μ model is completely free from the hazard concept. We would interpret the intervention causal effect as saying the intervention causes the degradation rate to increase or decrease. With the same initial health level, a large degradation rate

means an earlier first-hitting-time, hence a shorter survival time.

Besides focusing on interpreting the causal meaning of μ, estimating other causal measures such as Restricted Mean Survival Time (RMST) can also be done in TR. For a fixed time horizon τ, RMST is defined with respect to the survival function. It's the area under the survival curves with a prespecified restrictive follow-up time. We have also derived TR RMST formulas in previous publications (Lee et al., 2022). In general, the RMST for TR is calculated by

$$\text{RMST}(\tau) = \int_0^\tau S(t|\mu, y_0)dt. \tag{5.7}$$

Theoretically, there are three formulas for TR RMST, depending on the sign of deterioration rate. When $\mu < 0$, a convenient and exact computational formula for evaluating RMST(τ) is given by:

$$\text{RMST}(\tau) = \tau S(\tau|\mu, y_0) + (y_0/|\mu|)S(1/\tau| - y_0, -\mu). \tag{5.8}$$

Two other formulas when $\mu = 0$ and $\mu > 0$ can be easily obtained by variable transformation (Lee et al., 2022).

5.4 Simulation I - Collapsibility for TR with Linear Predictors

Now, we will use simulations to demonstrate the collapsibility of the TR model with linear predictors. To investigate whether an intervention postpones an individual's time to the first CVD event by slowing the deterioration rate, studies can be conducted to examine

the relationship among intervention, potential confounder, and time-to-event outcome. In this simulation, we considered hypothesized studies with sample size $N = 1000$. We simulated the dichotomous intervention A and two potential confounders Z_1, Z_2. The intervention A was generated from a Bernoulli distribution, with a probability of receiving the intervention as 0.5. Confounder Z_1 is a binary variable generated from a Bernoulli distribution, with $P(Z_1 = 1) = 0.5$. Some baseline characteristics such as gender, follow this distribution. Confounder Z_2 is a continuous variable, generated from a Normal distribution, with mean = 1 and standard deviation = 1. Following this setting, A and Z_1, Z_2 are independent. The time-to-event outcome T is simulated by an Inverse Gaussian (IG) distribution for TR as shown in (5.1) with μ and $\log y_0$ given later. The study is scheduled to end in 3 years. Thus, the endpoint is potentially right-censored due to the limited follow-up length. The observed time variable $Y = \min(T, C)$. We use $\Delta = \mathbf{1}\{T < C\}$ to denote whether an event is observed or not. Later, to make Z_2 a confounder for the intervention and outcome, we create an association between the intervention and confounder. This can be achieved by using $\text{logit}(P(A = 1)) = (\gamma Z_2)$ to replace the aforementioned Bernoulli distribution which generates the intervention. The coefficient γ thus measures the magnitude of positive/negative association between A and Z_2.

To start with, we consider Design 1 where the initial health status is the same for all individuals in the population. Intervention A and covariate Z_2 both have causal effects on the deterioration rate μ. The linear predictors of μ and $\log y_0$ are simulated as:

Design 1:

$$\begin{aligned} \mu &= -2 + 0.2 \times A + 0.2 \times Z_2, \\ \log y_0 &= 1. \end{aligned} \tag{5.9}$$

For a slightly more complicated design (Design 2), Z_1 is related to the initial health status but not to μ by construction. Examples of such predictors include the genotype of an individual. If Z_1 stands for gender as suggested earlier, the model is reasonable if the female has $Z_1 = 1$ and females have a certain genetic makeup that protects them from CVD events. In conclusion, the simulation for deterioration rate remains unchanged whereas Z_1 is added to the generation of $\log y_0$:

Design 2:

$$\begin{aligned} \mu &= -2 + 0.2 \times A + 0.2 \times Z_2, \\ \log y_0 &= 1 + Z_1. \end{aligned} \tag{5.10}$$

As a result, the intervention causal effect is 0.2 for both simulated studies. The effect can be interpreted as the intervention slowing the deterioration rate by 0.2. Now, we consider the fitting of μ and $\log y_0$ separately. For the deterioration rate, we consider a reduced model that includes A only versus a correct model that includes both A and Z_2. Then for the initial health status, two models for $\log y_0$ were considered. One does not adjust for Z_1 while the other does. In conclusion, for each design, with a specific γ, four TR models were explored. The simulation results of the intervention effect are summarized in Table 5.1, see the first six rows for Design 1 results and rows 7 - 12 for Design 2 results. The

first two columns present results of the μ model with A only, and the next two columns present results from the μ model which includes both A and Z_2. The different $\log y_0$ models are summarized by rows, where "1" indicates the $\log y_0$ model with the intercept only and "1,Z_1" indicates the $\log y_0$ model with intercept and Z_1. The standard error was computed based on 1000 repetitions.

			μ Model			
			(A)		(A, Z_2)	
Simulation Design	γ	$\log y_0$ Model	Estimates	s.d.	Estimates	s.d.
Design 1	0	1	0.196	0.050	0.200	0.049
		1,Z_1	0.196	0.050	0.200	0.049
	0.2	1	0.238	0.049	0.202	0.048
		1,Z_1	0.238	0.049	0.202	0.048
	-0.2	1	0.157	0.049	0.199	0.049
		1,Z_1	0.157	0.049	0.199	0.049
Design 2	0	1	0.087	0.050	0.087	0.049
		1,Z_1	0.193	0.054	0.201	0.055
	0.2	1	0.118	0.029	0.098	0.029
		1,Z_1	0.231	0.036	0.202	0.036
	-0.2	1	0.076	0.029	0.095	0.029
		1,Z_1	0.152	0.036	0.200	0.036

Table 5.1: TR model estimated intervention effect under Design 1 and 2. For μ model consider including Z_2 and A or A only, for y_0 model consider including Z_1 and intercept or intercept only.

The simulation results of Design 1 first demonstrated that, when both the intervention and the potential confounder are included in the deterioration model, the coefficient estimations are unbiased. Whether or not Z_2 is included in the $\log y_0$ component the estimation is not affected because the true initial health status is a constant. Now if the deterioration model only includes the intervention A in Design 1, the relationship between the intervention and the potential confounder Z_2 needs to be further examined. Note that Z_2 is not a confounder for the intervention and outcome if it's uncorrelated with A (i.e.

$\gamma = 0$). By examining the first two rows of the table, the intervention causal effect is estimated consistently by the model, whether omitting Z_2 or not. This is due to the independence between A and Z_2 which further indicates that $E_Z(Z|A = 1) = E_Z(Z|A = 0)$. As a result, $\mathcal{T} = \mathcal{T}_m$ and the estimated coefficient for A provides the causal meaning. The Cox model would fail to provide unbiased estimation when the uncorrelated covariate is omitted. The covariate Z_2 has a beneficial effect on outcome by setting, same as the intervention. Thus, if covariate Z_2 is positively associated with the intervention ($\gamma = 0.2$), exclusion of this confounder enlarges the intervention effect as the confounder beneficial effect has to be absorbed by the solely included intervention. If the association between the confounder and intervention is in a reverse direction ($\gamma = -0.2$), the intervention effect is attenuated to account for the exclusion. For Design 2, the previous discoveries remain true, but now we have to correctly account for Z_1 for the $\log y_0$ component. All reduced models failed to provide unbiased intervention estimation (serious underestimation), regardless of the format of μ. If both intervention and potential confounder are included in the deterioration rate model, we still obtain an unbiased estimation of the intervention effect. The unbiasedness is preserved if Z_2 is not a confounder ($\gamma = 0$).

5.5 TR G-computation Procedure with Time-independent covariates

G-computation serves as a useful tool for the marginal causal effect estimation. It's straightforward to conduct G-computation for TR with linear predictors. The first step of a regular G-computation is to fit a regression model of the outcome on the exposure and relevant covariates. The model fitted in this step is frequently called the "Q-model"

(Snowden et al., 2011). In TR G-computation, the Q-model is nothing but a standard threshold regression of the time-to-event outcome with potential predictors such as A, Z. A correct Q-model is required for G-computation to estimate the intervention causal effect unbiasedly. To improve the chance of finding the correct model, the investigator is able to apply classical model selection tools (for example, AIC or BIC) in picking the most plausible TR model. Model selection techniques for TR were discussed previously (Saegusa et al., 2020).

Once estimated, the Q-model is used to predict all counterfactual outcomes for each subject under all possible intervention regimes. If the intervention is dichotomous, this is accomplished by plugging a 1 and then subsequently a 0 as the intervention into the TR fit to obtain predicted outcomes. Depending on the research question, the counterfactual outcomes μ^a or other estimands of interest can be calculated. Specifically, $\mu^{a=1}$ corresponds to the potential deterioration rate that should be observed under intervention and $\mu^{a=0}$ is the counterpart that would be observed without intervention. Predicting all potential outcomes for each subject provides researchers with a full data set that is free of confounding under causal assumptions. In other words, comparison/contraction among different potential outcomes now can be fulfilled at the individual or population level.

Generally speaking, G-computation can be used to generate the full data under all possible intervention levels if an intervention with discrete levels is under investigation. If the research question is straightforward as to identify the intervention's causal effect on the deterioration rate, we can subsequently merely take the difference between $\mu^{a=1}$ and $\mu^{a=0}$. Then averaging the differences across the observed distribution of the confounders leads us to the marginal causal effect of the intervention on the deterioration rate ($\mathcal{T}_\mu$). For

instance, the marginal intervention causal effect on μ for a population with size N can be written as:

$$\begin{aligned}\mathcal{T}_{\mu} &= E_Z(\mu|\text{do}(A=1), Z) - E_Z(\mu|\text{do}(A=0), Z) \\ &= \frac{1}{N}\sum_{i=1}^{N}\mu_i^{a=1} - \frac{1}{N}\sum_{i=1}^{N}\mu_i^{a=0}.\end{aligned} \tag{5.11}$$

Alternatively, a Marginal Structural Model (MSM) (Robins et al., 2000) can be used to fit the counterfactual outcomes. The estimates from MSM are also equipped with causal meanings under certain regularization conditions. Under the binary intervention case, these two approaches yield the same estimation. Nevertheless, if the researcher plans to use a more clinical-related measure such as RMST, estimated μ and $\log y_0$ can be plugged into Formula (5.7) to obtain the counterfactual individual RMST. The marginal intervention causal effect on RMST with a specific τ can be calculated similarly to what has been implemented for the deterioration rate:

$$\begin{aligned}\mathcal{T}_{RMST} &= E_Z(\text{RMST}(\tau)|\text{do}(A=1), Z) - E_Z(\text{RMST}(\tau)|\text{do}(A=0), Z) \\ &= \frac{1}{N}\sum_{i=1}^{N}\text{RMST}(\tau)_i^{a=1} - \frac{1}{N}\sum_{i=1}^{N}\text{RMST}(\tau)_i^{a=0}.\end{aligned} \tag{5.12}$$

5.6 Machine Learning Assisted TR

TR models with linear predictors discussed in previous sections enjoy benefits such as easy interpretation and collapsibility. Nevertheless, the linear format of predictors is often criticized for the unrealistic linearity assumption. To mitigate the risk of model misspecification, one may propose adding all relevant covariates along with all possible

interactions and high-order terms into the model. Yet, a model with linear predictors is also likely to fail when the input dimension is high, especially when it is close to or even higher than the observation number (Johnstone and Titterington, 2009). A recent research work (Bin and Stikbakke, 2022) introduced a novel approach to fit the TR model with linear predictors in a high-dimensional framework. Using their suggested boosting algorithm, coefficients estimation can be obtained.

In causal survival analysis, one may not only have to deal with cases with high dimensional inputs but also have to consider possible non-linearity among predictors. To address these challenges, we propose to develop a neural network extension for the TR model, which is called TRNN. TRNN includes two separate neural networks to predict μ and y_0 in a TR model. Thus, the neural network inputs as well as the structures can be completely different for μ and y_0. These two neural networks are connected by a log-likelihood-loss for right censored data. Trainable parameters in these two networks are updated simultaneously while the model is trained with respect to the objective function. This is the first time that neural networks are employed to fit a TR. See Section 3.1 for background information on neural networks.

To start, the linear predictors (5.3) and (5.4) can be replaced by neural networks without a hidden layer:

$$\mu = \sigma_\mu(\boldsymbol{W}_\mu x + b_\mu), \tag{5.13}$$

and

$$y_0 = \sigma_y(\boldsymbol{W}_y z + b_y). \tag{5.14}$$

In above formulas, σ_μ and σ_y are activation functions. To echo the parameter range for

TR, we can choose an activation function with output on the real line (such as linear activation function) for μ and an activation function with a strictly positive output range (such as exponential activation function) for y_0. When the activation is an identity function and the logarithm, Formulas (5.13) and (5.14) simply reduce to (5.3) and (5.4), which have linear formats. More generally, hidden layers can be added between the input and output layer with user-specified neuron numbers. With increased complexity, neural networks can approximate any functions (Pinkus, 1999). The number of neurons in the output layer decides the output dimension. Regardless of the number of hidden layers or neurons in each layer, the number of neurons in the output layer for predicting μ and y_0 should always be one. As a result, each network outputs a one-dimensional value for μ or y_0 in TRNN. Formula (5.13) and (5.14) can be generalized to $\mu = f_\mu(\boldsymbol{x})$ and $y_0 = f_y(\boldsymbol{z})$, in which f_μ and f_y are some neural networks. An example of TRNN which takes a 3-dimensional input is demonstrated in Figure 5.2. This TRNN has one hidden layer with 5 units for both μ and y_0.

The log-likelihood function for TR with the right censored data is the sum of the log density or log survival functions:

$$\begin{aligned}
\log\{\mathcal{L}(\mu, y_0)\} &= \sum_{i=1}^{N} \log\{\mathcal{L}_i(\mu, y_0)\} \\
&= \sum_{i=1}^{N} \Delta_i f(t_i; \mu, y_0) + (1 - \Delta_i) S(t_i; \mu, y_0) \\
&\propto \sum_{i=1}^{N} \Delta_i \left(\log(y_0) - \frac{3}{2}\log(t) - \frac{(y_0 + \mu t)^2}{2t} \right) \\
&+ (1 - \Delta_i) \log\left(\Phi\left(\frac{\mu t + y_0}{\sqrt{t}} \right) - \exp(-2y_0\mu)\Phi\left(\frac{\mu t - y_0}{\sqrt{t}} \right) \right).
\end{aligned} \tag{5.15}$$

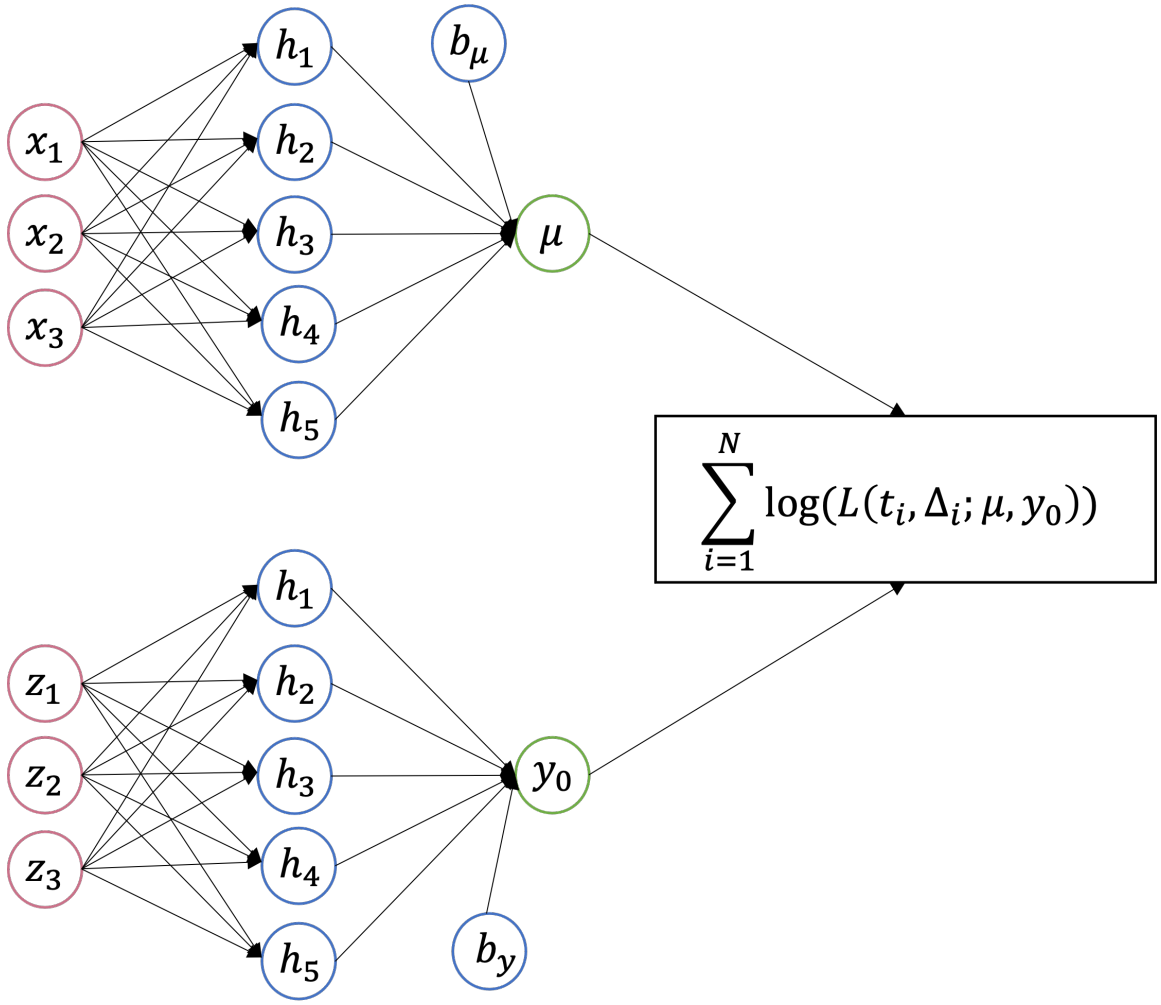


Figure 5.2: A TRNN with a 3 dimensional input and 5 units in the hidden layer for both μ and y_0.

The neural network solves the optimization problem by minimizing the loss. Consequently, the negative log-likelihood is the final loss function for TRNN. The MLE of Inverse Gaussian with censored observations is unique (Whitmore, 1983) and thus creates no challenge for the optimization problem. Optimization algorithms we introduced earlier in Section 3.1, such as Adam, can be used to train the model. To facilitate model convergence, we can first fit an Inverse Gaussian with intercept only (for both μ and y_0) using data, then use the estimated intercept terms as the initial bias for each network. Or, we

can simply set the initial bias for μ as -1 and for y_0 as 1. Usual neural network training techniques are all applicable to TRNN. For instance, the validation loss can be monitored during training and the algorithm will be stopped early if the validation loss decrements between iterations do not exceed a prespecified value (delta), this can effectively avoid over-fitting. See Section 5.8 for a real example.

The benefits of using TRNN are multi-fold, either when applying TR for a conventional survival analysis or a causal survival analysis. TRNN is capable of handling high-dimensional data such as gene expression. In Section 5.5, we have claimed that a correct Q-model is the prerequisite for the unbiased marginal effect estimation using G-computation. A TRNN can substitute a regular TR model for the Q-model for TR G-computation. Once the TRNN is optimized, it can be used to predict μ and y_0 under all possible regimes. Once the Q-model provides the complete data set with μ and y_0 under different regimes for each subject, the marginal causal effect can be evaluated as before, depending on the causal estimands of interest.

5.7 Simulation II - TR G-Computation

This section uses simulated data to demonstrate the application of G-computation with TR, either with the deterioration rate or RMST as the target causal estimand wherein the sample size $N = 1000$. The outcome of interest is still the time until the subject's first CVD event. The observed outcome is potentially right-censored at a prespecified study end time - 10 years. We further simulated two potential confounders. Similar to Simulation I, the binary variable gender (Z_1) was generated from a Bernoulli distribution, with

a probability of female being 0.5, and Z_2 is a continuous baseline variable that was generated from a Normal (1, 1). An example of such a predictor is population-standardized blood pressure. The covariates Z_1 and Z_2 are independent of each other. The binary intervention variable A was assigned by:

$$\text{logit}(P(A = 1|Z_1, Z_2)) = 0.5 - 0.2Z_1 + 0.3Z_2 \tag{5.16}$$

This assignment mechanism indicates that subjects with higher blood pressure and/or male are more likely to receive the intervention. This is just a sample intervention assignment mechanism. Nevertheless, the estimation procedure described in this section follows even if other parametrizations that create associations between covariates and intervention are considered. In observational studies, it's unreasonable to assume intervention is given at random. Instead, the intervention can be viewed as randomly assigned but the intervention probability depends on some baseline covariates. This echoes the conditional exchangeability assumption in Section 2.2.2. The time-to-event outcome T is simulated by an Inverse Gaussian (IG) distribution as shown in (5.1). Linear predictors of μ and $\log y_0$ are formulated as:

$$\begin{aligned} \mu &= -2 + 0.2 \times A + 0.2 \times A \times Z_2, \\ \log y_0 &= 1 + 0.5 \times A + Z_1. \end{aligned} \tag{5.17}$$

Due to a limited follow-up length, the observed time variable $Y = \min(T, C)$. We further use $\Delta = \mathbf{1}\{T < C\}$ to denote whether an event is observed or not. Figure 5.3 presents the DAG for the simulated study.

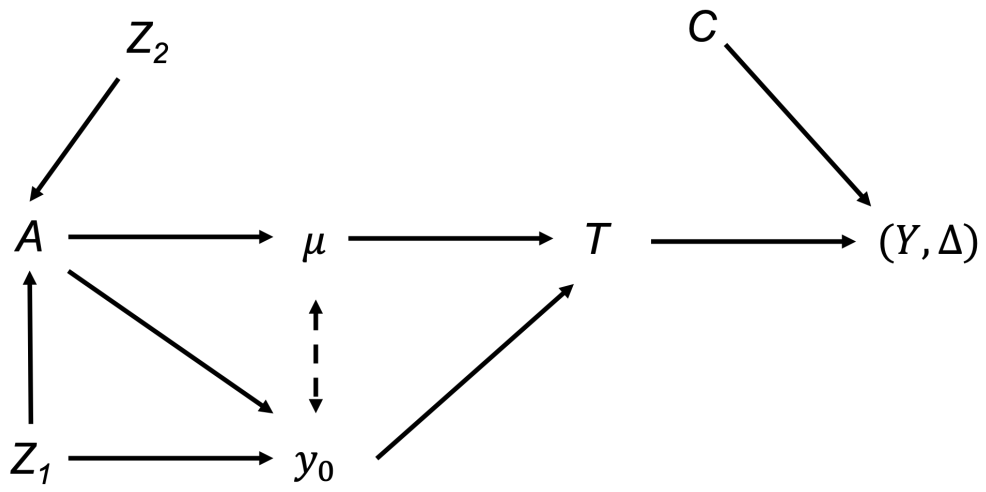


Figure 5.3: DAG of covariates effects on deterioration rate μ for the simulated study.

Under this simple simulation setting, two covariates (Z_1 and Z_2) must be considered when analyzing the effects of intervention A on the outcome Y. Gender (Z_1) affects both A and the initial health status (subsequently Y), hence it's a confounder for the intervention and outcome. Without any adjustments, one may incorrectly conclude that the intervention A has a causal effect on Y even when there is no such causal pathway from A to Y. The association appears only due to the unadjusted confounder Gender. In contrast, blood pressure (Z_2) affects the intervention, yet has no independent effect on the deterioration rate μ (subsequently Y). Thus, Z_2 is just an effect modifier for the intervention effect on μ, evident by the interaction term between Z_2 and A and lacking the causal pathway from Z_2 to μ. This data generation schema implies that the intervention has a positive effect on slowing the deterioration rate and the effect is further modified by Z_2 while the intervention benefit increases blood pressure increases. In terms of the initial health status, subjects who received the intervention and/or were female have larger initial health values.

We explored TR G-computation with four TR models (with different linear predictors for μ and/or $\log y_0$) as well as a TRNN as the Q-model. Model 1 assumes that the intervention influences both the initial health status and the deterioration rate. Thus one only adjusts the intervention in both μ and $\log y_0$.

Model 1 (linear predictors with intervention only):

$$\begin{aligned} \mu &= \alpha_0 + \alpha_1 A, \\ \log y_0 &= \beta_0 + \beta_1 A. \end{aligned} \tag{5.18}$$

Model 1 is incorrect for μ as it ignores the interaction effect and is also incorrect for $\log y_0$ by omitting the gender main effect. Model 2 is another incorrect model with a similar assumption for the covariate Z_2.

Model 2 (linear predictors with Z_2 only):

$$\begin{aligned} \mu &= \alpha_0 + \alpha_1 A + \alpha_2 Z_2, \\ \log y_0 &= \beta_0 + \beta_1 A + \beta_2 Z_2. \end{aligned} \tag{5.19}$$

This model includes the main effect of Z_2 when estimating μ but this causal pathway is lacking, Z_2 only influences μ by interacting with intervention A. Further, Z_2 wrongly replaces Z_1 for $\log y_0$. Thus, the model fails to model both the deterioration rate and initial health status flawlessly. Next, Model 3 is the correct model.

Model 3 (true simulated model as specified in Formula (5.17)):

$$\begin{aligned} \mu &= \alpha_0 + \alpha_1 A + \alpha_2 A \cdot Z_2, \\ \log y_0 &= \beta_0 + \beta_1 A + \beta_2 Z_1. \end{aligned} \tag{5.20}$$

Given the fact that picking the correct TR model may not be an easy task for the user, it is more realistic to consider a full model which includes all potential confounders as well as the interaction term.

Model 4 (full model with intervention, Z_1, Z_2, interaction):

$$\begin{aligned} \mu &= \alpha_0 + \alpha_1 A + \alpha_2 Z_1 + \alpha_3 Z_2 + \alpha_4 A \cdot Z_2, \\ \log y_0 &= \beta_0 + \beta_1 A + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 A \cdot Z_2. \end{aligned} \tag{5.21}$$

Table 5.2 summarizes the intervention effect estimated by TR, and the intervention causal effect estimation provided by the G-computation. The estimation standard deviation was calculated by nonparametric bootstrapping with 1000 times bootstrap samples.

RMST may be more interpretable for clinicians and patients. With the same simulated data, using $\tau = 10$, Table 5.3 summarized the RMST estimation results for each group and the group-wise difference. The estimations were obtained by using TR Model 1 - 4, TRNN (same parameter settings as used in previous G-computation), Kaplan Meier curves, and the Cox model. The Cox model used in the simulation is similar to the TR full model which includes A, Z_1, Z_2 and an interaction between intervention A and Z_2 as covariates. The baseline hazard was estimated by the Breslow method (Breslow, 1972).

The simulated true marginal effect on deterioration rate μ is $0.2 + 0.2 = 0.4$ per

	Model 1		Model 2		Model 3		Model 4	
	TR Parameter Estimates	Intervention Causal Effect Estimates	TR Parameter Estimates	Intervention Causal Effect Estimates	TR Parameter Estimates	Intervention Causal Effect Estimates	TR Parameter Estimates	Intervention Causal Effect Estimates
μ								
Int.	-0.917 (0.033)		-0.993 (0.048)		-2.011 (0.086)		-2.025 (0.198)	
A	0.273 (0.039)	0.273 (0.039)	0.264 (0.039)	0.264 (0.039)	0.203 (0.110)	0.405 (0.101)	0.211 (0.227)	0.412 (0.102)
Z_1							0.002 (0.085)	
Z_2			0.078 (0.035)				0.004 (0.177)	
$A \cdot Z_2$					0.202 (0.035)		0.201 (0.202)	
$\log y_0$								
Int.	0.871 (0.039)		0.865 (0.055)		1.004 (0.042)		1.007 (0.091)	
A	0.354 (0.045)	0.354 (0.045)	0.357 (0.045)	0.357 (0.045)	0.499 (0.049)	0.499 (0.049)	0.499 (0.109)	0.496 (0.050)
Z_1					1.000 (0.019)		1.000 (0.046)	
Z_2			0.007 (0.042)				0.001 (0.080)	
$A \cdot Z_2$							-0.003 (0.098)	

Table 5.2: Model estimation and G-computation results of marginal intervention causal effect by TR Models 1 - 4. TR Model 3 is the correct model, TR Model 4 is the full model.

Estimation Method	RMST Intervention (A) = 1 (s.d.)	RMST Intervention (A) = 0 (s.d.)	RMST Group Difference (s.d.)
TR Model 1	4.946 (0.093)	2.596 (0.074)	2.350 (0.120)
TR Model 2	4.932 (0.092)	2.611 (0.075)	2.321 (0.120)
TR Model 3 (True Model)	5.167 (0.092)	2.527 (0.054)	2.640 (0.078)
TR Model 4 (Full Model)	5.167 (0.092)	2.527 (0.054)	2.640 (0.078)
TRNN	5.156 (0.258)	2.576 (0.133)	2.580 (0.155)
KM	5.105 (0.106)	2.607 (0.076)	2.494 (0.132)
Cox PH (Full Model)	5.448 (0.148)	2.828 (0.110)	2.620 (0.111)

Table 5.3: Group specific and group-wise difference RMST estimations from TR Models 1 - 4, TRNN, KM, and Cox model. The RMST difference is calculated by the intervention group minus the control group.

Formula (5.17). Specifically, the true marginal intervention causal effect on the deterioration rate is the summation of the main effect and the interaction effect then averaged over the population distribution of the effect modifier. We observed that the intervention causal effects estimated by TR are the same as using TR with G-computation, for TR Models 1 and 2. This is because the TR with G-computation is estimating the same marginal effect as the TR models if no interaction term was captured in the predictor. Nonetheless, these TR models are biased due to model misspecifications. Both TR Model 1 and 2 have the intervention effect slightly overestimated because part of the intervention effect modified by BP is absorbed by the only included main intervention effect. As a result, the parameter estimations drift away from the truth. Though TR is collapsible, these TR models failed to provide consistent estimation as the excluded interaction term is correlated with the included main effect. Further, intervention A correlates with both Z_1 and Z_2 under this simulation setting. This emphasizes again the importance of confounder adjustment. Next, when the TR model was correctly specified (Model 3), TR with G-computation

returned a consistent estimation of the marginal intervention causal effect. This marginal causal effect has not only accounted for the main intervention effect but also the effect modified by Z_2 (BP). After adjusting for gender effect in the initial health level and the effect modifier (BP) for intervention in the deterioration rate, we were able to conclude that receiving the intervention will slow the deterioration rate by 0.4 marginally, for this population. The TR full model (Model 4) with G-computation also provided an unbiased estimation of the marginal intervention effect. Though the standard error for the TR model 4 estimations increased due to unnecessary terms have been added to the model, the marginal causal intervention effect estimation has very a similar standard error as the correct model.

Furthermore, as suggested in Section 5.6, we used a TRNN to fit μ and y_0. The algorithm took A, Z_1, Z_2 as input, for both μ and y_0 networks. We chose a network with one hidden layer (10 units, activation is `tanh`) for the μ network and a network without a hidden layer for the y_0 network. This TRNN can be conceptualized as TR **Model 5** in which there were no specific forms of μ and y_0 assumed. Each simulated data were first split into a training set (90%) and a testing set (10%). 25% of the training data were randomly selected as validation data and the model was trained with the rest training set. The optimization process would be potentially stopped early if the validation loss no longer decreases. Once the μ and y_0 networks were trained and we obtained the predictions of μ and y_0, the G-computation for the marginal intervention effect followed as demonstrated in Section 5.5. However, the estimation will only be evaluated using the test data. The marginal intervention causal effect on the deterioration rate we obtained by using TRNN as the Q-model is 0.358 (0.067), and the logarithm of initial health status is

0.488 (0.033). Therefore, the estimations were pretty accurate. Though only three main effects were fed into the model, using neural networks allows us to capture the nonlinear relationship among covariates. The standard deviations of marginal effect estimation were even smaller than those obtained from TR Model 3/4.

We would like to further quantify the marginal causal intervention effect in terms of RMST. Table 5.3 provides the RMST estimation results of TR Models 1-4, TRNN, KM, and the Cox Model, with G-computation. First, the results obtained by KM were from group-specific Kaplan-Meier curves nonparametrically. They serve as the true values and were the average RMST in each intervention group with diverse gender and BP values. The RMST difference was simply calculated as the difference of the area under group-specific survival curves. Next, for TR/Cox models, the group-specific estimations came directly from the estimated survival probabilities while the between-group difference requires an extra step of G-computation. The TR Models 3 and 4 provide nearly identical results for both group-specific estimations as well as the between-group difference. Moreover, these two TR models provided estimations that are close to the nonparametric estimation. Thus, these two models with G-computation both provided unbiased marginal intervention causal effect on RMST as the intervention effect modified by Z_2 has been correctly accounted for. Yet, TR Models 1 and 2 with G-computation fail this purpose as their underlying Q-models were incorrect. The intervention effect on RMST was underestimated. It can be seen from the TR Models 3 and 4 estimation results, the intervention was beneficial as the RMST increased by 2.640 years when the follow-up horizon was set to 10 years. In other words, the population-level RMST will be almost doubled if everyone switches to the intervention. Finally, TRNN outperforms all other TR models as it

provides the closest results to the nonparametric method, though the standard deviations of estimation increase significantly as well. The Cox model estimated the group-specific RMST inaccurately, however, the estimated RMST between-group difference is almost the same as TR estimated. Note that the standard error of Cox model estimation is much higher than that of TR models 3 and 4.

5.8 Application - CARDIA

The CARDIA study has been introduced in Section 4.2. In this section, we use the baseline measurement and the individual's first CVD event time measured until the end of the study to demonstrate the TRNN application and the following G-computation. We are interested in the marginal causal effect of smoking on the heart health deterioration rate μ. In total, there are 5,111 available complete baseline records. Following the discussion of CARDIA data in Chapter 4, we first use IPCW to adjust for the potential informative censoring process. A subject who is competing risk censored (death or lost-to-follow-up) will receive weights of 0, and the likelihood-based loss function used in TRNN has been updated accordingly to incorporate these weights. Figure 5.4 shows the Kaplan-Meier survival curves of CVD events of smokers vs. non-smokers. The outcome is the subject's time to his/her first CVD event. We truncate the follow-up time at month 400. As can be seen from the survival curves, there is an obvious trend of delayed exposure effect. The trend is expected as participants' likelihood of experiencing their first CVD event grows as they age.

To identify the optimal structure of the TRNN, we started with determining the

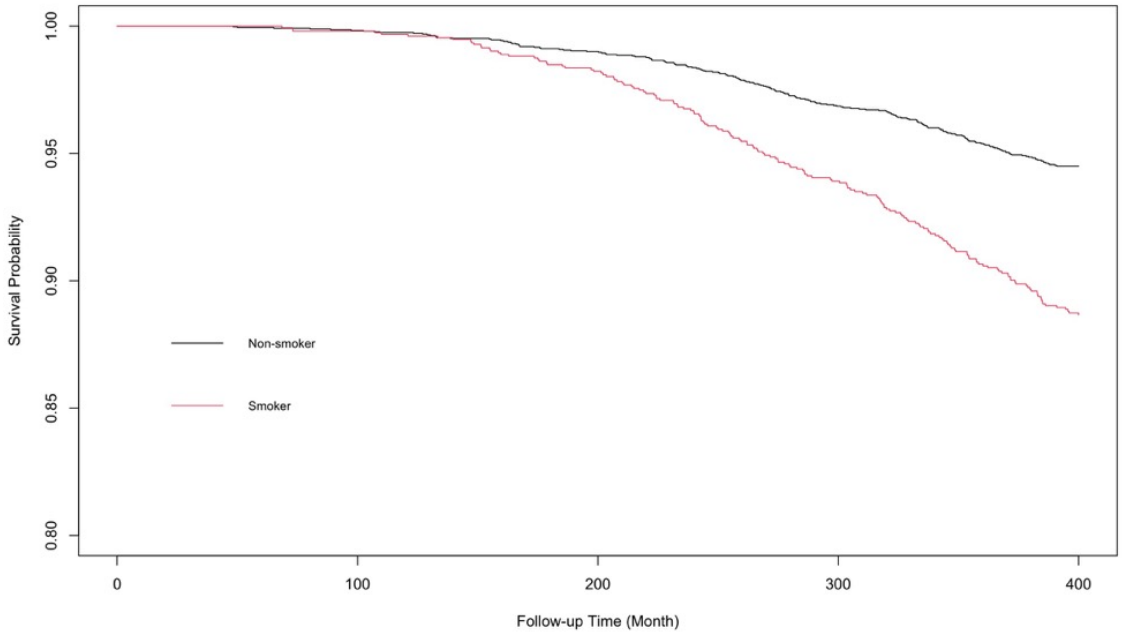


Figure 5.4: Kaplan-Meier Curves of Time-to-First CVD Event, Smokers (red) vs Non-smokers (black), CARDIA. Note the y-axis does not start from 0.

hyperparameters to be used in the network. Parameter tuning has been mentioned in Section 3.9. 5-fold cross-validation (Hastie et al., 2001) was used for the tuning process. From the simulation, we observed that one hidden layer is sufficient to capture the non-linear nature of the covariates. In this application, we would like to consider `sigmoid` and `tanh` activation function for the hidden layer. Additional hyperparameters we tuned included hidden unit number for μ network $(10, 15, 20)$, hidden unit number for y_0 network $(2, 5, 10)$, and learning rate $(0.002, 0.005)$ and delta for early stopping $(0.01, 0.02)$. There are many other hyperparameters such as batch size to be tuned. The truth is that it's impossible to tune all hyperparameters exhaustively as the set of hyperparameters must be large. Therefore, we only included a working example here with a limited set of hyperparameters. The choice of network structure and hyperparameter set are mostly

experience-based. For instance, in this dataset the event rate is low. We don't encourage using a small batch size - the algorithm may not perform well if a batch of data does not include a single event. Thus, we set the batch size to 256 and refrain from tuning this parameter. The number of hidden units also depends on the input dimension. For the μ network, we considered baseline covariates including the smoking status (smoker vs non-smoker), gender (female vs male), race (white vs non-white), age at enrollment (younger than 18, $19-25$, $25-29$ and older than 30), the marital status (married or living as married, divorced or separated or widowed, never married), the baseline BMI and the baseline SBP. For the y_0 network, we considered gender (female vs male), race (white vs non-white), age at enrollment (younger than 18, $19-25$, $25-29$, and older than 30), the marital status (married or living as married, divorced or separated or widowed, never married), the baseline BMI, the baseline SBP, the exercise intensity score, the working status (full-time/part-time/unemployed), the education level (less than high school or GED/some college or college/college above), self-reported diabetic, hypertension, high cholesterol, whether father/mother had a heart attack and whether father/mother had a stroke. Nevertheless, there are many more potentially important covariates that can be included and TRNN is capable of handling high-dimensional input. Using domain knowledge, we narrowed down the range of covariates as well as the hyperparameter set, for demonstration purposes. The 5-fold cross-validation selects units numbers 15 and 2, activation function `sigmoid` and `tanh` for μ and y_0 network, respectively. We set the learning rate to 0.005 and the delta for early stopping to 0.01.

With selected hyperparameters, the data were further split into the training set (70%), the validation set (20%), and the testing set (10%). Once the model is trained, the

G-computation is performed with 10% testing data that the model never used previously. Note without setting a random seed, the resultant neural networks parameter estimations are different due to the stochastic optimization. With 100 refits of the neural networks, smoking changes the deterioration rate by -0.024 $(-0.029, -0.019)$ in this population.

In terms of RMST ($\tau = 400$ months) of the first CVD event free time, the difference between smokers and non-smokers on average is -24.589 $(-29.418, -19.761)$ months. The confidence interval was computed using the bootstrapped standard deviation. The marginal smoking causal effect we detected using G-computation and TRNN is significant but not as dramatic as that we found in Chapter 4 using data from longitudinal visits. This is partially due to the fact that we only used records from the baseline as covariates. In other words, we didn't use the longitudinal data and assumed these characteristics stay the same from the enrollment to the end of the follow-up. As shown in previous chapters, a time-dependent model is more powerful than the baseline model in detecting the smoking effect. Though the difference is not huge per other methods (Table 4.3), our proposed algorithm GE-SCORE and GE-MIMIC detected that the baseline smoking effect is most significant in Section 4.4.4.

5.9 Discussion

The TR model provides profound clinical interpretations because the TR model uses two parameters, i.e., the initial health status y_0 and the deterioration rate μ, to describe the FHT of a latent health process. Predictors can be linked to these two parameters by using appropriate link functions. We argued that the TR model with linear predictors

for the deterioration rate μ is collapsible if y_0 is correctly controlled. Collapsibility is a desirable property for a causal model because one often feels uncomfortable interpreting the conditional effect estimation. Collapsibility at least ensures an unbiased estimation of the intervention's causal effect on the outcome, if an uncorrelated covariate is omitted from the model. On the other hand, the Cox model does not have such a property.

When analyzing the intervention effect, any proposed model will generally not be exactly correct, and results can be difficult to interpret if the model is misspecified and treatment effects substantially differ across subgroups (U.S. Food and Drug Administration, 2021). To make things more complicated, the subgroup information may not be collected at all. As a result, the marginal intervention effect is of research interest as it provides population summaries with a simple interpretation. A typical randomized controlled trial attempts to estimate the marginal effect of the investigated intervention, which is the intervention effect on the population level regardless of the diverse characteristics. Collapsibility is a particularly desirable model property because it provides a model with a much higher likelihood to estimate the marginal effect unbiasedly.

Furthermore, a collapsible model allows for other more advanced causal inference techniques such as the instrumental variable (IV) analysis (Tchetgen et al., 2015). It has been shown that TR offers an appealing opportunity to apply IV analysis under survival context (Hellier et al., 2020).

The causal aspect of TR is underrated with minimum availability of TR causal inference (Hellier et al., 2020). Section 5.3 and Simulation 5.4 have demonstrated the collapsibility of TR with linear predictors when one would like to make causal inferences for the deterioration rate and the initial health status is correctly controlled. Though we

focus on interpreting the deterioration rate, the $\log y_0$ part in TR allows one to control for the population baseline heterogeneity. Thus, we don't treat the necessity of controlling $\log y_0$ in TR deterioration rate analysis as a disadvantage. Previous research has shown that the intervention effect may interact with the baseline genotypes (Mok et al., 2009). This is a common cause for crossing survivals observed in medical research. Subjects in the intervention group seem to experience the event sooner than the control group in the early stage of follow-up due to subgroup initial health status heterogeneity. But, the intervention benefit emerges later because the intervention has significant effects on the deterioration rate. The Cox model would lose power significantly when the PH assumption is violated, not to mention the lack of collapsibility and interpretation issues. Alternatively, TR is capable of addressing distinct hazard types (Chen et al., 2022). As we have shown in Section 5.4, the genotype or other baseline characteristics can be adjusted as part of the initial health status. Once the baseline heterogeneity has been effectively controlled, the marginal intervention causal effect can be viewed as slowing the health process deterioration, providing a solid causal interpretation of the intervention on the deterioration rate μ.

Though the nonparametric marginal causal estimation is feasible under limited cases such as when all covariates are discrete and there are data collected for each subcategory, parametric models are more realistic choices in practice. One may have to seek help from the G-computation method when confounders for both the intervention and outcome or an effect modifier (interaction) are present. G-computation is a statistically reliable method of covariate adjustment for an unconditional treatment effect that produces a resulting estimator (U.S. Food and Drug Administration, 2021). We have demonstrated

that, with correct model specifications, TR is capable of estimating the intervention effect marginally besides the usual conditional estimation with G-computation. G-computation with TR can also provide a valid marginal causal estimation if other estimands such as RMST are chosen as the target causal estimand. Simulation 5.7 advocates using a comprehensive model as the Q-model to mitigate the risk of model misspecification. Plus, it's possible to derive a doubly robust estimation that combined the outcome model (TR in our case) and IPW. The combined strategies are generally superior to either strategy alone (Little and Rubin, 2000).

TRNN is the neural network expansion of the TR model in which the linear predictor structure is replaced by neural networks. TRNN is capable of capturing and analyzing more sophisticated relationships between covariates and deterioration rate and/or initial health status. From our observations, TRNN serves as a competitive Q-model for the G-computation. As we have shown with simulations in Tables 5.2 and 5.3, using TRNN can free practitioners from selecting the "true" model. Yet, as in other ML methods, the network structures and hyperparameters have to be tuned before implementation. With an increasing pool of ML toolkits, we only expect this method to become more accessible in the future. Last, the convergence of TRNN is fairly quick. Under the simulated setting, 100 training epochs are sufficient, thus no significant increased computation time was observed compared to a non-NN-based TR model.

As a flexible parametric survival model, surrogate estimands, such as RMST, can be readily obtained for TR. Using this estimand also solves the dilemma that an investigated intervention may provide reverse effects for the initial health status and the deterioration rate. Given a fixed end of follow-up horizon (τ), slowing the degradation rate and/or

improving the initial health status directly improves the survival rate, subsequently the RMST. We can also consider other causal estimands such as causal survival probability difference.

There exist other aspects of TR, which make it more valuable in practice. For instance, when $\mu > 0$, TR with a cure rate can handle the case where a substantial proportion of subjects are believed "cured" or immune from the endpoint events (Lee et al., 2022).

5.10 Limitations and Future Work

As mentioned in Section 3.12, the discovery of the correct DAG falls beyond the scope of this dissertation. Yet, the correctness of DAG is often the prerequisite of an unbiased G-computation and the prerequisite of any valid causal inference. Though G-computation enables one to draw marginal causal effect conclusions, the generalization of the results requires extra attention because the confounder distribution may vary significantly across multiple populations. Even within the same population, making inferences for unobserved subgroups requires extrapolations that are often dangerous. Applying G-computation to TR in this chapter requires the assumption that the underlying survival distribution is Inverse Gaussian. In practice, the assumption is not necessarily true. The advantage of the model is its rich parameterization that provides good opportunities for modeling the underlying health process. There are TR models using other processes in the literature, but it's outside the scope of this dissertation.

The G-computation for longitudinal studies with time-dependent covariates was

introduced in Section 2.5. We also have introduced the structural nested AFT model in Chapter 3. How to do a Structural Nested Threshold Regression Model (SNTRM) is an open question. In order to consider longitudinal TR with time-varying covariates, the deterioration rate at an earlier interval is a causal factor of initial health status at the beginning of the next interval. It thus requires a rethink of the model from the random process perspective.

5.11 Conclusion

Threshold Regression treats subjects' latent health status as a random process and models the process's first-hitting-time of a boundary. The TR model is straightforward to implement for time-to-event outcomes and provides clinically meaningful interpretations. There are two parameters in the model: the initial health status y_0 and the deterioration rate μ. Different causal questions can be posed for μ and/or y_0, depending on the domain knowledge. In this section, we focus on using TR to model the time to the occurrence of the subject's event of interest, without considering the longitudinal case. The research interest centers on whether an intervention causes a change in the deterioration rate when y_0 is correctly controlled. It is shown that a TR model with linear predictors for μ is collapsible if y_0 is correctly controlled. The collapsibility makes this model a useful tool in causal survival analysis. Furthermore, TR G-computation can provide the marginal causal effect estimation on the deterioration rate. The neural network expansion of TR - TRNN, not only provides a new working direction for TR in the ML era but also allows TR to model a wide variety of scenarios. Thus, it has the benefit to use TRNN as the

Q-model for the TR G-computation. In conclusion, TR has proved its effectiveness in causal survival analysis.

Chapter 6: Conclusion

Most questions in medical research are causal problems, not association problems. Survival (time-to-event) outcome is a commonly seen measure as it naturally reflects the disease process. Many interventions are developed to delay the event if the undesirable event cannot be avoided (such as death). Models that analyze time-to-event outcomes capture intervention effects on lifetimes. However, survival analysis has another barrier as the observed data suffer from censoring due to limited follow-up and/or other factors. Causal survival analysis has only begun shedding light recently on these problems firstly due to the existence of randomized controlled trials (RCTs) in which association models serve as causal models, and secondly due to the complexity of analyses. Nevertheless, causal survival analysis methods are necessary whenever an RCT is not applicable, or the trial fails to be compliant with its protocol. The non-compliance may be either due to uncontrollable post-randomization factors, or unexpected protocol deviations. Even in a perfectly planned RCT, participants may deviate from the protocol. Whenever the conduct of a clinical trial is in doubt, the estimated effect is highly likely biased from the desired causal meaning. Another complication of current medical research is the increasing amount of data which calls for methods that are still robust and applicable with high-dimensional input.

This dissertation focused on causal survival analysis when the outcome of interest is time-to-event. We first explored the common but complex time-varying confounding issue in longitudinal observational studies. The Structural Nested Accelerated Failure Time Model (SNAFTM) is an effective method to correct the time-vary confounding bias. This model quantifies the intervention causal effect by the acceleration factor. Two neural network-based algorithms (GE-SCORE and GE-MIMIC) have been proposed to estimate SNAFTM. From simulations and application, the two proposed algorithms are shown to be more flexible than the existing G-estimation method. A significant advantage of these proposed algorithms is that they don't require a pre-specified functional format. At the same time, they are able to identify potential effect modifiers, thus providing inference at individual levels.

In addition to SNAFTM investigated in Chapters 2 to 4, we also explored causal applications of the Threshold Regression (TR) model for the first hitting time in a Wiener process in Chapter 5. The exploration was limited to cases with time-independent confounding measured at baseline without additional information being gathered before the occurrence of the first event. This model conceptualizes two parameters of the subject's latent health process, namely, the initial health status y_0 and deterioration rate μ. It is shown that a TR model with linear predictors for μ is collapsible if y_0 is correctly controlled. Furthermore, the proposed neural network expansion of TR (TRNN) allows the model to handle high-dimensional input as well as complex nonlinear functions of predictors. TRNN appears to be both novel and promising for practical applications of TR. We expect more research and applications of this model in causal survival analysis in the future.

Bibliography

Aalen, O. O., Cook, R. J., and Røysland, K. (2015). Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, 21(4):579–593.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Anderson, K. M., Wilson, P. W., Odell, P. M., and Kannel, W. B. (1991). An updated coronary risk profile. a statement for health professionals. *Circulation*, 83(1):356–362.

Bellot, A. and van der Schaar, M. (2019). Conditional independence testing using generative adversarial networks.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. (2020). Estimating counterfactual treatment outcomes over time through adversarially balanced representations.

Bin, R. D. and Stikbakke, V. G. (2022). A boosting first-hitting-time model for survival analysis in high-dimensional settings. *Lifetime Data Analysis*.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Breslow, N. E. (1972). Discussion on professor cox's paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):202–220.

Burch, P. (1974). DEATH AND CORONARY ATTACKS AFTER GIVING UP CIGARETTE SMOKING. *The Lancet*, 304(7896):1564.

Burgess, S. (2015). Commentary. *Epidemiology*, 26(3):411–413.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Cauchy, A. (1847). Methode generale pour la resolution des systemes d'equations simultanees. *C.R. Acad. Sci. Paris*, 25:536–538.

Chen, Y., Lawrence, J., and Lee, M.-L. T. (2022). Group sequential design for randomized trials using "first hitting time" model. *Statistics in Medicine*, 41(13):2375–2402.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chollet, F., Allaire, J., et al. (2017). R interface to keras. `https://github.com/rstudio/keras`.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall/CRC.

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

Crowther, M. J., Royston, P., and Clements, M. (2022). A flexible parametric accelerated failure time model and the extension to time-dependent acceleration factors. *Biostatistics*.

Daniel, R., Cousens, S., Stavola, B. D., Kenward, M. G., and Sterne, J. A. C. (2012). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, 8(4).

Didelez, V. and Stensrud, M. J. (2022). On the logic of collapsibility for causal effect measures. *Biometrical Journal*, 64(2):235–242.

Ding, P. and Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33(2).

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Esteban, C., Staeck, O., Baier, S., Yang, Y., and Tresp, V. (2016). Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Ford, I., Norrie, J., and Ahmadi, S. (1995). Model inconsistency, illustrated by the cox proportional hazards model. *Statistics in Medicine*, 14(8):735–746.

Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jacobs, D. R., Liu, K., and Savage, P. J. (1988). Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41(11):1105–1116.

Gidding, S. S., Xie, X., Liu, K., Manolio, T., Flack, J. M., and Gardin, J. M. (1995). Cardiac function in smokers and nonsmokers: the cardia study. the coronary artery risk development in young adults study. *Journal of the American College of Cardiology*, 26(1):211–6.

Gill, R. D. and Robins, J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *The Annals of Statistics*, 29(6).

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, 7(5):498–501.

Greenland, S. and Pearl, J. (2011). Adjustments and their consequences-collapsibility analysis using graphical models. *International Statistical Review*, 79(3):401–426.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1).

Hackshaw, A., Morris, J. K., Boniface, S., Tang, J.-L., and Milenković, D. (2018). Low cigarette consumption and risk of coronary heart disease and stroke: meta-analysis of 141 cohort studies in 55 study reports. *BMJ*, 360.

Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York.

Heath, D. and Sudderth, W. (1976). De finetti's theorem on exchangeable variables. *The American Statistician*, 30(4):188–189.

Hellier, J., Emsley, R., and Pickles, A. (2020). Estimating dose-response for time to remission with instrumental variable adjustment: the obscuring effects of drug titration in genome based therapeutic drugs for depression trial (GENDEP): clinical trial data. *Trials*, 21(1).

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13–15.

Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M., and Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 14(7):477–491.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman and Hall/CRC.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366.

Howe, C. J., Cole, S. R., Chmiel, J. S., and Muñoz, A. (2011). Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *American Journal of Epidemiology*, 173(5):569–577.

Joffe, M. M., Yang, W. P., and Feldman, H. (2011). G-estimation and artificial censoring: Problems, challenges, and applications. *Biometrics*, 68(1):275–286.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Lachin, J. M. (2015). Fallacies of last observation carried forward analyses. *Clinical Trials*, 13(2):161–168.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, M.-L. T., Lawrence, J., Chen, Y., and Whitmore, G. A. (2022). Accounting for delayed entry into observational studies and clinical trials: length-biased sampling and restricted mean survival time. *Lifetime Data Analysis*.

Lee, M.-L. T. and Whitmore, G. A. (2006). Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statist. Sci.*, 21(4):501–513.

Lee, M.-L. T. and Whitmore, G. A. (2010). Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Analysis : An International Journal Devoted to Statistical Methods and Applications for Time-to-Event Data*, 16(2):196–214.

Li, R., Shahn, Z., Li, J., Lu, M., Chakraborty, P., Sow, D., Ghalwash, M., and wei H. Lehman, L. (2020). G-net: A deep learning approach to g-computation for counterfactual outcome prediction under dynamic treatment regimes.

Lim, B., Alaa, A., and van der Schaar, M. (2018). Forecasting treatment responses over time using recurrent marginal structural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Little, R. J. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, 21(1):121–145.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Liu, K., Daviglus, M. L., Loria, C. M., Colangelo, L. A., Spring, B., Moller, A. C., and Lloyd-Jones, D. M. (2012). Healthy lifestyle through young adulthood and the presence of low cardiovascular disease risk profile in middle age. *Circulation*, 125(8):996–1004.

Liu, Y. (2020). *Python Machine Learning By Example - Third Edition*. Packt Publishing, [S.l.].

Lok, J., Gill, R., van der Vaart, A., and Robins, J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica*, 58(3):271–295.

Loria, C. M., Liu, K., Lewis, C. E., Hulley, S. B., Sidney, S., Schreiner, P. J., Williams, O. D., Bild, D. E., and Detrano, R. (2007). Early adult risk factor levels and subsequent coronary artery calcification. *Journal of the American College of Cardiology*, 49(20):2013–2020.

Luehrs, R. E., Zhang, D., Pierce, G. L., Jacobs, D. R., Kalhan, R., and Whitaker, K. M. (2021). Cigarette smoking and longitudinal associations with blood pressure: The CARDIA study. *Journal of the American Heart Association*, 10(9).

Martí, R., Resende, M. G., and Ribeiro, C. C. (2013). Multi-start methods for combinatorial optimization. *European Journal of Operational Research*, 226(1):1–8.

Martinsson, E. (2016). Wtte-rnn: Weibull time to event recurrent neural network. Master's thesis, Chalmers University Of Technology.

Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in cox and aalen regression models. *Lifetime Data Analysis*, 19(3):279–296.

McCulloch, W. and Pitts, W. (2021). A logical calculus of the ideas immanent in nervous activity (1943). In *Ideas That Created the Future*, pages 79–88. The MIT Press.

Mertens, K. and Vansteelandt, S. (2017). Augmented and doubly robust g-estimation of causal effects under a structural nested failure time model. *Biometrics*, 74(2):472–480.

Mok, T. S., Wu, Y.-L., Thongprasert, S., Yang, C.-H., Chu, D.-T., Saijo, N., Sunpaweravong, P., Han, B., Margono, B., Ichinose, Y., Nishiwaki, Y., Ohe, Y., Yang, J.-J., Chewaskulyong, B., Jiang, H., Duffield, E. L., Watkins, C. L., Armour, A. A., and Fukuoka, M. (2009). Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *The New England journal of medicine*, 361(10):947–57.

Ockene, I. S. and Miller, N. H. (1997). Cigarette smoking, cardiovascular disease, and stroke. *Circulation*, 96(9):3243–3247.

Ockene, J. K. (1987). Smoking intervention: the expanding role of the physician. *American Journal of Public Health*, 77(7):782–783.

Ockene, J. K., Kuller, L. H., Svendsen, K. H., and Meilahn, E. (1990). The relationship of smoking cessation to coronary heart disease and lung cancer in the multiple risk factor intervention trial (MRFIT). *American Journal of Public Health*, 80(8):954–958.

Oganisian, A. and Roy, J. A. (2020). A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2):518–551.

Patel, V., Chisholm, D., Dua, T., Laxminarayan, R., and Medina-Mora, M. E. (2016). *Disease Control Priorities, Third Edition (Volume 4): Mental, Neurological, and Substance Use Disorders*. Washington, DC: World Bank.

Pearl, J. (2009a). *Causality*. Cambridge University Press.

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Pinkus, A. (1999). Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195.

Pletcher, M. J., Varosy, P., Kiefe, C. I., Lewis, C. E., Sidney, S., and Hulley, S. B. (2005). Alcohol consumption, binge drinking, and early coronary calcification: Findings from the coronary artery risk development in young adults (CARDIA) study. *American Journal of Epidemiology*, 161(5):423–433.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Robins, J. (1986a). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. (1986b). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Lecture Notes in Statistics*, pages 69–117. Springer New York.

Robins, J. M. (2002). Analytic methods for estimating HIV-treatment and cofactor effects. In *Methodological Issues in AIDS Behavioral Research*, pages 213–288. Springer US.

Robins, J. M. (2005). *Structural Nested Failure Time Models*. American Cancer Society.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, 3(4):319–336.

Robins, J. M. and Hernán, M. A. (2008). Estimation of the causal effects of time-varying exposures. In *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 553–599. Chapman and Hall/CRC.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5).

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Robins, J. M. and Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics - Theory and Methods*, 20(8):2609–2631.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1).

Ruder, S. (2016). An overview of gradient descent optimization algorithms.

Saegusa, T., Ma, T., Li, G., Chen, Y. Q., and Lee, M.-L. T. (2020). Variable selection in threshold regression model with applications to hiv drug adherence data. *Statistics in Biosciences : Journal of the International Chinese Statistical Association*, 12(3):376–398.

Sen, R., Shanmugam, K., Asnani, H., Rahimzamani, A., and Kannan, S. (2018). Mimic and classify : A meta-algorithm for conditional independence testing.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Sjölander, A. (2016). Regression standardization with the r package stdReg. *European Journal of Epidemiology*, 31(6):563–574.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738.

Stensrud, M. J., Aalen, J. M., Aalen, O. O., and Valberg, M. (2018). Limitations of hazard ratios in clinical trials. *European Heart Journal*, 40(17):1378–1383.

Sun, S., Cao, Z., Zhu, H., and Zhao, J. (2020). A survey of optimization methods from a machine learning perspective. *IEEE Transactions on Cybernetics*, 50(8):3668–3681.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6).

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2021). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, pages 1–12.

Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology*, 26(3):402–410.

Tieleman, T. and Hinton, G. (2015). Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. Technical report.

Tilling, K., Sterne, J. A. C., and Didelez, V. (2012). G-estimation for accelerated failure time models. In *Modern Methods for Epidemiology*, pages 243–260. Springer Netherlands.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., and Wei, L.-J. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 32(22):2380–5.

U.S. Food and Drug Administration (2021). Adjusting for covariates in randomized clinical trials for drugs and biological products guidance for industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-pro Last accessed on 2022-11-11.

van der Laan, M. J. and Robins, J. M. (2003). Unified approach for causal inference and censored data. *Springer Series in Statistics*, page 311–370.

van der Wal, W. M. and Geskus, R. B. (2011). ipw: An R package for inverse probability weighting. *Journal of Statistical Software*, 43(13):1–23.

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219.

VanderWeele, T. J. and Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical methods in medical research*, 21(1):7–30.

Vansteelandt, S. and Joffe, M. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4).

Vock, D. M., Tsiatis, A. A., Davidian, M., Laber, E. B., Tsuang, W. M., Copeland, C. A. F., and Palmer, S. M. (2013). Assessing the causal effect of organ transplantation on the distribution of residual lifetime. *Biometrics*, 69(4):820–829.

Vowels, M. J., Camgoz, N. C., and Bowden, R. (2022). D'ya like dags? a survey on structure learning and causal discovery. *ACM Comput. Surv.* Just Accepted.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

Whitmore, G. A. (1983). A regression method for censored inverse-gaussian data. *Canadian Journal of Statistics*, 11(4):305–315.

Witteman, J. C. M., D'Agostino, R. B., Stijnen, T., Kannel, W. B., Cobb, J. C., de Ridder, M. A. J., Hofman, A., and Robins, J. M. (1998). G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the framingham heart study. *American Journal of Epidemiology*, 148(4):390–401.

Wolf, P. A. (1988). Cigarette smoking as a risk factor for stroke. *JAMA*, 259(7):1025.

Xie, D., Yang, W., Jepson, C., Roy, J., Hsu, J. Y., Shou, H., Anderson, A. H., Landis, J. R., and Feldman, H. I. (2017). Statistical methods for modeling time-updated exposures in cohort studies of chronic kidney disease. *Clinical Journal of the American Society of Nephrology*, 12(11):1892–1899.

Yang, Y., Fasching, P. A., and Tresp, V. (2017). Modeling progression free survival in breast cancer with tensorized recurrent neural networks and accelerated failure time models. In Doshi-Velez, F., Fackler, J., Kale, D., Ranganath, R., Wallace, B., and Wiens, J., editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 164–176. PMLR.

Young, J. G., Hernán, M. A., Picciotto, S., and Robins, J. M. (2009). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis*, 16(1):71–84.

Zhang, M., Joffe, M. M., and Small, D. S. (2011). Causal inference for continuous-time processes when covariates are observed only at discrete times. *The Annals of Statistics*, 39(1).