Cumulative Probability Models for Semiparametric G-Computation

By

Caroline Isabelle Birdrow

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

August 31, 2021

Nashville, Tennessee

Approved:

Andrew J. Spieker, Ph.D.

Bryan E. Shepherd, Ph.D.

ACKNOWLEDGEMENTS

During my time in the Biostatistics Graduate Program at Vanderbilt, I have had the privilege of taking a course with, serving as a teaching assistant for, and being advised on my master's thesis by Dr. Andrew Spieker. His patience, guidance, and wealth of statistical knowledge have been an essential component of my development as a biostatistician, and I am incredibly grateful for his continued dedication to his students and instructional excellence. Furthermore, I would like to thank Dr. Bryan Shepherd for his collaborative work on cumulative probability models which motivated this project, for serving as the reader on my thesis committee, and for championing the efforts of this research. Thank you also to Dr. Robert Greevy for recognizing my potential as a candidate for this program, providing multiple avenues for developing my skills, and giving indispensable advice along the way.

Additionally, I would like to extend a special thanks to each professor whose course I have taken, the teaching assistants who went above and beyond to support my progress in those courses, and my classmates whose comradery and perspectives have been invaluable. I particularly would like to thank Jamie Joseph and her dog, Bernie, for their friendship, debugging assistance, and endless cheerleading. Thank you also to Elizabeth Sigworth for sharing her passion for data visualizations and offering support in that area for this project.

My mother, Sandra, and my sister, Cynthia, deserve all my thanks and more for continuously encouraging me to pursue my dreams even when the obstacles seemed insurmountable and the road unclear. Last, but certainly not least, I would like to thank Collin at Three Brothers Coffee for always remembering my name, pretending not to remember my order, and providing the caffeine needed to sustain me throughout this process.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Time-varying confounding is a commonly encountered challenge in longitudinal observational studies that seek to evaluate the causal effect of a time-dependent treatment [1]. A time-varying confounder is one that lies on the causal pathway from the (prior) treatment to the outcome but also influences future treatment. Simple approaches to account for confounding (e.g., traditional covariate adjustment in a regression model) do not allow one to identify the total causal effect because such confounders also are mediators on the pathway from prior treatment to the outcome. Collider-stratification bias may be introduced if the covariate is caused by a prior exposure and by a prior value of the covariate.

G-computation is a generalization of standardization that can be used to identify causal effects of longitudinal treatment regimes. Specifically, g-computation allows estimation of the expected outcome under the hypothetical scenario in which everyone in the population receives a particular longitudinal treatment regime [2,3]. The standardization formula expresses the mean potential outcome as a marginalized version of the conditional expectation of the outcome (i.e., conditional on baseline confounders). G-computation generalizes the method of standardization to a longitudinal setting in which treatment and covariate values change over time.

G-computation can readily accommodate commonly encountered complex challenges such as censoring and truncation by death. Typical implementations of g-computation are sometimes criticized for their heavy reliance on parametric models, such that there is possible non-robustness to model misspecification. This motivates the use of a semiparametric approach within g-computation such as the cumulative probability model (CPM); briefly, a CPM treats continuous outcomes ordinally and is semiparametric in that it presumes that the error term of a linear model follows a specified distribution up to some unspecified monotonic transformation, an assumption much weaker than the fully parametric models typically used with g-computation [4]. Estimating the parameters of a CPM amounts to estimating the conditional distribution of the outcome, for reasons we will see in Chapter 2. CPMs have several advantages including their ability to accommodate semicontinuous data (arising, for instance, when dealing with observations that fall below a detection limit). CPMs result in only a minor loss in statistical efficiency when compared to properly specified parametric models. Although not as computationally efficient as parametric models, CPMs can be fit with large sample sizes (e.g., N = 40,000).

The objective of this research is to illustrate the utility of CPMs within g-computation as a tool for causal inference in longitudinal studies. We will seek to evaluate degree of concordance between the estimated and theoretical distributions of potential outcomes. We will further assess finite-sample properties such as bias in estimating the mean potential outcome. We will also apply the CPM-based g-computation approach to study three-month cost outcomes in endometrial cancer patients under various adjuvant radiation therapy strategies based on a fully simulated data set designed to mirror key aspects the true distribution of and correlation between variables from a SEER (Surveillance, Epidemiology, and End Results)-Medicare linked database of women with endometrial cancer. A fully simulated version of these data was generated due to the proprietary nature of the data; results from this analysis are intended to be used only for illustrative purposes. Because cost outcomes are measured intermittently, we use the nested g-computation adaptation appropriate for settings in which the outcome of interest is a sum of repeated outcome measures over time [3].

CHAPTER 2

METHODS

2.1 Notation and Causal Assumptions

In this section, we briefly introduce the notation used throughout this thesis and summarize the key assumptions required to identify the mean potential outcome under a longitudinal treatment regime. We let *Y* denote the observed value of the outcome, *L* denote measured covariates, and *A* denote observed treatment administered. Let $i = 1, \ldots, n$ index independently sampled study subjects, and let $t = 1, \ldots, T$ index uniformly spaced time points measured for each independently sampled study subject. We further use overbar notation to denote variable history up until the indexed time (or entire variable history absent a time index).

Figure 2.1 illustrates the temporal ordering of the variables in a hypothetical study involving three treatment and covariate measurements prior to a single outcome:

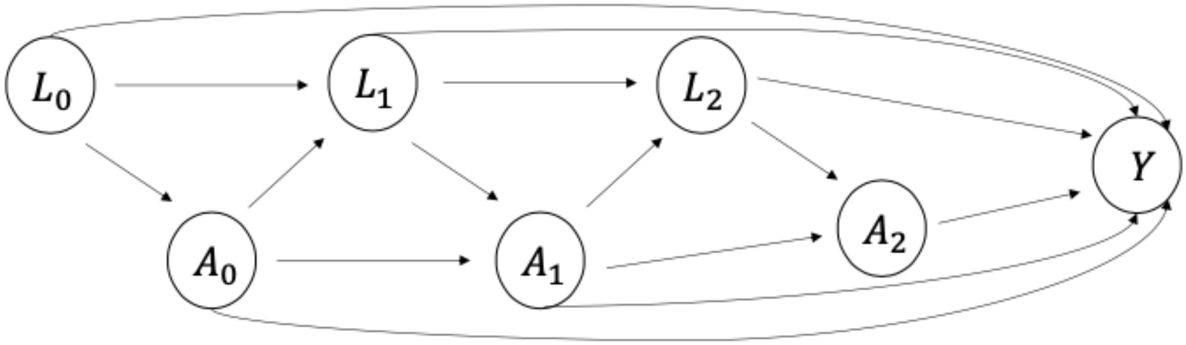

Figure 2.1: Causal diagram for a longitudinal study with three time points. Treatment is denoted as A_t for $t = 0, \ldots, 2$; covariate values are denoted as L_t for $t = 0, \ldots, 2$; and the outcome is denoted as *Y*. This is based on a similar diagram by Daniel et al.

This causal diagram makes clear, for instance, that L_1, a common cause of future treatment A_1 and future confounder L_2, is also a mediator on the path from prior treatment, A_0 to the outcome, Y. Further, all treatment and confounder values occurring from times $t = 1$ and $t = 2$ are colliders (i.e., they have two causes). While we will later conduct a simulation study involving three time points, we will formulate the identifying assumptions under the more general case of T time points, which is represented in the DAG of Figure 2.2 (shown below).

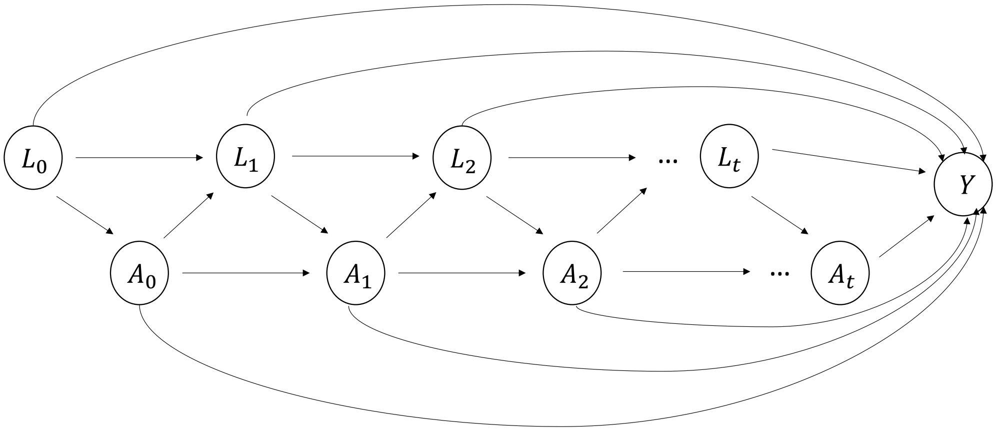

Figure 2.2: Causal diagram for longitudinal study with a time-dependent confounder influenced by previous treatment/exposure, based on a similar diagram by Daniel et al.

The identifying assumptions are described by Hernán and Robins [10]; generalizations to repeated outcomes are further summarized by Spieker et al. [3], which are summarized as follows:

1. *Sequentially ignorable treatment*: The observed treatment at time t is independent of the potential outcome and the measured covariate at time t, conditional on measured covariate history through time t – 1. This also is referred to as “no unmeasured confounding” or “conditional exchangeability.” Formally, this assumption is expressed as:

$$(Y_t^{\bar{a}}, L_t^{\bar{a}}) \perp A_t \mid \bar{L}_{t-1}, \bar{A}_{t-1}, Y_{t-1} \ \forall \ \bar{a} \in \bar{A}, \forall \ t \in \{0, \dots, T\}$$

2. *Positivity*: Each subgroup defined by their covariate history has a non-zero probability of receiving each treatment regime under consideration. Mathematically, this is written as:

$$0 < P(A_t|\, \bar{L}_t, \bar{A}_{t-1}, \bar{Y}_{t-1}) < 1, \forall\ (\bar{L}_t, \bar{A}_{t-1}, \bar{Y}_{t-1}){:}\, P(\bar{L}_t, \bar{A}_{t-1}, \bar{Y}_{t-1}) > 0;\ 1 \leq t \leq\ T.$$

3. Stable-unit-treatment-value assumption (SUTVA):
   a. *No interference*: $Y_{it}^{\bar{a}}$, the potential outcome for subject *i* under treatment regime $\bar{a}$ at time *t* is independent of the observed treatment assignment of others. This can be expressed as:

$$Y_{it}^{\bar{a}}\ \perp\ \bar{A}_{i'}, 1 \leq i \neq i' \leq N, \forall\, t.$$

   b. *Consistency*: Under the treatment actually received, the observed outcome is precisely the potential outcome under the observed treatment. Formally, this can be represented as:

$$Y_i = Y_i^{\bar{A}_i}\,.$$

The standardization formula is given as follows, and is valid under the cross-sectional version of the assumptions listed above:

$$\mathrm{E}(Y^a) = \sum_{\ell} \mathrm{E}(Y|A =\ a, L = \ell) p_L(\ell).$$

The g-formula is the longitudinal generalization of the above formula, given by:

$$\mathrm{E}(Y^{\bar{a}}) = \sum_{\bar{\ell}} \mathrm{E}(Y|\bar{A} = \bar{a}, \bar{L} = \bar{\ell}) \prod_{t=0}^{T} p_{L_t}(\ell_t|\bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{\ell}_{t-1}).$$

Note that the above expression is suitable for the setting in which there is a single outcome measured following the time-varying confounders and treatment. While our simulation study will reflect this simpler setting, it will be useful in our application (in which the outcome of interest is characterized by $Y = \sum_{t=1}^{T} Y_t$) to understand the *nested* g-formula, which amounts to applying the above expression to each of the repeated outcome measures, each time treating prior values of the outcome as covariates [3]:

$$\mathrm{E}(Y^{\bar{a}}) = \sum_{t=0}^{T} \sum_{\bar{\ell}} \mathrm{E}(Y_t | \bar{A}_t = \bar{a}_t, \bar{L}_t = \bar{\ell}_t) \prod_{k=0}^{t} P_{L_k}(\ell_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{\ell}_{k-1}).$$

2.2 Model Specification

Neither standardization nor the g-formula requires specification of the treatment assignment mechanism. However, the conditional distribution of the outcome and the confounders need to be modeled in some fashion. Notably, the distribution of the baseline covariate, L_0, can be modeled empirically as it is not influenced by any prior variables on the DAG. The covariate L_1, depends on L_0 and A_0. Modeling this covariate requires us to condition on the covariate and treatment history; we propose in this work to use a cumulative probability model to do this as flexibly as possible. By similar logic, we will need to do this for future values of the covariates and for the outcome.

We now introduce the CPM, which will be used throughout this work to model the conditional distributions necessary for implementation of the g-formula. We will describe it specifically for a general outcome, Y, and general covariate profile, X. We assume that Y is realized as a monotonic transformation of some latent Y^* [4]. Expressed formally, $Y = H(Y^*)$ where $H(\cdot)$ is an increasing function, $Y^* = \beta^T X + \epsilon$, and $\beta^T X$ is a linear combination of the input variables

X. Also note that ϵ denotes the error term, for which the distribution, F_ϵ, is specified. Without loss of generality, it is assumed that $y_1 \leq \ldots \leq y_n$ and that $\alpha(\cdot)$ is increasing. The CPM formula for *Y* conditional on *X* can be derived as follows:

$$\begin{aligned} F(y|X) &= P(Y \leq y|X) \\ &= P[H(\beta^T X + \epsilon) \leq y|X] \\ &= P[\epsilon \leq H^{-1}(y) - \beta^T X|X] \\ &= F_\epsilon[H^{-1}(y) - \beta^T X]. \end{aligned}$$

This can be simplified by setting $G = F_\epsilon^{-1}$ (we refer to *G* as the link function), and $\alpha = H^{-1}$:

$$G[P(Y \leq y\,|X)] = \alpha(y) - \beta^T X.$$

Cumulative probability models are considered semiparametric because, while $H(\cdot)$ is unspecified, the conditional distribution of *Y* is modeled parametrically up to the unspecified transformation (possibly along with the association between *X* and *Y*) [4].

We return to the nested g-formula and describe how CPMs can be used in this scenario. Recall that the L_1 is a descendent of both L_0 and A_0. If we believe the error distribution to be normal (e.g., with a *probit* link function, we can express the conditional cumulative distribution function (CDF) of L_1 as:

$$P(L_1 \leq \ell_1|L_0 = \ell_0, A_0 = a_0) = \Phi\big(\alpha(\ell_1) - (\beta_1\ell_0 + \beta_2 a_0)\big).$$

Analogous CPMs can be specified for all time-varying confounders that require specification of a conditional distribution. When concerns regarding overfitting become apparent, Markov assumptions can be invoked. For instance, we may assume a variable to only depend on covariate values in the same and directly preceding interval [3]. If there is a single outcome measured at the end of the study, it is typical to presume it depends upon all prior treatments and confounders. If presuming a probit link, the CPM could be expressed, for instance, as:

$$P\left(Y \leq y \middle| \bar{L} = \bar{\ell}, \bar{A} = \bar{a}\right) = \Phi\left(\alpha(y) - \left(\sum_{t=1}^{T} \beta_t^{\ell} \ell_t + \sum_{t=1}^{T} \beta_t^{a} a_t\right)\right).$$

2.3 Estimation Procedure

The process by which the mean value of a potential outcome as expressed by the g-formula is estimated is referred to a *g-computation*. It is a computational/numeric tool that is used to evaluate what is in most realistic cases an analytically intractable high-dimensional integral (the exception to this is the setting in which there few time points with only discrete covariates, each having few categories). We now discuss how to implement this procedure when CPMs are used to estimate all conditional distributions.

1. Set the number of Monte-Carlo iterations, *M* (say, *M* = 5,000).
2. Set the longitudinal treatment regime to be studied: $\bar{a} = (a_0, \ldots, a_T)$.
3. Estimate the parameters of all necessary CPMs (details by Liu et al. [4]).
4. Generate *M* random draws from the empirical distribution of L_0; call them ℓ_0^1, …, ℓ_0^M.
5. Generate *M* random draws from the distribution of $L_1^{\bar{a}}$, call them ℓ_1^1, …, ℓ_1^M. This procedure can be broken down into the sub-steps:
    a. Construct an estimate of the observation-specific CDF, $\hat{P}(L_1 \leq \ell_1 | L_0 = \ell_0^1, A_0 = a_0)$, based on the estimated parameters of the CPM for L_1.

   b. Use the inverse CDF method to take a random sample from the estimated CDF of Step 5(a).

6. Repeat Step 5 to take random draws from the estimated distribution of each of $L_2^{\bar{a}}$, …, $L_T^{\bar{a}}$, and $Y^{\bar{a}}$. Call the random draws from the latter distribution y^1, …, y^M.
7. The average of the randomly drawn outcomes, $M^{-1}\sum_{m=1}^{M} y^m$, is a point estimate of $\mathrm{E}[Y^{\bar{a}}]$.

One can cycle through this procedure to study the mean of any number of longitudinal treatment regimes, from which various longitudinal causal effects can be studied.

2.4 Extensions of the Estimation Procedure

As previously mentioned, this procedure also can be adapted to include repeated outcome measurements as will be done within the application to the fully simulated SEER-Medicare data set. Estimating repeated outcome measurements requires the use of nested g-computation so that the original g-computation formula can be used within each time interval of the study. Nested g-computation amounts to cycling through this procedure for each outcome and treats past outcomes as confounders.

CHAPTER 3

SIMULATION

In this chapter, we use Monte Carlo techniques to demonstrate the utility of the CPM within g-computation.

3.1 Description of Simulation Setup

We conduct a simulation study under the setting of three time points (i.e., $T = 2$, with three values of A and L), and a sample size of $n = 1{,}000$. The general form of the data generation mechanism for this simulation is described as follows:

1. $L_{0i} \sim N(\mu = 0, \sigma^2 = 1)$.
2. $A_{0i} \sim \text{Bernoulli}(p = \text{expit}(\alpha_0 + \alpha_1 L_{0i}))$.
3. $L_{1i} = \begin{cases} 0 & \text{if } L_{1i}^* < -1 \\ \text{expit}(L_{1i}^*) & \text{otherwise} \end{cases}$, where $L_{1i}^* \sim N(\mu = \beta_0 + \beta_1 L_{0i} + \beta_2 A_{0i}, \sigma^2 = 0.3^2)$.
4. $A_{1i} \sim \text{Bernoulli}(p = \text{expit}(\alpha_0 + \alpha_1 L_{0i} + \alpha_2 A_{0i} + \alpha_3 L_{1i}))$.
5. $L_{2i} = \begin{cases} 0 & \text{if } L_{2i}^* < -1 \\ \text{expit}(L_{2i}^*) & \text{otherwise} \end{cases}$, where $L_{2i}^* \sim N(\mu = \beta_0 + \beta_1 L_{1i} + \beta_2 A_{1i}, \sigma^2 = 0.3^2)$.
6. $A_{2i} \sim \text{Bernoulli}(p = \text{expit}(\alpha_0 + \alpha_1 L_{1i} + \alpha_2 A_{1i} + \alpha_3 L_{2i}))$.
7. $Y_i = \begin{cases} 0 & \text{if } Y_i^* < -1 \\ \text{expit}(Y_i^*) & \text{otherwise} \end{cases}$, where $Y_i^* \sim N(\mu = \gamma_0 + \sum_{t=0}^2 \gamma_t^L L_{ti} + \sum_{t=0}^2 \gamma_t^A A_{ti}, \sigma^2 = 1)$.

Note that this data generation mechanism features a Markov assumption in which, conditional on covariate history in the concurrent and prior interval, treatment and confounders are conditionally independent of variables further back in time than a single interval. In this particular simulation, we set the following parameter values:

- $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-0.3, 0.7, 0.5, 0.7)$.
- $(\beta_0, \beta_1, \beta_2) = (-0.6, 0.25, 0.2)$.
- $(\gamma_0, \gamma_1^L, \gamma_2^L, \gamma_3^L, \gamma_1^A, \gamma_2^A, \gamma_3^A) = (-2.5, 0.2, 0.4, 0.6, 0.1, 0.3, 0.5)$.

The treatment assignment is generated as a binary variable at each time point, resulting in a total of 2^3 = 8 possible longitudinal treatment regimes. For ease of presentation, we restrict our attention to four distinct treatment regimes defined by the time at which treatment commences. The regimes can be realized as follows:

Regime 1: $\bar{a} = (0,0,0)$ (i.e., never treated).

Regime 2: $\bar{a} = (0,1,1)$ (i.e., treatment commences after one period of no treatment).

Regime 3: $\bar{a} = (0,0,1)$ (i.e., treatment commences after two periods of no treatment).

Regime 4: $\bar{a} = (1,1,1)$ (i.e., always treated).

There is no closed-form expression that allows us to derive a closed-form expression for the true mean potential outcome under a particular longitudinal treatment regime. However, the true values can be ascertained computationally by sampling under the data generating mechanism under the very large sample size of *N* = 1,000,000 subjects and taking the mean of the generated outcomes; of course, we plug in the treatment regime of interest, $\bar{a}$, in order to accomplish this (as opposed to actually generating the treatment randomly in a fashion that depends upon time-dependent covariates). Being able to computationally ascertain the value of the true mean potential outcome will allow us to characterize finite sample properties such as bias when using CPMs to estimate their values.

3.2 Proof of Concept Regarding the CPM

In evaluating the performance of the CPM method, it first is illustrative to use Monte-Carlo techniques to graphically compare the theoretical marginal distributions of the potential outcome and the time-dependent confounders to their respective estimated marginal distributions as determined by the CPM method under a large sample. To achieve a numeric representation of the true distributions, we simulate data under the true parameter values using the procedure discussed in the final paragraph of Section 3.1 under each of the four longitudinal treatment regimes. We then use the CPM-based g-computation method under a single simulation replicate (with the data generation mechanism described at the start of Section 3.1) with a total sample size of N = 1,000 independent observations and M = 5,000 Monte Carlo iterations. The resulting distributions are displayed in Figures 3.1 through 3.4.

For each treatment regime, we found that the estimated marginal distributions very closely reflected their respective theoretical distributions in almost all cases. One notable discrepancy is observed in estimating the distribution of L_1 (Figure 3.4), which requires further investigation within future work. Overall, these distributions suggest that the method of implementing CPMs within g-computation is extremely effective in estimating the distributions of interest.

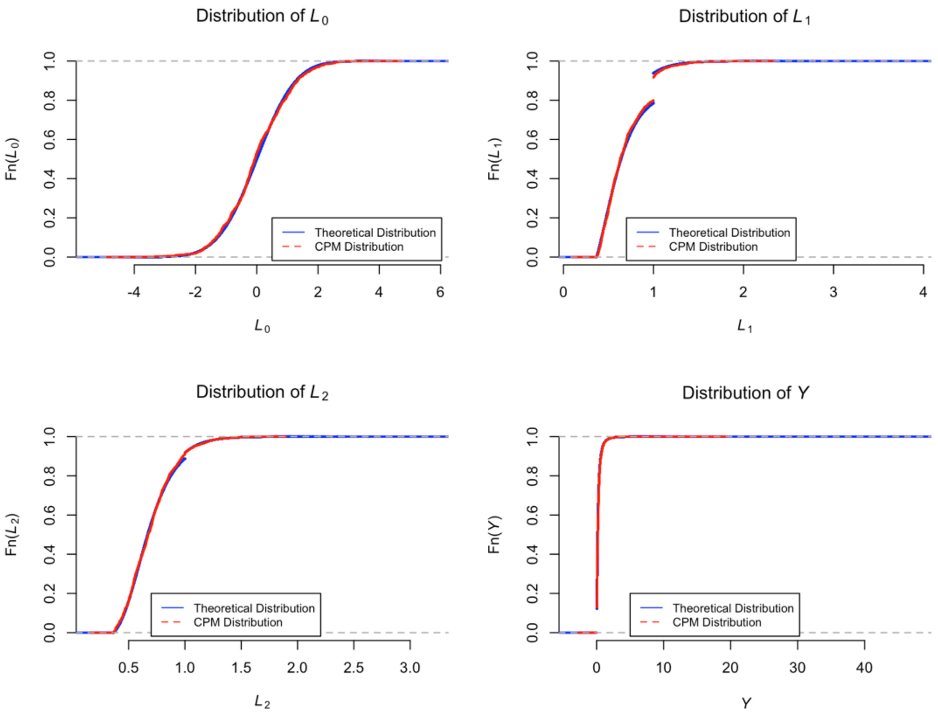

Figure 3.1: Theoretical and CPM distributions of the time-dependent cofounder (all three time points) and the outcome, Y, under Regime 1 ($\bar{a} = (0,0,0)$; never treated).

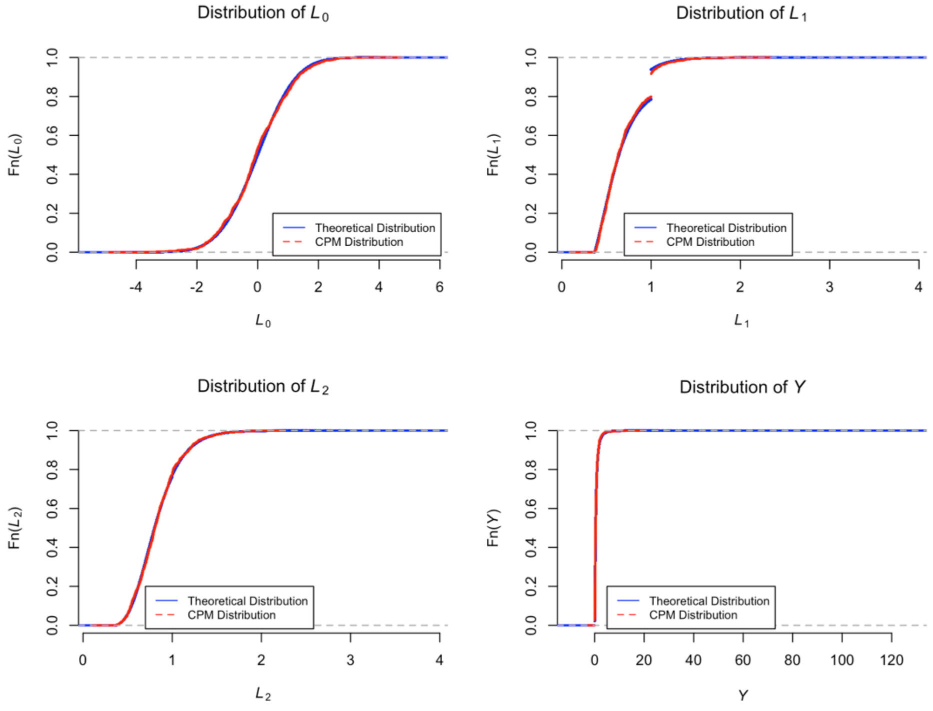

Figure 3.2: Theoretical and CPM distributions of the time-dependent cofounder (all three time points) and the outcome, Y, under Regime 2 ($\bar{a} = (0,1,1)$; treatment commences after a period of no treatment).

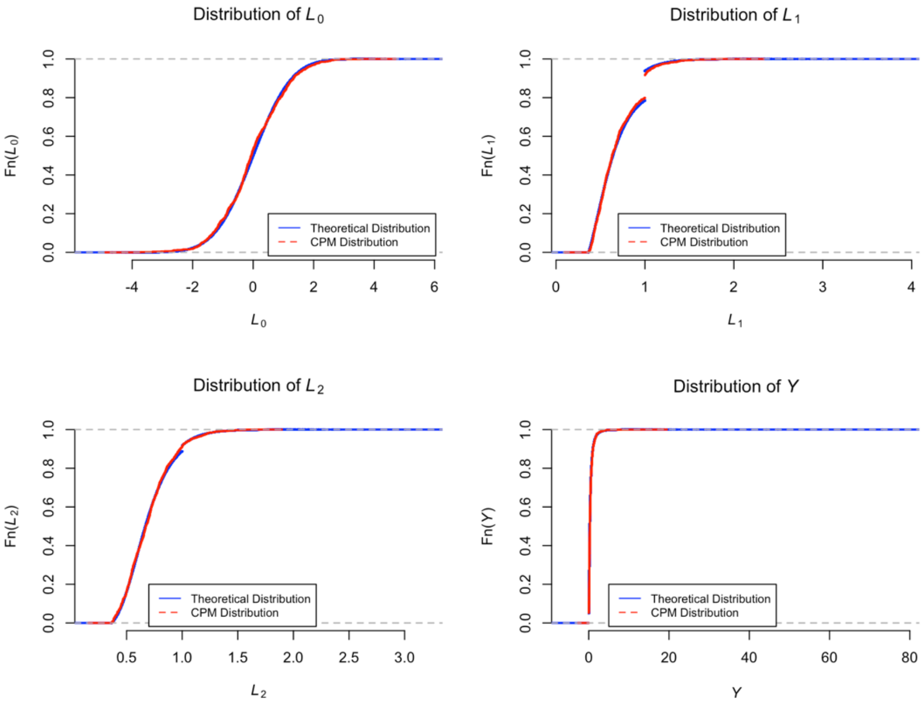

Figure 3.3: Theoretical and CPM distributions of the time-dependent cofounder (all three time points) and the outcome, Y, under Regime 3 ($\bar{a} = (0,0,1)$; treatment commences after two periods of no treatment).

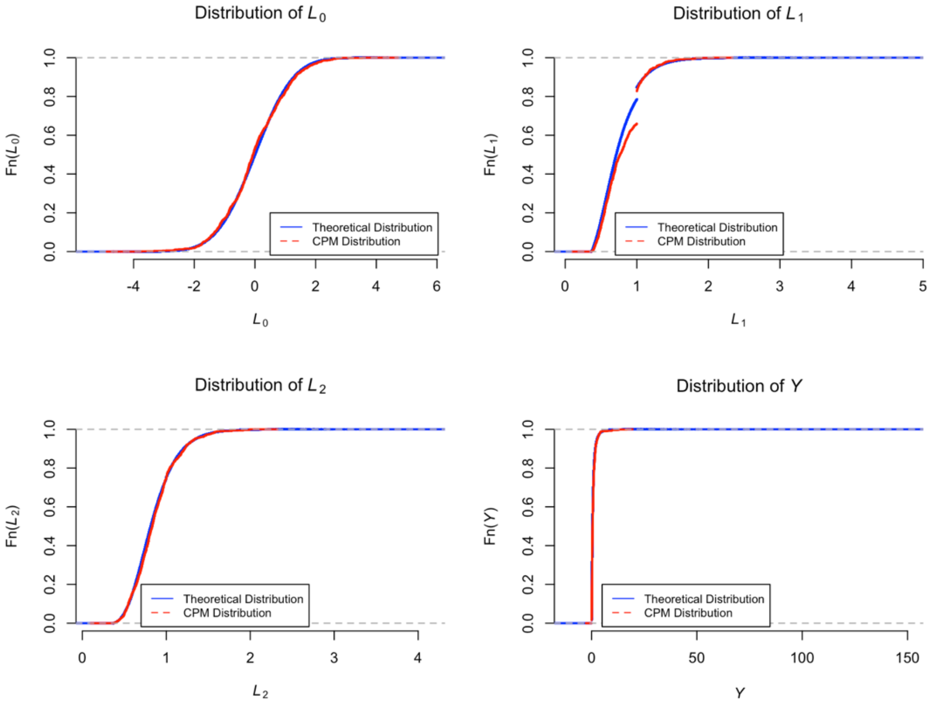

Figure 3.4: Theoretical and CPM distributions of the time-dependent cofounder (all three time points) and the outcome, Y, under Regime 4 ($\bar{a} = (1,1,1)$; always treated).

The above is more or less a "proof of concept" simulation that illustrates the mechanism by which we hypothesize CPMs will have high utility in g-computation. Specifically, CPMs appear to be able to flexibly and adequately allow random draws from marginal distributions that are unlikely to be well captured by simple parametric models.

3.3 Further Simulation-based Assessment of Finite Sample Properties

We further seek to evaluate the bias of this approach, as given by the difference between $\mathrm{E}[Y^{\bar{a}}]$ and $\widehat{\mathrm{E}}[Y^{\bar{a}}]$. We observe in Table 3.1 below the estimated means for each potential outcome under each of the four longitudinal treatment regimes. This result provides a further indication that the CPM approach is a suitable method for inclusion in the g-computation procedure for estimating the mean value of a potential outcome.

Table 3.1: Bias of the CPM-based g-computation approach.

Regime	$\mathbf{E}[\boldsymbol{Y}^{\bar{\boldsymbol{a}}}]$	$\widehat{\mathbf{E}}[\boldsymbol{Y}^{\bar{\boldsymbol{a}}}]$	$\widehat{\mathbf{E}}[\boldsymbol{Y}^{\bar{\boldsymbol{a}}}]$ - $\mathbf{E}[\boldsymbol{Y}^{\bar{\boldsymbol{a}}}]$
$\bar{\boldsymbol{a}} = (\mathbf{0}, \mathbf{0}, \mathbf{0})$	0.281	0.283	0.00188
$\bar{\boldsymbol{a}} = (\mathbf{0}, \mathbf{1}, \mathbf{1})$	0.692	0.696	0.00412
$\bar{\boldsymbol{a}} = (\mathbf{0}, \mathbf{0}, \mathbf{1})$	0.469	0.473	0.00477
$\bar{\boldsymbol{a}} = (\mathbf{1}, \mathbf{1}, \mathbf{1})$	0.804	0.805	0.000598

Our simulation study of the finite-sample properties of implementing CPMs within g-computation suggests that this approach is useful for estimating the mean potential outcome in a longitudinal setting with time-varying confounders. Not only do the theoretical distributions of the cofounder and outcome nearly match the CPM distributions, but also the bias in estimating the mean potential outcome is negligible.

CHAPTER 4

APPLICATION TO CUMULATIVE MEDICAL COSTS IN ENDOMETRIAL CANCER

4.1 Background: Endometrial Cancer

In the United States, endometrial cancer is the fourth most common type of cancer in women, and fifth most common cancer-related cause of death overall [5]. A total of 63,230 women in the United States were diagnosed with endometrial cancer in 2018, with approximately 11,350 cases resulting in death [5]. Additionally, because endometrial cancer is usually a postmenopausal disease and the U.S. population is continuing to increase in age, the prevalence of this disease is only expected to rise; in spite of this trend, older patients tend to receive treatments that are less aggressive than those received by younger individuals, as some physicians believe that older adults have less tolerance for therapy [6].

To further complicate this situation, spending on medical care in the United States has been found to exceed other high-income countries almost two-fold [7]. In 1996, total health care expenditures were approximately $1.4 trillion, and this number increased to about $3.1 trillion by 2016 [8]. Despite this high level of spending, rates of medical care utilization in the U.S. are comparable with other high-income countries that have lower levels of spending [7].

It is of interest to identify cost-effective treatments for endometrial cancer for the purposes of policy development, planning, and budgeting [3]. Considerations regarding safety, efficacy, and tolerability all enter into patient-specific treatment decisions. Though in an ideal world cost would be no factor, the burden of treatment costs also needs to be understood on a population level for the purposes of resource allocation. Since patient-specific treatment is typically adaptive and iterative in nature, principled methods for estimating costs of longitudinal treatment strategies are of critical importance.

4.2 Description of Simulated Data Set

The SEER (Surveillance, Epidemiology, and End Results)-Medicare Health Outcomes Survey connects information from surveys of Medicare recipients with cancer registry data [9]. The data set to which we will apply CPM-based g-computation approach is applied is fully simulated to protect the proprietary nature of the SEER-Medicare data but was designed to reflect key aspects of the overall variable distributions in the original data base. Our results are intended to be taken only as an illustration and are not suitable for particular policy recommendations. For the remainder of the thesis, our discussion of the data will be based on the simulation version.

All women in these data had undergone hysterectomy, the only gold standard approach to evaluate cancer grade and stage. We focus in this particular project on patients with stage 1A (N = 8,618) or 1B (N = 2,924) cancer. There are 11,542 participants in total who each have a unique ID number. The database participants' age and cancer stage were recorded. Additional information collected includes the Charlson comorbidity index recorded at the beginning of the first, second, and third months post-hysterectomy; treatment within the first, second, and third months of hysterectomy; medical cost (in US dollars, $) accumulated within the first, second, and third months of hysterectomy; and the number of hospitalizations within the first, second, and third months of hysterectomy.

Possible values of the Charlson comorbidity score are 0, 1, 2, and 3+. These data are aggregated based on different kinds of radiation therapy and chemotherapy. The different kinds of radiation therapy that might have been received by the patients include three-dimensional conformal radiation, proton therapy, stereotactic body radiation therapy (SBRT), cyberknife, interstitial radiation, and intensity-modulated radiation therapy (IMRT). As for chemotherapy, the various types available include paclitaxel, carboplatin, doxorubicin, cisplatin, and docetaxel. For the purposes of this analysis, subjects were categorized as having received one of the following treatment options: no treatment, radiation therapy (RT), chemotherapy (CT), or both. Because patients are managed in large part according to prior treatment exposure, each treatment is considered an absorbing state such that treated patients will continue to be considered as treated.

The purpose of this study is to estimate the mean total costs accumulated over time under various hypothetical treatment regimes. This longitudinal study includes two time-varying confounders: the Charlson comorbidity index and the number of hospitalizations. These variables likely are associated with both the treatment received, as well as the cost of this treatment over time. Figure 4.1 represents the relationships present in this setting at baseline and at the first time point, which can be expanded to all three time points:

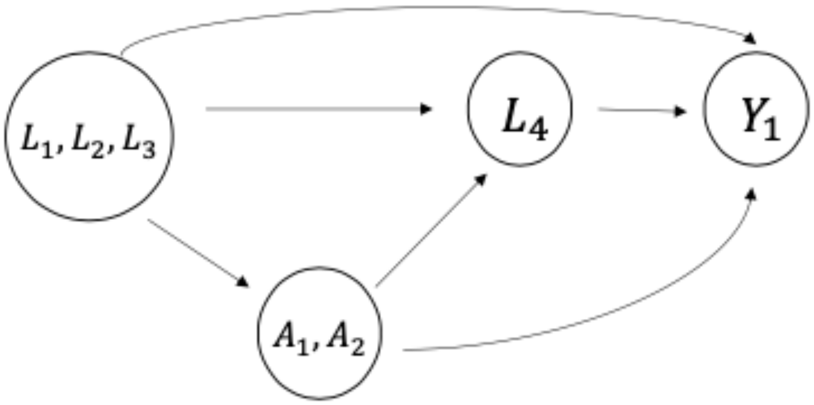

Figure 4.1: Causal diagram for the SEER-Medicare study to demonstrate the temporal ordering in a single time interval. L_1 and L_2 are the time-stable covariates of age and cancer stage, measured at baseline. L_3 is the time-varying Charlson comorbidity index measured at the beginning of the first month, A_1 and A_2 are the time-varying treatments administered within the first month, and L_4 is the time-varying covariate of hospitalizations that accumulates value within the first month. Y_1 represents the cost accumulated within that first month.

The Charlson comorbidity score is a time-varying confounder that is recorded at the beginning of the month, and hospitalizations is a time-varying confounder that accumulates value throughout the month. The outcome, cost, also is not only recorded at the end of the study but also after each month has passed. The cost accumulated within each month is estimated because these values may be of interest if determining when a patient incurs the most costs. Cost also may be a time-varying confounder, as it theoretically could be associated with future treatment decisions and is associated with the final cost.

Regardless of the differences between this setting and the simulation, the CPM approach still can accommodate this scenario. As in the simulation, a first-order Markov assumption is invoked to allow for a current variable to only depend on covariate values in the same and directly preceding interval [3].

4.3 Descriptive Statistics

In order to characterize the variables in the data set, a table of descriptive statistics, stratified by cancer stage, is presented in Table 4.1. Within the table, p-values are provided to indicate differences in covariate values between subjects with stage 1A endometrial cancer and stage 1B endometrial cancer. For categorical variables, these p-values are obtained from the Chi-square test. A Fisher's exact test is performed for categorical variables with sparse cell counts. A Kruskal-Wallis test is performed to yield the p-values for continuous variables. For categorical variables, the number of observations and frequency are presented in the table. For continuous variables, mean (standard deviation) and median (min, max) are presented. Table 4.2 presents an additional table of descriptive statistics, stratified by baseline values of the Charlson comorbidity index. No tests of significance are performed to create this second table, and p-values are not provided.

For every variable in the data set except for the cost within three months, covariate values between participants with stage 1A cancer and stage 1B cancer differ statistically ($p < 0.05$). Those with stage 1B cancer tend to be older than those with stage 1A cancer (75.1 (6.94) vs. 73.1 (6.37)), and the overall mean age regardless of cancer stage is 73.6 years.

For each category of the Charlson comorbidity index and at every time point, there are more patients with stage 1A cancer than stage 1B. Overall, the majority of patients fall in the 0 category of this Charlson comorbidity index at baseline ($N = 7{,}915$), and the two categories with the greatest number of patients at $t = 1$ and $t = 2$ are 0 ($N = 4{,}643$ and $N = 4{,}509$, respectively) and 2 ($N = 5{,}942$ and $N = 5{,}184$, respectively).

For each treatment at each time point, except for radiation therapy within the first month, it also is true that there are more participants with stage 1A cancer than stage 1B cancer. At each time point, the majority of all subjects receive no treatment, and the second greatest number of subjects receive radiation.

For readability, cost is scaled by 1,000 prior to summarizing its distribution. Cost within one month is higher for the stage 1B subjects, and cost within two months is comparable between the two groups, but the distribution of cost for the stage 1A subjects displays a greater variance. For participants overall, cost increases from $t = 1$ to $t = 2$ and decreases from $t = 2$ to $t = 3$.

While the two cancer stage groups differ statistically in terms of the number of hospitalizations, this variable still appears to be comparable between them. A notable trend is that the overall number of hospitalizations decreases with time.

Table 4.1: Descriptive statistics, stratified by cancer stage.

	1A	1B	Overall	
	(*N* = 8618)	(*N* = 2924)	(*N* = 11542)	P-value
Age at Hysterectomy				< 0.001
Mean (SD)	73.1 (6.37)	75.1 (6.94)	73.6 (6.57)	
Median [Min, Max]	72.0 [60.0, 99.0]	74.0 [60.0, 98.0]	73.0 [60.0, 99.0]	
Charlson Comorbidity Index, *t* = 1				< 0.001
0	6010 (69.7%)	1905 (65.2%)	7915 (68.6%)	
1	1931 (22.4%)	713 (24.4%)	2644 (22.9%)	
2	545 (6.3%)	246 (8.4%)	791 (6.9%)	
3+	132 (1.5%)	60 (2.1%)	192 (1.7%)	
Charlson Comorbidity Index, *t* = 2				< 0.001
0	3588 (41.6%)	1055 (36.1%)	4643 (40.2%)	
1	567 (6.6%)	190 (6.5%)	757 (6.6%)	
2	4325 (50.2%)	1617 (55.3%)	5942 (51.5%)	
3+	138 (1.6%)	62 (2.1%)	200 (1.7%)	
Charlson Comorbidity Index, *t* = 3				< 0.001
0	3486 (40.5%)	1023 (35.0%)	4509 (39.1%)	
1	591 (6.9%)	202 (6.9%)	793 (6.9%)	
2	3764 (43.7%)	1420 (48.6%)	5184 (44.9%)	
3+	777 (9.0%)	279 (9.5%)	1056 (9.1%)	
Treatment within Month 1				< 0.001
None	8379 (97.2%)	2571 (87.9%)	10950 (94.9%)	
Radiation Therapy (RT)	196 (2.3%)	343 (11.7%)	539 (4.7%)	
Chemotherapy (CT)	43 (0.5%)	10 (0.3%)	53 (0.5%)	
Treatment within Month 2				< 0.001
None	6579 (76.3%)	2047 (70.0%)	8626 (74.7%)	
RT	1730 (20.1%)	790 (27.0%)	2520 (21.8%)	
CT	222 (2.6%)	58 (2.0%)	280 (2.4%)	
Both RT & CT	87 (1.0%)	29 (1.0%)	116 (1.0%)	
Treatment within Month 3				< 0.001
None	6070 (70.4%)	1880 (64.3%)	7950 (68.9%)	
RT	1998 (23.2%)	889 (30.4%)	2887 (25.0%)	
CT	343 (4.0%)	94 (3.2%)	437 (3.8%)	
Both RT & CT	207 (2.4%)	61 (2.1%)	268 (2.3%)	

Cost within Month 1				< 0.001
Mean (SD)	15.0 (19.3)	16.3 (19.9)	15.4 (19.5)	
Median [Min, Max]	9.22 [0, 122]	11.3 [0, 135]	9.78 [0, 135]	
Cost within Month 2				< 0.001
Mean (SD)	36.1 (94.2)	36.1 (87.8)	36.1 (92.6)	
Median [Min, Max]	1.94 [0, 2110]	3.76 [0, 766]	2.33 [0, 2110]	
Cost within Month 3				0.154
Mean (SD)	9.69 (60.9)	8.00 (33.5)	9.26 (55.3)	
Median [Min, Max]	1.10 [0.00302, 4330]	1.17 [0.00273, 683]	1.13 [0.00273, 4330]	
Hospitalizations within Month 1				
Mean (SD)	0.925 (0.265)	0.937 (0.244)	0.928 (0.260)	0.032
Median [Min, Max]	1.00 [0, 2.00]	1.00 [0, 2.00]	1.00 [0, 2.00]	
Hospitalizations within Month 2				
Mean (SD)	0.443 (0.539)	0.490 (0.551)	0.455 (0.543)	< 0.001
Median [Min, Max]	0 [0, 3.00]	0 [0, 3.00]	0 [0, 3.00]	
Hospitalizations within Month 3				
Mean (SD)	0.0899 (0.296)	0.0663 (0.261)	0.0840 (0.288)	< 0.001
Median [Min, Max]	0 [0, 3.00]	0 [0, 3.00]	0 [0, 3.00]	

In general, age appears to display a positive association with higher categories of the baseline Charlson comorbidity index. Those with a baseline Charlson comorbidity index of 2 tend to be the oldest, with a mean age of 76.2 years. For both stages of cancer, the greatest number of subjects have a baseline Charlson comorbidity index of 0. Additionally, the overall number of patients who have stage 1A (*N* = 8,618) cancer is greater than the number who have stage 1B (*N* = 2,924).

At *t* = 2, for the Charlson comorbidity index of 0 and 2, the greatest number of individuals have a baseline Charlson comorbidity index of 0. Also at this time point, for the Charlson comorbidity index of 1, the greatest number of subjects have a baseline Charlson comorbidity index of 1. For the Charlson comorbidity index of 3+ at this time point, the largest portion of subjects have a baseline Charlson comorbidity index of 3+. At *t* = 3, for the Charlson comorbidity index of 0, 2, and 3+, the majority of participants have a baseline Charlson comorbidity index of 0. For a Charlson comorbidity index of 1 at this time point, most subjects have a baseline Charlson comorbidity index of 2.

For each treatment level at each time point, the greatest number of subjects have a baseline Charlson comorbidity index of 0. Excluding the third time point, those with a baseline Charlson comorbidity index of 2 have accumulated the greatest costs. Within the third month post-hysterectomy, subjects with a baseline Charlson comorbidity index of 1 have accumulated the most costs.

The number of hospitalizations is comparable between the different baseline values of the Charlson comorbidity index. For each value, the number of hospitalizations decreases with time.

Figure 4.2 illustrates the distribution of age at hysterectomy within the study sample. A right-skewness is observed, with a defined peak at approximately age 65. This likely is due to the fact that U.S. citizens become eligible for Medicare at the age of 65 years.

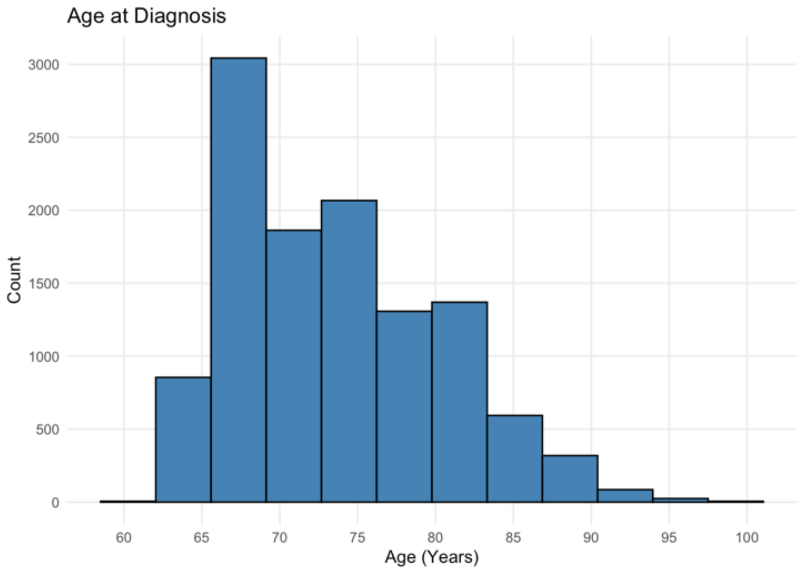

Figure 4.2: Distribution of age at hysterectomy.

Table 4.2: Descriptive statistics, stratified by baseline Charlson comorbidity index.

	0 (*N* = 7915)	1 (*N* = 2644)	2 (*N* = 791)	3+ (*N* = 192)	Overall (*N* = 11542)
Age at Hysterectomy					
Mean (SD)	72.7 (6.36)	75.4 (6.55)	76.2 (6.63)	76.1 (6.71)	73.6 (6.57)
Median [Min, Max]	71.0 [60.0, 98.0]	75.0 [64.0, 99.0]	76.0 [65.0, 98.0]	75.0 [65.0, 93.0]	73.0 [60.0, 99.0]
Stage at Hysterectomy					
1A	6010 (75.9%)	1931 (73.0%)	545 (68.9%)	132 (68.8%)	8618 (74.7%)
1B	1905 (24.1%)	713 (27.0%)	246 (31.1%)	60 (31.2%)	2924 (25.3%)
Charlson Comorbidity Index, *t* = 2					
0	4643 (58.7%)	0 (0%)	0 (0%)	0 (0%)	4643 (40.2%)
1	80 (1.0%)	677 (25.6%)	0 (0%)	0 (0%)	757 (6.6%)
2	3189 (40.3%)	1963 (74.2%)	790 (99.9%)	0 (0%)	5942 (51.5%)
3+	3 (0.0%)	4 (0.2%)	1 (0.1%)	192 (100%)	200 (1.7%)
Charlson Comorbidity Index, *t* = 3					
0	4509 (57.0%)	0 (0%)	0 (0%)	0 (0%)	4509 (39.1%)
1	145 (1.8%)	648 (24.5%)	0 (0%)	0 (0%)	793 (6.9%)
2	2748 (34.7%)	1737 (65.7%)	699 (88.4%)	0 (0%)	5184 (44.9%)
3+	513 (6.5%)	259 (9.8%)	92 (11.6%)	192 (100%)	1056 (9.1%)
Treatment within Month 1					
None	7520 (95.0%)	2493 (94.3%)	755 (95.4%)	182 (94.8%)	10950 (94.9%)
Radiation Therapy (RT)	360 (4.5%)	134 (5.1%)	36 (4.6%)	9 (4.7%)	539 (4.7%)
Chemotherapy (CT)	35 (0.4%)	17 (0.6%)	0 (0%)	1 (0.5%)	53 (0.5%)
Treatment within Month 2					
None	5922 (74.8%)	1960 (74.1%)	593 (75.0%)	151 (78.6%)	8626 (74.7%)
RT	1727 (21.8%)	580 (21.9%)	178 (22.5%)	35 (18.2%)	2520 (21.8%)
CT	181 (2.3%)	81 (3.1%)	13 (1.6%)	5 (2.6%)	280 (2.4%)
Both RT & CT	85 (1.1%)	23 (0.9%)	7 (0.9%)	1 (0.5%)	116 (1.0%)

Treatment within Month 3					
None	5466 (69.1%)	1789 (67.7%)	550 (69.5%)	145 (75.5%)	7950 (68.9%)
RT	1976 (25.0%)	673 (25.5%)	203 (25.7%)	35 (18.2%)	2887 (25.0%)
CT	281 (3.6%)	123 (4.7%)	25 (3.2%)	8 (4.2%)	437 (3.8%)
Both RT & CT	192 (2.4%)	59 (2.2%)	13 (1.6%)	4 (2.1%)	268 (2.3%)
Cost within Month 1					
Mean (SD)	12.8 (17.7)	19.9 (20.9)	24.3 (23.9)	20.9 (23.1)	15.4 (19.5)
Median [Min, Max]	6.40 [0, 115]	14.8 [0, 122]	17.9 [0, 135]	15.2 [0, 84.8]	9.78 [0, 135]
Cost within Month 2					
Mean (SD)	32.6 (89.7)	43.2 (97.8)	49.0 (103)	31.8 (84.9)	36.1 (92.6)
Median [Min, Max]	1.08 [0, 2110]	6.33 [0, 1100]	9.04 [0, 823]	3.10 [0, 686]	2.33 [0, 2110]
Cost within Month 3					
Mean (SD)	9.39 (41.1)	9.70 (90.2)	7.18 (20.2)	6.49 (22.3)	9.26 (55.3)
Median [Min, Max]	0.987 [0.00343, 1590]	1.37 [0.00273, 4330]	1.55 [0.00515, 249]	1.34 [0.0267, 208]	1.13 [0.00273, 4330]
Hospitalizations within Month 1					
Mean (SD)	0.911 (0.287)	0.968 (0.175)	0.973 (0.161)	0.911 (0.285)	0.928 (0.260)
Median [Min, Max]	1.00 [0, 2.00]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 2.00]
Hospitalizations within Month 2					
Mean (SD)	0.391 (0.501)	0.581 (0.601)	0.659 (0.604)	0.505 (0.560)	0.455 (0.543)
Median [Min, Max]	0 [0, 3.00]	1.00 [0, 3.00]	1.00 [0, 3.00]	0 [0, 3.00]	0 [0, 3.00]
Hospitalizations within Month 3					
Mean (SD)	0.103 (0.315)	0.0393 (0.211)	0.0341 (0.182)	0.0990 (0.299)	0.0840 (0.288)
Median [Min, Max]	0 [0, 3.00]	0 [0, 3.00]	0 [0, 1.00]	0 [0, 1.00]	0 [0, 3.00]

The trajectories of Charlson comorbidity index and number of hospitalizations are plotted in Figure 4.3. By construction, the comorbidities that comprise the Charlson score are absorbing states, and they therefore are non-decreasing over time.

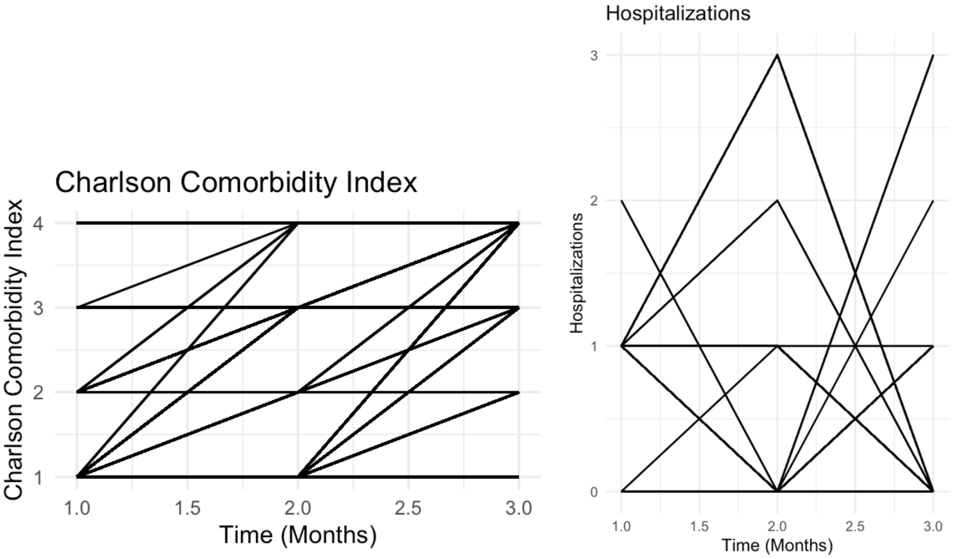

Figure 4.3: All possible trajectories for Charlson comorbidity index and hospitalization trajectories over three months.

The plot (right panel of Figure 4.3) of hospitalizations for each subject over time shows that, for some individuals, the number of hospitalizations increases after 2 months post-hysterectomy, followed by a decrease. For other participants, the number of hospitalizations decreases after 2 months post-hysterectomy and then increases after 3 months post-hysterectomy.

The treatment changes that occurred within the sample over the course of three months are shown in Figure 4.4. From this plot, we observe a decrease in the percentage of individuals who receive no treatment at a given time point out of all the individuals present at that time point. This percentage changes from 94.87% to 68.88% during the study. Conversely, for each treatment option, we can see an increase in the percentage of patients who receive that treatment at a given time point out of all the patients in the sample at that time. For instance, the percentage of individuals who receive radiation changes from 4.67% at t = 1 to 25.01 % at t = 3. Within any given month, the greatest percentage of subjects receive no treatment, and the second greatest percentage receive radiation therapy. The lowest percentage receive both radiation and chemotherapy.

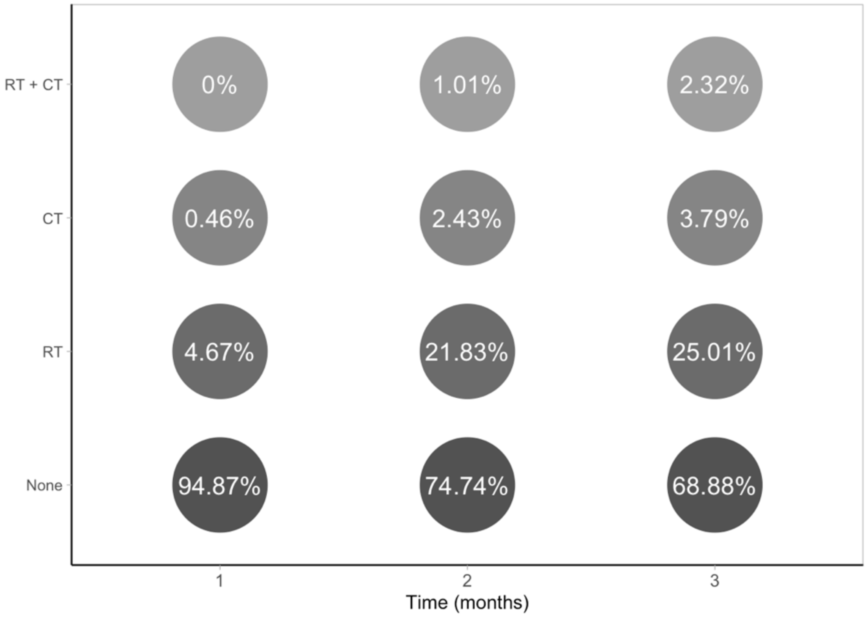

Figure 4.4: Treatment distribution over three months.

Figure 4.5 shows the overall distribution of cost in U.S. dollars scaled by 1,000 after each month post-hysterectomy. From these histograms, it can be seen that cost is severely right-skewed which is an indication that this variable could be log-transformed prior to the analysis. Figure 4.6 shows the overall distribution of cost after log-transforming, and while the distributions still are fairly skewed, they are more symmetric than those in Figure 4.5.

The overwhelming trend of cost accumulated over time, stratified by subject, is that cost increases after 2 months post-hysterectomy and then decreases after 3 months post-hysterectomy. For some individuals, cost spikes after three months post-hysterectomy.

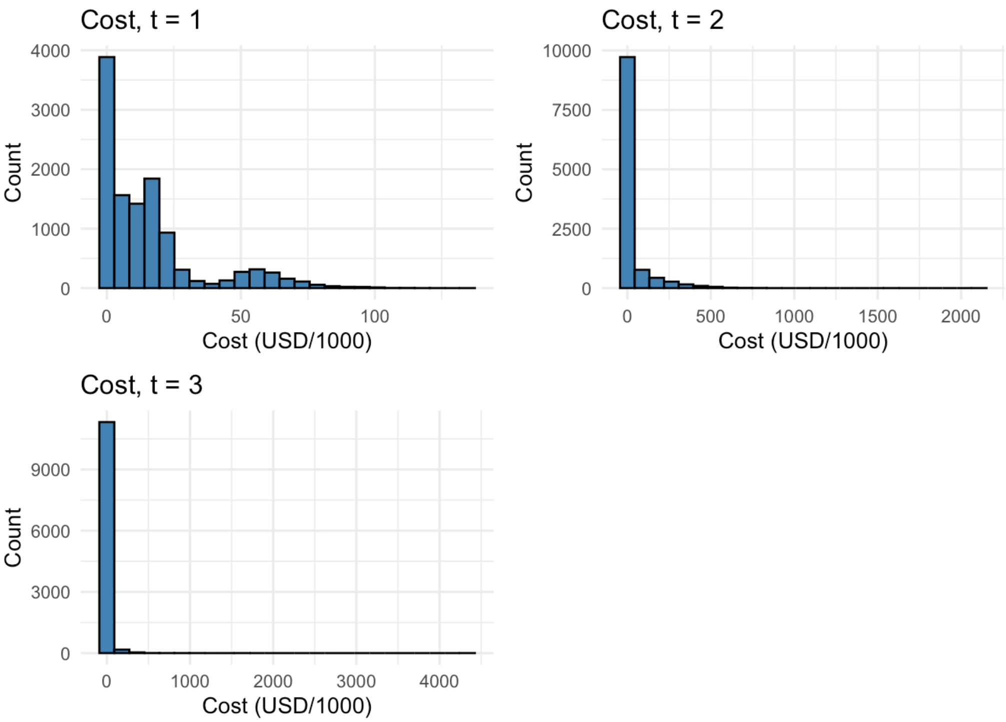

Figure 4.5: Overall distribution of cost after each month post-hysterectomy.

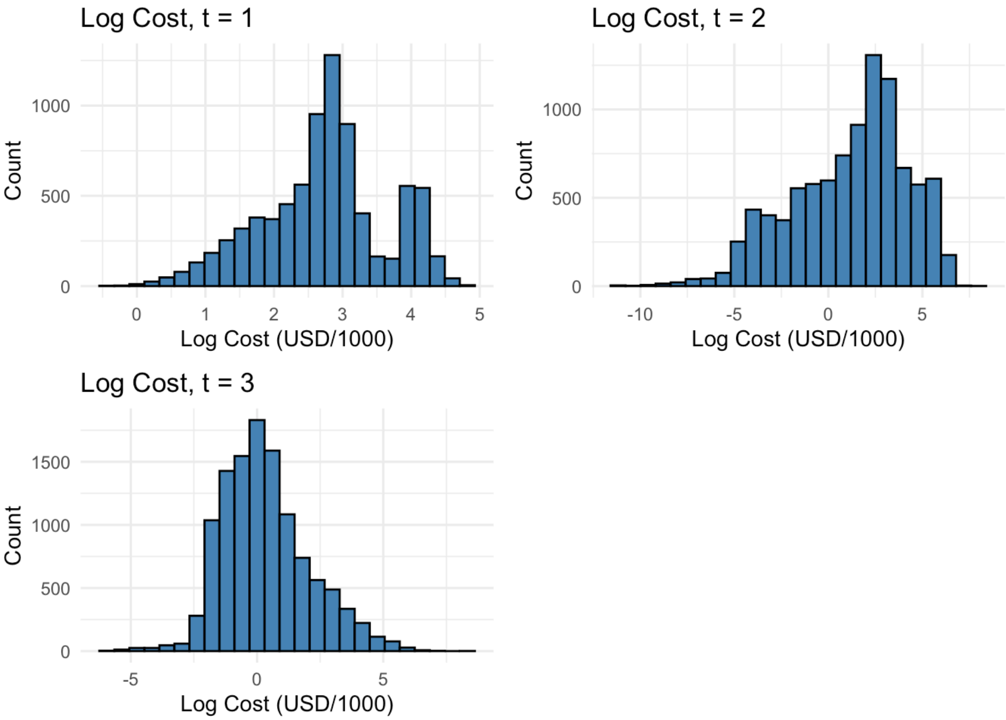

Figure 4.6: Overall distribution of log cost after each month post-hysterectomy.

4.4 Model Specification for G-computation Procedure

In formulas and models moving forward, L again will represent the covariate values, but a subscript of j will be introduced to denote the type of covariate being referenced. Recall that independent subjects are indexed by i, and time points/observations are represented by t. Time-stable covariates of interest will be labeled as follows:

- L_1: Baseline age (at time of hysterectomy).
- L_2: Cancer stage (1A vs. 1B).

Further, time-dependent covariates will be denoted as follows:

- L_{3t}: Charlson comorbidity index (for $t = 1, \ldots , 3$).
- L_{4t}: Hospitalizations (for $t = 1, \ldots , 3$).

Treatment will be denoted by A_{jt}, where radiation therapy is given by A_{1t} and chemotherapy as A_{2t}. The cost outcomes are represented by Y_t, with $t = 1, \ldots, 3$. We will use the notation $I(Y_t = 0)$ to denote a zero-cost at time t. Let $\boldsymbol{L}_t = (L_1, L_{3t}, L_{4t})$ and $\boldsymbol{A}_t = (A_{1t}, A_{2t})$ for ease of notation. Before using the CPM approach to estimate the cost, it is essential to establish and clarify the temporal ordering within the g-computation formula (see Section 2.1 for a presentation of the nested g-formula and Section 2.3 for the estimation procedure). In particular, we will need to be able to take random draws from certain distributions of interest, some of which will be conditional and therefore will need to be estimated using CPMs. Within the first time interval, the distributions of interest are as follows:

- $P(L_1, L_{31})$, estimated empirically.
- $P(L_{41}|\boldsymbol{A}_1, L_1, L_{31})$, estimated with a CPM.
- $P(Y_1|L_{41}, \boldsymbol{A}_1, L_1, L_{31})$, estimated with a zero-inflated CPM.

For the times *t* = 2 and *t* = 3, we invoke a first-order Markov assumption that allows us to estimate the models based on first order lagged variables; these can each be estimated with a CPM:

- $P(L_{3t}|\overline{\boldsymbol{A}}_{t-1}, \overline{\boldsymbol{L}}_{t-1}, \overline{Y}_{t-1})$.
- $P(L_{4t}|\overline{\boldsymbol{A}}_t, \overline{\boldsymbol{L}}_{t-1}, \overline{Y}_{t-1}, L_{3t})$.
- $P(Y_t|\overline{\boldsymbol{A}}_t, \overline{\boldsymbol{L}}_t, \overline{Y}_{t-1})$.

All cumulative probability models are expressed using a probit link function, and all terms in the model are presumed linear.

4.5 Treatment Regimes and Comparisons of Interest

Recall that treatment is considered as an absorbing state such that a patient is considered to receive a treatment from the time of initiation onward. Defining treatment is this manner reduces the dimensionality of the possible treatment options; without such a simplification, the positivity assumption would be more easily violated. Further, this application is limited to the most common treatment trajectories present within the data set. These trajectories are as follows:

- No treatment.

- Radiation therapy commences during third month following hysterectomy.
- Radiation therapy commences during second month following hysterectomy.
- Radiation therapy commences within one month of hysterectomy.

We further stratify our analysis by cancer stage, with mean costs at each time point estimated for both stage 1A patients and stage 1B patients separately.

4.6 Results

Tables 4.3 and 4.4 summarize the results for each cancer stage and treatment trajectory.

Table 4.3: Mean cumulative cost ($/1000) for cancer stage 1A.

	$\hat{\mathbf{E}}[\boldsymbol{Y_1^{\bar{a}}}]$ **(SD)**	$\hat{\mathbf{E}}[\boldsymbol{Y_2^{\bar{a}}}]$ **(SD)**	$\hat{\mathbf{E}}[\boldsymbol{Y_3^{\bar{a}}}]$ **(SD)**	$\hat{\mathbf{E}}[\boldsymbol{Y^{\bar{a}}}]$ **(SD)**
$\boldsymbol{\bar{a}}$ **= (0, 0, 0)**	14.6 (18.4)	18.7 (62.1)	22.7 (73.5)	56.1 (154)
$\boldsymbol{\bar{a}}$ **= (0, 0, RT)**	14.6 (18.6)	18.7 (62.6)	17.3 (85.7)	50.7 (167)
$\boldsymbol{\bar{a}}$ **= (0, RT, RT)**	15.1 (18.9)	16.8 (62.5)	36.6 (142)	68.4 (224)
$\boldsymbol{\bar{a}}$ **= (RT, RT, RT)**	31.8 (23.9)	43.5 (107)	56.2 (137)	131 (268)

Table 4.4: Mean cumulative cost ($/1000) for cancer stage 1B.

	$\hat{\mathbf{E}}[\boldsymbol{Y_1^{\bar{a}}}]$ **(SD)**	$\hat{\mathbf{E}}[\boldsymbol{Y_2^{\bar{a}}}]$ **(SD)**	$\hat{\mathbf{E}}[\boldsymbol{Y_3^{\bar{a}}}]$ **(SD)**	$\hat{\mathbf{E}}[\boldsymbol{Y^{\bar{a}}}]$ **(SD)**
$\boldsymbol{\bar{a}}$ **= (0, 0, 0)**	14.0 (17.5)	16.6 (56.0)	20.2 (62.2)	50.8 (136)
$\boldsymbol{\bar{a}}$ **= (0, 0, RT)**	14.5 (18.0)	17.3 (56.5)	14.9 (54.9)	46.7 (129)
$\boldsymbol{\bar{a}}$ **= (0, RT, RT)**	14.1 (17.4)	11.3 (42.5)	19.9 (64.2)	45.3 (124)
$\boldsymbol{\bar{a}}$ **= (RT, RT, RT)**	31.7 (24.7)	25.1 (69.1)	33.4 (84.7)	90.1 (178)

Among patients with stage 1A cancer, the overall cost tends to be higher for those patients who receive more treatment which, intuitively, is a logical observation. For instance, treatment 1 (which entails no treatment received at each time point) is associated with a lower overall cost than treatment 3 (no treatment at $t = 1$ and radiation therapy at $t = 2$ and $t = 3$) which itself has a lower cost than treatment 4 (radiation therapy at all time points). However, treatment 2 (no treatment at $t = 1$ or $t = 2$ and radiation therapy at $t = 3$) for these patients has the lowest overall cost of all the treatment trajectories. While the exact reason for this is unknown, this could be an indication that receiving some treatment (vs. none) may result in more mild symptoms and, thus, lower total costs. Simultaneously, receiving treatment at one time point vs. more than one time point intuitively will lead to lower total costs.

For stage 1B cancer patients, overall cost is lower for treatment regimes that include more treatment, with the exception of treatment regime 4. For instance, treatment 1 is associated with a higher overall cost than treatment 2 which itself has a higher overall cost than treatment 3. However, treatment 4 for these patients has the highest overall cost of all the treatment trajectories. This could be the case if treatment regimes 2 and 3 are more effective than treatment regimes 1 and 4 in terms of reducing symptoms, which would lead to lower total costs. It also is logical that receiving no treatment would incur a lower cost than receiving treatment at every time point.

For treatment regime 1, among patients in both cancer stage groups, cost tends to increase over the course of three months. This trend also holds for treatments 3 and 4 among subjects with stage 1A cancer. This could be due to either worsening symptoms or more treatment received over time. Treatment 2, for all cancer patients, corresponds with an increase in cost from $t = 1$ to $t = 2$ and a decrease from $t = 2$ to $t = 3$. If receiving treatment at $t = 3$ helps alleviate symptoms, this could be the reason for that decline in cost. For treatments 3 and 4, among patients with stage 1B cancer, costs decrease from $t = 1$ to $t = 2$ and increase from $t = 2$ to $t = 3$. This may be the result if receiving treatment two months post-hysterectomy is effective in decreasing symptoms and, therefore, costs. Perhaps treatment three months post-hysterectomy is not the most effective or symptoms recur.

For all treatments, the overall cost of treating stage 1A cancer patients is higher than the cost of treating stage 1B patients. This could be the case if more treatments are administered to subjects with less-aggressive cancer as a preventative measure.

We display the distribution of log cost at each time point for each cancer stage under treatment trajectory (Stage 1A in Figure 4.7 and Stage 1B in Figure 4.8).

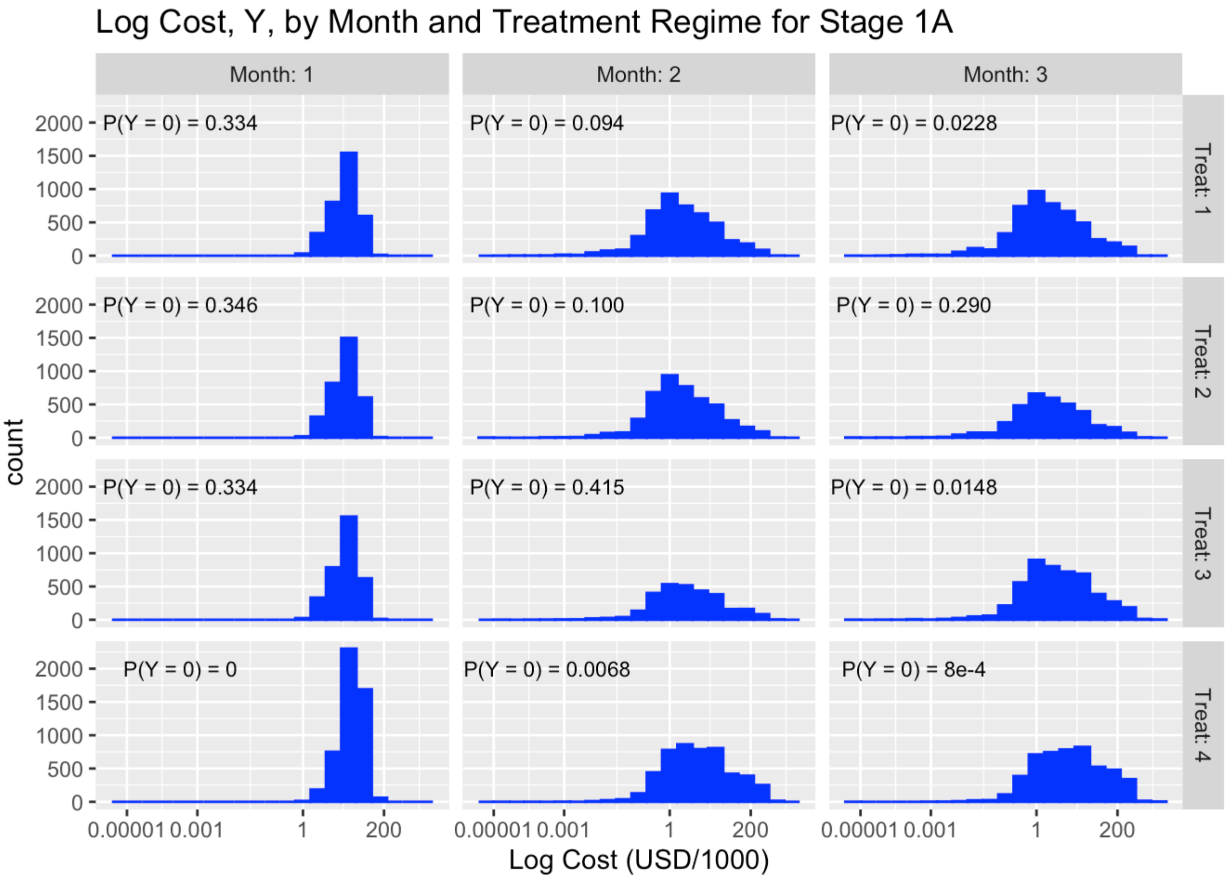

Figure 4.7: Distribution of non-zero log cost with exponentiated *x*-axis labels, stratified by month and treatment trajectory among patients with stage 1A cancer.

The log cost distributions for treatments 1-3 within month 1 display a slight left-skewness and look similar (if not exactly the same). For treatment 4 within month 1, the distribution of log cost also is left-skewed but includes more of the higher log cost values. Because each treatment regime, except for 4, receives no treatment within the first month, it seems likely that their log cost distributions would be similar.

Within the second and third months for treatment 1, the distributions of log cost are flatter and appear to display more non-zero log cost values than within month 1. They both are right-skewed and appear to match one another. It is logical that these patients in treatment regime 1 would generate more log costs over time and that cost at t = 2 and t = 3 would have similar distributions as the patients still do not receive treatment within months 2 and 3. Also similar (if not the same) to these distributions is that for log cost within month 2 for treatment regime 2. This again makes sense, as those patients also do not receive any treatment within the second month.

The distribution of log cost within month 3 for treatment regime 2 is slightly right-skewed, meaning that there are more of the lower log cost values in this distribution. This is a reasonable observation, as we observed in Table 4.3 that the mean cost decreases from t = 2 to t = 3 for cancer stage 1A patients receiving treatment 2. Treatment regime 3's log cost distribution within month 2 is very flat and slightly right-skewed. This could be due to the effectiveness of the treatment in reducing symptoms and, therefore, the number of non-zero log cost values. For treatment regime 3 within month 3, there are more non-zero log cost values in the distribution and there is a right-skewness. It makes sense that more log cost would be accumulated as more radiation therapy is received. The log cost distributions for treatment regime 4 within months 2 and 3 are somewhat symmetric and similar to one another. These distributions may be similar due to the fact that radiation therapy is received at both times.

Additionally, it generally is true for stage 1A cancer patients that the probability of having a zero cost value decreases with increased treatment. It makes sense that receiving more treatment results in more non-zero costs incurred. Two exceptions are the probability of a zero cost for treatment regime 2 within month 3 and treatment regime 3 within month 2. These time points happen to be when treatment begins for these regimes, and it could be that the initial treatment period results in lower (more zero) costs.

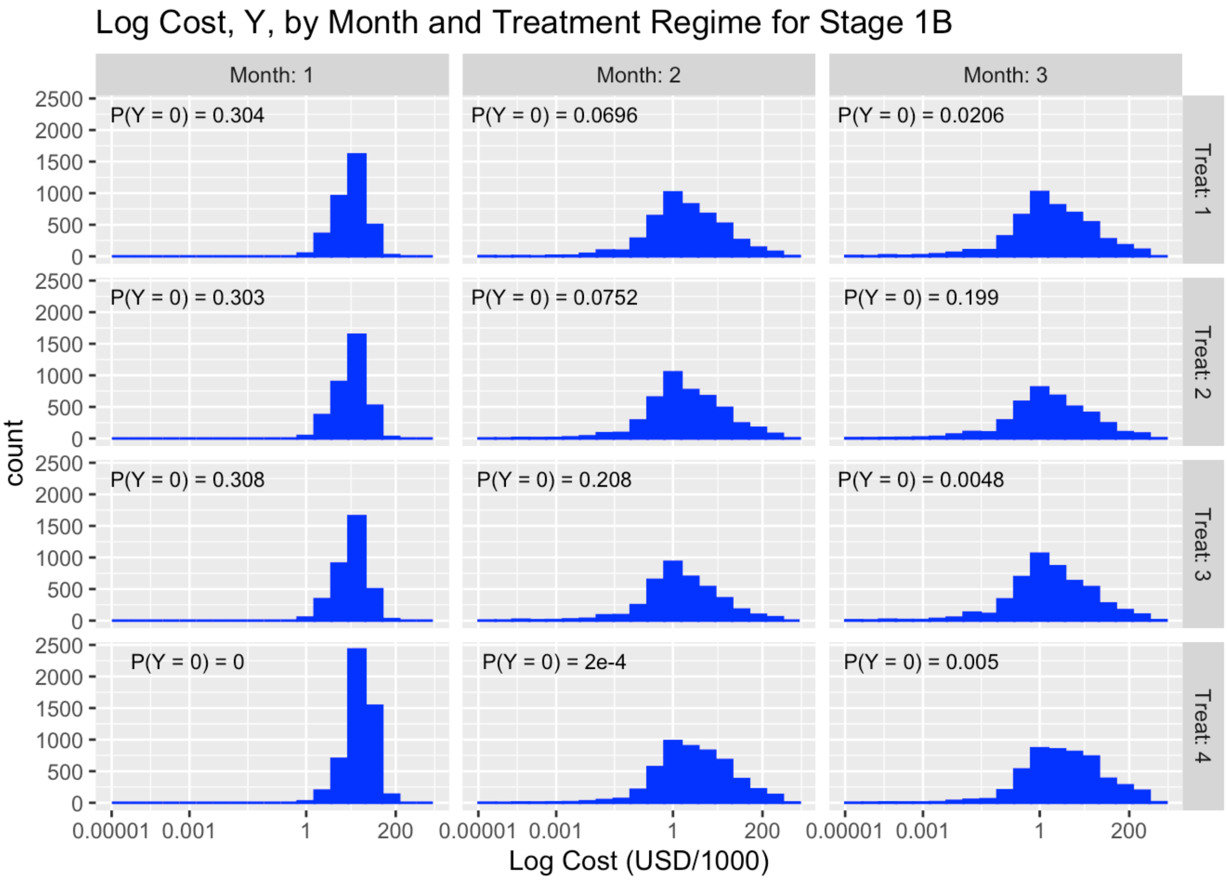

Figure 4.8: Distribution of non-zero log cost with exponentiated *x*-axis labels, stratified by month and treatment trajectory among patients with stage 1B cancer.

As was the case for stage 1A patients, the log cost distributions among stage 1B patients for treatments 1-3 within month 1 display a slight left-skewness and look similar (if not exactly the same). For treatment 4 within month 1, the distribution of log cost is also left-skewed but includes more of the higher log cost values. The reasoning for these observations for stage 1A cancer patients holds for stage 1B cancer patients.

The remaining distributions for stage 1B subjects, except for the distribution of log cost within month 3 for treatment regime 4, are right-skewed and look very similar. This may indicate that time and treatment do not influence log cost among patients with stage 1B cancer. However, further statistical investigation is necessary to make such a conclusion. The plot for month 3 for treatment regime 4 is slightly less skewed than the other plots at times 2 and 3.

Similar to stage 1A cancer patients, it is true that for stage 1B cancer patients that the probability of having a zero cost value decreases with increased treatment. One exception is the probability of a zero cost for treatment regime 2 within month 3. As with the stage 1A subjects, this time point is when treatment begins for this group, and it could be that the initial treatment period results in lower (more zero) costs.

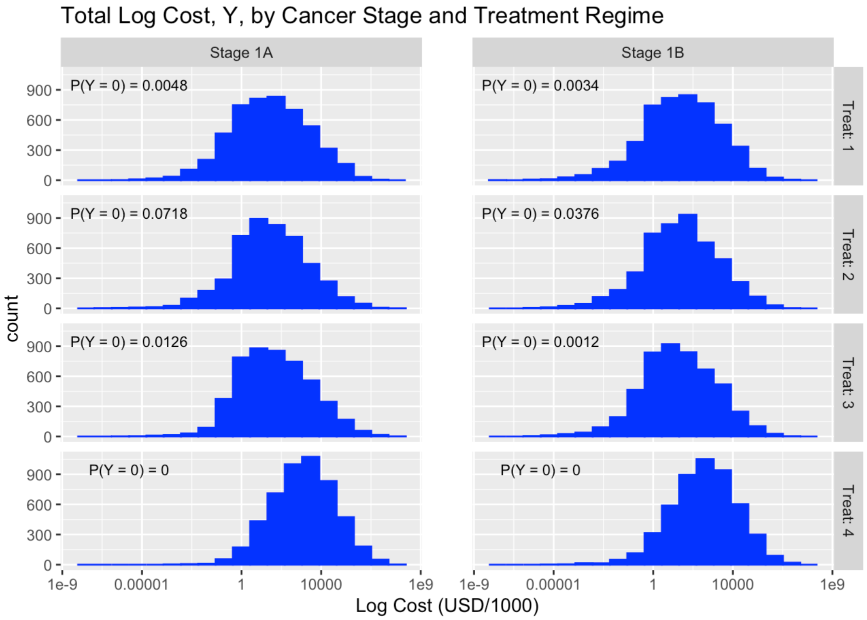

Figure 4.9: Distribution of non-zero overall log cost with exponentiated *x*-axis labels, stratified by cancer stage and treatment trajectory.

Figure 4.9 displays the total log cost for each treatment regime by cancer stage. Comparing distributions of overall accumulated log cost between cancer stages (see Figure 4.9), it appears as if these distributions are approximately the same within each treatment group. Thus, there is not a noticeable difference in total log cost between stage 1A and 1B cancer. The distributions of log cost within treatment regime 1, 2 and 4, for both cancer stages, are fairly symmetric. The log cost distribution under treatment regime 3, for both cancer stages, is right-skewed which means that this treatment results in more of the lower total log cost values.

The probability of having a zero cost is comparable between the two cancer stages for all the treatment regimes, with stage 1B patients having only a minimally smaller probability for treatment regimes 1-3. This indicates that, within a given treatment regime, cancer stage may not influence the probability of having a zero-cost value.

CHAPTER 5

DISCUSSION

Given the rising costs in health care in the United States and the difficulty in estimating the full treatment effect in a longitudinal study with time-varying confounders, it is critical that methods are developed to effectively measure these costs in such settings. While parametric g-computation is a suitable approach, the more flexible option of implementing CPMs within g-computation is shown to be viable, as well. From the simulation outlined in this paper, the resulting bias in estimating the mean potential outcome is extremely low, and the CPM distributions of the time-varying confounder and the outcome are nearly the same as the corresponding theoretical distributions. When this method is applied to fully simulated SEER-Medicare data, modifications are made to accommodate two time-varying confounders, additional baseline covariates, and longitudinal measurements of the outcome (instead of measuring just the final outcome). Regardless of these changes, the general framework of the CPM approach holds, and the mean potential outcome at each time point and overall can be estimated.

Future work on this project could entail determining the uncertainty of the mean potential outcome estimates by performing bootstrap replicates of the estimation procedure to determine the standard error of those estimates. Additionally, it only is known now that flexible g-computation is possible, but it is not clear how this approach compares to parametric g-computation in terms of efficiency. Determining this would entail repeating the estimation procedure with parametric models (such as multiple linear regression models) in place of the cumulative probability models and comparing the approaches based on bias, bootstrap standard error, coverage probability, and robustness to model misspecification.

Additional expansions of this research include applying the CPM approach to settings in which the time points are not evenly spaced, there is missing data, and other semiparametric models are considered. Furthermore, the process of programming flexible g-computation within statistical software is lengthy and complex. For this approach to be more easily implemented within applied collaborations, a package could be written for R to include functions that perform the steps which are coded manually for this project.

Continuing to explore flexible and varied methods of approaching common problems can only help to advance the efforts of modern research, allowing investigators to reach conclusions with less assumptions and, in turn, less room for error. When trying to understand a topic as imperative as rising medical costs, this endeavor is even more valuable. Exploring the CPM approach further should lead to significant contributions to longitudinal research.

BIBLIOGRAPHY

[1] Daniel, R., Cousens, S., Stavola, B. D., Kenward, M. G., & Sterne, J. A. (2012). Methods for dealing with time-dependent confounding. Statistics in Medicine, 32(9), 1584-1618.

[2] Robins, J. (1986). 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.

[3] Spieker, A. J., Ko, E. M., Roy, J. A., & Mitra, N. (2020). Nested g-computation: a causal approach to analysis of censored medical costs in the presence of time-varying treatment. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1189–1208.

[4] Liu, Q., Shepherd, B. E., Li, C., & Harrell, F. E. (2017). Modeling continuous response variables using ordinal regression. *Statistics in Medicine, 36*(27), 4316-4335.

[5] Brooks, R. A., Fleming, G. F., Lastra, R. R., Lee, N. K., Moroney, J. W., Son, C. H., Tatebe, K., & Veneris, J. L. (2019). Current recommendations and recent progress in endometrial cancer. *CA: A Cancer Journal for Clinicians*.

[6] Moore, K., & Brewer, M. A. (2017). Endometrial Cancer: Is This a New Disease? American Society of Clinical Oncology Educational Book, 37, 435–442.

[7] Papanicolas, I., Woskie, L. R., & Jha, A. K. (2018). Health Care Spending in the United States and Other High-Income Countries. JAMA, 319(10), 1024.

[8] Dieleman, J. L., Cao, J., Chapin, A., Chen, C., Li, Z., Liu, A., Horst, C., Kaldjian, A., Matyasz, T., Scott, K. W., Bui, A. L., Campbell, M., Duber, H. C., Dunn, A. C., Flaxman, A. D., Fitzmaurice, C., Naghavi, M., Sadat, N., Shieh, P., … Murray, C. J. (2020). US Health Care Spending by Payer and Health Condition, 1996-2016. *JAMA*, *323*(9), 863.

[9] Ambs, A., Warren, J. L., Bellizzi, K. M., Topor, M., Haffer, S. C., & Clauser, S. B. (2008). Overview of the SEER—Medicare Health Outcomes Survey Linked Dataset. *Health Care Financ Rev.,* 29(4), 5–21.

[10] Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

APPENDIX

The R code to perform this project's simulation and application can be found in a public repository at the following link: https://github.com/BirdrowC/Semiparametric-G-Computation.