QUALITY ADJUSTED Q-LEARNING AND CONDITIONAL STRUCTURAL MEAN MODELS FOR OPTIMIZING DYNAMIC TREATMENT REGIMES

by

Geoffrey S Johnson

M.S. Statistics, George Mason University, 2011

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Geoffrey S Johnson

It was presented on

May 4th, 2016

to

Abdus S. Wahed, Phd
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Joyce Chang, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Jong-Hyeon Jeong, Phd
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Yu Cheng, Phd
Associate Professor
Department of Statistics
Dietrich School of Arts and Sciences
University of Pittsburgh

Dissertation Director: Abdus S. Wahed, Phd
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

QUALITY ADJUSTED Q-LEARNING AND CONDITIONAL STRUCTURAL MEAN MODELS FOR OPTIMIZING DYNAMIC TREATMENT REGIMES

Geoffrey S Johnson, PhD

University of Pittsburgh, 2016

ABSTRACT

The focus of this work is to investigate a form of Q-learning using estimating equations for quality adjusted survival time, and to generalize these methods to quality adjust other outcomes. We use the m-out-of-n bootstrap and threshold utility analysis to show how the patient-specific optimal regime varies according to treatment characteristics (e.g. cost, side effects). Methodologies investigated are demonstrated to construct optimal treatment regimes for the treatment of children's neuroblastoma. We also propose a new method for optimizing dynamic treatment regimes using conditional structural mean models. The inverse-probability-of-treatment weighted (IPTW) or g-computation estimator is used at each stage to estimate what we call the 'preliminary' optimal treatment regime, given patient information up to the current stage and prior treatment assignment. Essentially this tailors the optimal treatment assignment at the current stage, and provides an optimal strategy for the remaining stages given the information currently available. We compare this method for optimizing a dynamic treatment regime to Q-learning. Additionally, we propose a two step prescriptive variable selection procedure that supports the tailored optimization of dynamic treatment regimes using conditional structural mean models by eliminating from consideration any suboptimal treatment regimes and sifting out the covariates that prescribe the optimal treatment regimes. The methods described herein are meant to advance the field of dynamic treatment regimes, a field that has a substantial impact on public health. The

treatment policies that come from DTRs, whether determined for the population as a whole or tailored for specific subgroups, can be used to guide and shape health policies that will ultimately lead to greater public health and safety.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my dissertation advisor, supervisor, professor, and mentor, Dr. Abdus S. Wahed. His guidance has shaped my understanding of statistical theory, and my skill in application. The opportunity to work with him and to earn my PhD has completely altered my career path, and will forever change my life. I would also like to thank Dr. Clifton D. Sutton for molding my foundation in statistical theory. My continued success rests on his training. This dissertation is dedicated to my parents, sister, wife, and son.

1.0 INTRODUCTION

Dynamic treatment regimes (DTRs) provide the basis for statistical analysis in personalized medicine. A DTR is a decision rule that guides the treatment choices over the course of therapy. The sequence of treatments a patient receives depends on the patient's health status, response to prior treatments, and other patient characteristics [22, 28, 5]. The goal is to find a DTR that optimizes the overall outcome, commonly taken as overall survival in cancer studies [34, 33, 18]. Techniques for analyzing DTRs are important to properly account for patient responses and sequence of treatments to correctly identify the optimal treatment at each stage. Consider a game of chess. Each player's turn corresponds to a stage of a DTR, each player's move corresponds to a treatment assignment, and achieving check mate corresponds to optimizing the outcome. The player's best move in each turn depends on his previous and future moves. If the chess player optimizes each move individually, without regard to past or future moves, he/she will likely not achieve check mate. If the treatments at each stage of a DTR are analyzed without regard to past and future treatments, biased results may occur. Assuming larger outcomes are better, it is natural to search for the optimal regime, the one with the largest expected outcome. To this end there are primarily two approaches: structural mean models (direct search) and nested mean models (inductive search).

Structural mean models use weighting techniques found in survey sampling to estimate the mean outcome of a regime had everyone sampled followed that regime, allowing a direct comparison of the outcome across DTRs. Inverse probability of treatment weighted (IPTW) estimators, as their name suggests, average the outcome for all subjects following a specific regime, while weighting each observation by the inverse of the probability of receiving the treatments prescribed by the regime, similar to the Horvitz-Thompson estimator

[17]. Alternatively, g-computation estimators first find the mean outcome for each path of a particular regime, and weight these means by the proportion who followed each path to form the mean of the regime [25, 27]. Conversely, nested mean models use backwards induction to find the optimal treatment at each stage. Murhpy (2003) [22], Robins (2004) [26], and others pioneered the use of backwards induction in statistics via Q-learning and g-estimation to identify such optimal regimes. These algorithms work backwards in time by identifying at each stage which treatment has the largest expected outcome, and creating pseudo data for each subject by replacing his/her observed outcomes with the estimated optimal expected outcome at each stage, given prior observed outcomes and covariate information. The optimal regime is the one with the largest expected value of this pseudo data.

Generally, to build the models that identify the optimal treatment regime clinical trials are designed, patients are recruited, and their information is gathered. These patients, though treated, do not get to benefit from the knowledge they collectively bring. It is the next patient in line, the prospective patient, that can use this information to make a more informed treatment decision. From the clinician's perspective all that is required is to collect demographic, blood marker, and genetic information, and the optimal treatment regime can be assigned for the prospective patient. The prospective patient, however, has an entirely different decision making process. He/she builds a nearly infinite dimensional model in his/her subconscious when choosing between treatments. Is the treatment painful? What are the side effects? How much does it cost? Does the treatment conflict with my ethical or religious beliefs? What is the treatment schedule and can I adhere to it? And so on. For each prospective patient the questions and answers are different. From this prospective patient's point of view, he/she is not here to offer a few data points, he/she is here to decide which treatment is best for him/her. To capture the decision making process of the prospective patient in the survival analysis setting, we investigate quality adjusted lifetime as the outcome in dynamic treatment regimes via Q-learning, and while this outcome is not specific to Q-learning, nor to dynamic treatment regimes, it is especially pertinent. This approach offers a glimpse into the mind of the prospective patient's decision making process, and allows us to see just how averse the patient must be to a particular treatment, or regime, before it is no longer optimal.

Clinical trials for cancer often measure a primary outcome and several secondary outcomes. The secondary outcomes may include, among others, measures of toxicity and adherence. Taken separately, these measures may sometimes lead to different optimal treatments. While one treatment may have the largest expected primary outcome, a second one may be less toxic, and a third might have the best adherence. Gelber et al. (1989) [10], Glasziou et al. (1989) [11], GoldHirsch et al. (1989) [12] and Korn (1993) [19] considered quality adjusted lifetime to adjust the length of life based on its quality. In its simplest form, quality adjusted life assigns a utility weight, ranging from 0 (death) to 1(perfect health), to separate states of health. If there are k health states, then $U_i = \sum_{j=1}^{k} q_j s_{ji}$ is the quality adjusted lifetime (QAL) for the i^{th} patient, where $s_{1i}, .., s_{ki}$ are the times spent in each state, and $q_1, ..., q_k$ are the utility coefficients assigned to each of the health states. Note that the quality adjusted lifetime U_i is simply a fraction of total lifetime for patient i. More recently, Zhao and Tsiatis [38, 39, 40, 41] have provided consistent and efficient estimators, and provided hypothesis tests for distributional features of quality adjusted lifetime in the presence of right censoring. Wang & Zhao (2007) [35] extended this work to the regression setting, using inverse weighting techniques to form consistent estimating equations for regression parameters.

The first goal of this dissertation is to develop an optimal dynamic treatment regime to maximize quality adjusted lifetime by using a Q-learning-type approach discussed in [18]. This method will be operationalized using the estimating equations of [35], and a threshold utility analysis will be used to show how the subject-specific optimal DTR not only depends on patient history and intermediate outcomes, but also on quality of life, monetary cost, and other factors during each treatment. Though we optimize quality adjusted lifetime, we provide suggestions on how the quality adjustment can be used for any continuous outcome via Q-learning. The utility weights capture the secondary outcomes as well as the unmeasurable decision making process of the prospective patient and discount the expected utility of treatments. We use a simulation study to evaluate these methods, and then apply them to COG study A3891 concerning 379 children receiving treatment for high-risk neuroblastoma [21].

Dynamic treatment regimes are a function of patient data, and as ever larger studies collect more patient data, it is natural to turn to variable selection methods when searching for the optimal regime. Common approaches to variable selection include, but are not limited to, the forward, fackward, and ftep-wise selection methods, which by their nature are discrete processes, and the least absolute shrinkage and selection operator (LASSO) and its derivatives, which are continuous processes. To operate, all of these methods rely on a measure of model fit or prediction error, such as the sum of squared errors, the leave-one-out cross-validation estimate of prediction error, or Akaike information criterion (AIC). These variable selection methods are designed to sift through a large collection of variables and identify those that most greatly reduce the variability and increase the accuracy of the estimator, which Gunter et al. (2011) [13] define as *predictive* variables. However, in the realm of dynamic treatment regimes, we are interested in variables that are not only predictive, but also help prescribe the optimal treatment for a given patient. Such variables are called *prescriptive* [15], and must qualitatively interact with treatment. For a nested mean model approach, Gunter et al. (2011) [13] propose two different ranking methods to sort variables according to how likely they will qualitatively interact with the outcome, and provide a four step algorithm involving LASSO regression on nested subsets of covariates for selecting important predictive variables. Zhang (2014) [37] generalizes from the least squares regression model and offers a simpler, more effective two step method involving Multivariate Adaptive Regression Splines (MARS) models and logistic regression with LASSO.

Most authors employing structural mean models perform a marginal analysis, comparing dynamic treatment regimes for the entire sample of patients [33]. Those that perform a subgroup analysis using conditional models do so by conditioning on baseline information only [14, 5]. While these conditional models shed some light on the regime effects across baseline covariates, they lack the ability of Q-learning and other backwards induction techniques to use past and current patient information to prescribe the optimal treatment at each stage. The second goal of this dissertation is two fold: i) to propose a new method for optimizing dynamic treatment regimes using conditional structural mean models that incorporates

current patient information at every stage (decision point), ii) and to provide an effective prescriptive variable selection method for these conditional structural mean models. The method in Zhang (2014) [37] for nested mean models is reviewed, and extended to structural mean models. We use a simulation study to evaluate these methods, and apply them to a phase II study concerning 215 patients with acute myeloid leukemia (AML) or high-risk myelodysplastic syndrome (MDS) [8].

2.0 QUALITY ADJUSTED Q-LEARNING AND THRESHOLD UTILITY ANALYSIS FOR OPTIMIZING DYNAMIC TREATMENT REGIMES

2.1 SETUP

2.1.1 Quality-adjusted lifetime

Describe the health history for the i^{th} patient with a continuous time stochastic process $\{V_i(t), t \geq 0\}$. $V_i(t)$ maps to the space of health states $S = \{0, 1, 2, ..., m\}$, where the state '0' corresponds to the absorbing state of death. Denote the health history up to time t by $V_i^H(t) = \{V_i(s) : s \leq t\}$. Let $V_i(s) = 0$ imply that $V_i(t) = 0$ for $t \geq s$. Let T_i denote the survival time for patient i. Naturally, $V_i(t) = 0$ for $t \geq T_i$. Then we see that $T_i = \inf\{t : V_i(t) = 0\}$. Let $q()$ be a quality of life function mapping $V_i(t)$ to $[0, 1]$, with $q(0) \equiv 0$. The quality adjusted lifetime for the i^{th} patient is defined as $Q(T_i) = \int_0^{T_i} q\{V_i(t)\}dt$.

In the presence of non-informative right censoring, one might consider the restricted survival time where total follow-up time is limited to L, where L is some value less than the maximum survival time for all patients. Therefore, the survival time for all patients will be truncated at L, $T^L = \min(T, L)$. For ease of notation, we will drop the superscript and simply use T. We will denote the i^{th} patient's censoring time by C_i, and the survival distribution of C by $K(t) = P(C > t)$. Define $U_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, respectively, to be the observed time to event (death or censoring), and the death indicator. Then $Q(U_i) = \int_0^{U_i} q\{V_i(t)\}dt$ represents the quality adjusted time to event for the i^{th} patient.

In this construction the quality function q is not patient specific (does not have a subscript i), and was assumed known. One view is that q exists at the population level. This means that every patient in the analysis, and all of the patients they represent, experience the same quality of life when in a particular health state. This allows for a threshold utility analysis, described in detail in Section 2.2.2, where quality adjusted lifetime (or a function of it) is considered over the entire range of possible values of q, to examine how the value of q affects the estimation of quality adjusted lifetime. As a convention we will take $Q(s,t)$ to refer to $\int_s^t q\{V(u)\}du$ and $Q(t)$ to refer to $\int_0^t q\{V(u)\}du$.

For example, consider a discrete-state health history process $V_i(t)$ with three states: treatment, response (well-being), and death. Suppose each of these states are mapped to [0,1] as $q\{V_i(t)\} = q_a I\{t \leq T_i^R\} + 1I\{T_i^R < t < T_i\} + 0I\{t > T_i\}$. Such a mapping may be reasonable as the quality is the least (zero) after death, one when healthy, and a constant, q_a, between zero and one when being treated due to toxicity related complications and/or monetary cost from receiving treatment $A = a$. Here, time from beginning of treatment to response is denoted by T_i^R. Under this scenario, $Q(T_i) = \int_0^{T_i^R} q_a dt + \int_{T_i^R}^{T_i} 1dt = T_i - (1 - q_a)T_i^R$. If the patient undergoes a maintenance treatment immediately after responding, and remains on maintenance treatment $B = b$ until death, $Q(T_i)$ could be written as $Q(T_i) = \int_0^{T_i^R} q_a dt + \int_{T_i^R}^{T_i} q_b dt = q_b T_i - (q_b - q_a)T_i^R$, where the constant q_b reflects the utility weight of treatment $B = b$ for toxicity, monetary cost, and other factors.

Since quality adjusted lifetime is the area under $q\{V_i(t)\}$ over the health states from 0 to T, for any function $q\{V_i(t)\}$ there exists a constant function in each health state that results in the same area, and produces the same quality adjusted lifetime. Not coincidentally, the example above has the health states of each patient correspond to the sequence of treatments received. When estimating mean quality adjusted lifetime in such settings, the utility weights q_a and q_b factor out, producing $E[Q(T_i)] = q_a E[T_i^R] + q_b E[T_i - T_i^R]$. When viewed in this way, not only can the utility weights be seen as population constants, they can alternatively be seen as adjustments to the expected utility of each treatment for the prospective patient, depending on his or her aversion to each treatment, with each prospec-

tive patient potentially having different values of the utility weights. Such an interpretation of the utility weights offers even more motivation for a threshold utility analysis.

For drawing inference on quality adjusted lifetime, the survival function of quality adjusted lifetime may be used the same way as as the survival function of overall survival. In the presence of non-informative censoring one might naturally turn to the Kaplan-Meier estimator, to estimate $S(t) = P\Big(Q(T_i) > t\Big)$, but Gelber et al. (1989) [10] and Pradhan & Dewanji (2009) [24] showed that this can result in biased estimation because the quality adjustment induces a dependence between the survival times and censoring times. Zhao & Tsiatis (1997) [38] offer an inverse-probability weighted estimator, similar to that proposed by Robins & Rotnitzky (1992) [29] and Robins et al. (1994) [30], $\hat{S}(t)^{cen} = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} I[Q(U_i) > t]$, where $\hat{K}(U_i)$ is the Kaplan-Meier estimator for the censoring random variable evaluated at U_i, and Δ_i and $\hat{K}(U_i)$ can depend on t to improve efficiency. Zhao & Tsiatis (1999) [39] improve the efficiency of their estimator by incorporating each patient's health history. In Zhao & Tsiatis (2000) [40] they used the same principles to estimate the mean quality adjusted lifetime.

Wang & Zhao (2007) [35] extended this work to the regression setting by constructing consistent estimating equations for mean quality adjusted lifetime in the presence of censoring, yielding $\mathcal{U}_n(\beta) = \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} h(X_i)\{Q(U_i) - g(\beta, X_i)\} = 0$, where X_i denotes a $(p+1)\times 1$ vector of covariates associated with patient i, with the first covariate being the constant 1, $h(X_i)$ is a $(p+1)\times 1$ vector of functions of X_i, β is a $(p+1)\times 1$ vector of parameters, and $g(\beta, X_i) = E[Q(T_i)|X_i]$. The estimator for β solving $\mathcal{U}_n(\beta)$ will be used to operationalize our search for the optimal dynamic treatment regime, described in Section 2.2.

2.1.2 Dynamic treatment regimes and corresponding terminology

Consider a two-stage sequential multiple assignment randomized trial (SMART) design where patients are randomized to one of two induction therapies, $\mathcal{A} = \{a_1, a_2\}$. Patients may be resistant to their initial treatment, or they may respond. For each of the induction therapies,

if treatment response is observed, patients are further randomized to one of two maintenance treatments, $\mathcal{B} = \{b_1, b_2\}$. This design allows for inference on four DTRs that might be carried out in clinical practice, namely, $d(A_i = a_j; B_i = b_k),\ j, k = 1, 2$, where $d(A_i; B_i)$ stands for "Treat with A_i, if the patient responds, treat with B_i." Our goal is to find the optimal treatment regime among these that maximizes expected quality adjusted lifetime.

Let $G_i^H(t)$ denote all information collected on patient i prior to time t. Some or all of the information in $G_i^H(t)$, for example serum biomarker levels, responses to questionnaires, or tumor size, is used to define $V_i^H(t)$, which then defines R_i and T_i^R, the observed response indicator and the observed time to response given $R_i = 1$, respectively. $G_i^H(t)$ may include additional patient information not used to define $V_i^H(t)$. Then, introducing further indicators for first and second stage treatment, the observed data for the i^{th} patient in the presence of censoring is written as

$$D_i^\delta = \Big(Z_{1i}^{(A)},\ Z_{2i}^{(A)},\ R_i,\ R_i T_i^R,\ R_i Z_{1i}^{(B)},\ R_i Z_{2i}^{(B)},\ U_i,\ \Delta_i,\ V_i^H(U_i),\ G_i^H(U_i)\Big),$$

where $Z_{ji}^{(A)}$=1 if patient i received the j^{th} induction therapy, $Z_{ji}^{(A)}$=0 otherwise, and $Z_{ki}^{(B)}$ denotes the b_k treatment assignment indicator $I\{B = b_k\}$, defined only if R_i=1. Note that $Z_{2i}^{(A)} = 1 - Z_{1i}^{(A)}$ and $Z_{2i}^{(B)} = 1 - Z_{1i}^{(B)}$, but we explicitly define them to facilitate the use of summation.

By design, treatments are assigned independently of prognosis or any observed data measured prior to the second stage. This condition is often referred to as no unmeasured confounders or sequential randomization assumption. This 'no unmeasured confounders' condition holds even if the second-stage randomization probabilities depend on the first-stage treatment assignments.

2.2 OPTIMIZATION OF DYNAMIC TREATMENT REGIMES ON QUALITY ADJUSTED SURVIVAL

2.2.1 Optimization

Following the work of Murphy (2003) [22], Robins (2004) [26], and Huang et al. (2014) [18], we describe a backward induction method to identify the optimal dynamic treatment regime, using mean quality adjusted survival time as the criterion of optimality. From the reinforcement learning literature in the field of DTRs, the typical $\mathcal{Q}$-functions for two stages of our SMART design, assuming no unmeasured confounders, would be

$$\begin{aligned} \mathcal{Q}_B\Big(A_i = a_j, G_i^H(T_i^R), B_i = b_k\Big) &= E\Big[Q(T_i^R, T_i)\Big|A_i = a_j, R_i = 1, G_i^H(T_i^R), B_i = b_k\Big] \\ \mathcal{Q}_A\Big(G_i^H(0), A_i = a_j\Big) &= E\Big[H_i^{(A)}\Big|G_i^H(0), A_i = a_j\Big], \end{aligned}$$

where

$$H_i^{(A)} = \begin{cases} Q(T_i^R) + \underset{b_k}{max}\ \mathcal{Q}_B\Big(A_i, G_i^H(T_i^R), B_i = b_k\Big), & \text{if } R_i = 1 \\ Q(T_i), & \text{if } R_i = 0. \end{cases}$$

Then, the optimal stage 1 treatment given baseline information is

$$A_i^{opt} = \underset{a_k}{argmax}\ E\Big[H_i^{(A)}\Big|G_i^H(0), A_i = a_j\Big],$$

and the optimal stage 2 treatment given stage 1 treatment assignment and information up to stage 2 is

$$B_i^{opt} = \underset{b_k}{argmax}\ E\Big[Q(T_i^R, T_i)\Big|A_i = a_j, R_i = 1, G_i^H(T_i^R), B_i = b_k\Big].$$

Below we walk through the backwards induction used to estimate the optimal treatment at each stage, with a different $H_i^{(A)}$ shown in Huang et al. (2014) that we use in our simulation and application.

We start with the second stage (include only those patients who responded $R_i = 1$). Under assumptions described in Section 2.1.2, the quality adjusted time from maintenance

therapy to death for those patients who responded is $Q(T_i^R, T_i) = \int_{T_i^R}^{T_i} q\{V_i(t)\}dt$, so that

$$\begin{aligned}\gamma_B \equiv\ & E\left[\ Q(T_i^R, T_i)\ \middle| A_i = a_j, B_i = b_1, R_i = 1, G_i^H(T_i^R)\right] \\ & -E\left[\ Q(T_i^R, T_i)\ \middle| A_i = a_j, B_i = b_2, R_i = 1, G_i^H(T_i^R)\right]\end{aligned}$$

is the difference in expected stage 2 outcomes, given prior information. We assume the following linear model for $\mathcal{Q}_B\Big(A_i, B_i, R_i = 1, \bar{X}_{Bi}, \boldsymbol{\beta}_B, \boldsymbol{\alpha}_B\Big)$

$$E\Big[Q(T_i^R, T_i)\Big| A_i, B_i, R_i = 1, \bar{X}_{Bi}, \boldsymbol{\beta}_B, \boldsymbol{\alpha}_B\Big] = \bar{X}_{Bi}'\boldsymbol{\beta}_B + Z_{1i}^{(B)}\bar{X}_{Bi}'\boldsymbol{\alpha}_B, \tag{2.1}$$

where $\bar{X}_{Bi}$ are the first stage treatment assignment indicators and covariates from $G_i^H(T_i^R)$, and includes an element equal to 1 corresponding to an intercept term, which implies that $\gamma_B = \bar{X}_{Bi}'\boldsymbol{\alpha}_B$, and the estimated optimal stage two treatment given stage 1 treatment assignment and patient information up to stage 2 is

$$\hat{B}^{opt}(\bar{X}_{Bi}) = \underset{b_k}{argmax}\ \hat{E}\left[\ Q(T_i^R, T_i)\ \middle| A_i = a_j, R_i = 1, B_i = b_k, \bar{X}_{Bi}, \boldsymbol{\beta}_B, \boldsymbol{\alpha}_B\right].$$

If γ_B is positive then b_1 is the optimal stage 2 treatment, otherwise, b_2 is optimal. Using fitted models corresponding to equation (2.1) we can estimate the optimal quality adjusted time from maintenance therapy to death as

$$H_i^{(B)}(\hat{\boldsymbol{\alpha}}_B) \equiv \begin{cases} Q(T_i^R, T_i) + \left|\bar{X}_{Bi}'\hat{\boldsymbol{\alpha}}_B\right|, & \text{if } B_i = b_k,\ \hat{B}_i^{opt} \neq b_k \\ Q(T_i^R, T_i), & \text{if } B_i = b_k,\ \hat{B}_i^{opt} = b_k. \end{cases}$$

Moving to the first stage, under assumptions described in Section 2.1.2 the quality adjusted survival time with observed stage one treatment and the estimated optimal stage two treatment can be written as

$$H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B) = \begin{cases} Q(T_i^R) + H_i^{(B)}(\hat{\boldsymbol{\alpha}}_B), & \text{if } R_i = 1 \\ Q(T_i), & \text{if } R_i = 0. \end{cases}$$

Let X_{Ai} denote important covariates in $G_i^H(0)$ predictive of residual survival. We assume

the following linear model for $\mathcal{Q}_A\Big(A_i, X_{Ai}, \boldsymbol{\beta}_A, \boldsymbol{\alpha}_A\Big)$

$$E\Big[H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B)\Big|A_i, X_{Ai}, \boldsymbol{\beta}_A, \boldsymbol{\alpha}_A\Big] = X_{Ai}'\boldsymbol{\beta}_A + Z_{1i}^{(A)}X_{Ai}'\boldsymbol{\alpha}_A,$$

where X_{Ai} includes an element equal to 1 corresponding to an intercept term, which implies that

$$\gamma_A \equiv E\Big[H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B)\Big|A_i = a_1, G_i^H(0)\Big] - E\Big[H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B)\Big|A_i = a_2, G_i^H(0)\Big] = X_{Ai}'\boldsymbol{\alpha}_A$$

is the difference in expected outcomes at stage 1, given that each patient received his estimated optimal stage 2 treatment. The estimated optimal stage one treatment is

$$\hat{A}^{opt}(X_{Ai}) = \underset{a_j}{argmax}\ \hat{E}\left[H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B)\middle|A_i = a_j, X_{Ai}, \boldsymbol{\beta}_A, \boldsymbol{\alpha}_A\right].$$

If γ_A is positive then a_1 is the optimal stage 1 treatment, otherwise, a_2 is optimal. Thus if one could estimate the quantities γ_A or γ_B, or equivalently, the parameters α_B and α_A, the optimal treatment regime could be constructed given the q function for each specific stage.

To estimate these parameters, the simple weighted regression models described in Section 2.1.1 by Wang & Zhao (2007) [35] can be used. Explicitly, for stage 2 we solve the estimating equation

$$\mathcal{U}_n^{(B)}(\boldsymbol{\beta}_B, \boldsymbol{\alpha}_B) = \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} R_i \begin{bmatrix} \bar{X}_{Bi} \\ Z_{1i}^{(B)}\bar{X}_{Bi} \end{bmatrix} \Big\{Q(T_i^R, U_i) - \bar{X}_{Bi}'\boldsymbol{\beta}_B - Z_{1i}^{(B)}\bar{X}_{Bi}'\boldsymbol{\alpha}_B\Big\} = 0,$$

for $\boldsymbol{\beta}_B$ and $\boldsymbol{\alpha}_B$. Similarly, for stage 1 we solve

$$\mathcal{U}_n^{(A)}(\boldsymbol{\beta}_A, \boldsymbol{\alpha}_A) = \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} \begin{bmatrix} X_{Ai} \\ Z_{1i}^{(A)}X_{Ai} \end{bmatrix} \Big\{H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B) - X_{Ai}'\boldsymbol{\beta}_A - Z_{1i}^{(A)}X_{Ai}'\boldsymbol{\alpha}_A\Big\} = 0$$

to obtain estimates of $\boldsymbol{\beta}_A$ and $\boldsymbol{\alpha}_A$.

2.2.2 Threshold Utility Analysis

Glasziou et al. (1990) [11] perform a threshold utility analysis when studying the effects of adjuvant chemotherapy on quality adjusted lifetime in patients with early breast cancer.

Each patient's survival time is quality adjusted based on periods of toxicity of treatment and relapse of disease. These quality weights, ranging from 0 to 1, are plotted against each other and the regions where each treatment is favored are identified via lines (planes) of indifference. This results in a type of sensitivity analysis, allowing one to see all possible treatment decisions drawn depending on the quality weights. In our DTR setting, a patient's course of treatment often depends on his/her state of health, be it response to treatment or relapse of the disease, so that his/her health states correspond to the stages of the DTR. In our approach, each patient's survival time will be weighted according to treatment received, allowing a threshold utility analysis among treatments, and ultimately among regimes.

Optimal decision rules for first and second stage treatments developed in Section 2.2 are not only a function of the observed data (patient level information), but also of the quality of life function q. In our development in the previous section, we assumed that this q function was known, and we offered two interpretations of its meaning. Rather than performing a single analysis with one q function, a sensitivity analysis can be performed using a variety of reasonable q functions to determine for which functions of q, if any, the choice of optimal regime changes. In the special case of constant q functions, q can be varied from 0 to 1, and a threshold utility plane can be plotted. This is of importance, since depending on the values of the q function, there may be different optimal treatment regimes.

To be explicit, consider quality adjusting each patient's survival time as

$$Q(T_i) = \begin{cases} T_i q_{a_j} & \text{, if } A_i = a_j,\ R_i = 0, \\ T_i^R q_{a_j} + (T_i - T_i^R) q_{b_k} & \text{, if } A_i = a_j,\ R_i = 1,\ B_i = b_k, \end{cases}$$

for $j = 1, 2$, $k = 1, 2$ where $q_{a_j}, q_{b_k} \in [0, 1]$. For those who responded ($R_i = 1$) and received maintenance treatment, the quality weights $q_{b_1}, q_{b_2} \in [0, 1]$ can be plotted against each other

on the x and y axes, with

$$\begin{aligned}\hat{\gamma}_B &= \hat{E}\left[(T_i - T_i^R)q_{b_1} \,\middle|\, A_i = a_j, B_i = b_1, R_i = 1, \bar{X}_{Bi}, \boldsymbol{\beta}_B, \boldsymbol{\alpha}_B \right] \\ &\quad -\hat{E}\left[(T_i - T_i^R)q_{b_2} \,\middle|\, A_i = a_j, B_i = b_2, R_i = 1, \bar{X}_{Bi}, \boldsymbol{\beta}_B, \boldsymbol{\alpha}_B \right] \\ &= \bar{X}_{Bi}^{'}\hat{\boldsymbol{\alpha}}_B\end{aligned}$$

from Section 2.2 plotted on the z axis. This forms a two-dimensional plane in a three-dimensional space. When quality adjusting in this way, the utility weights q_{b_1} and q_{b_2} factor out of the expectations and can be viewed as adjustments to the expected utility of each stage two treatment for the prospective patient, depending on his or her aversion to each treatment. The line where $\hat{\gamma}_B = 0$ is the estimated threshold at which the expected utility of b_1 and b_2 are equal, where the prospective patient is indifferent when choosing between stage two treatments.

Similarly, for those who received an induction treatment, the quality weights $q_{a_1}, q_{a_2} \in [0, 1]$ can be plotted against each other on the x and y axes, with

$$\hat{\gamma}_A = \hat{E}\Big[H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B)\Big|A_i = a_1, X_{A_i}, \boldsymbol{\beta}_A, \boldsymbol{\alpha}_A\Big] - \hat{E}\Big[H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B)\Big|A_i = a_2, X_{A_i}, \boldsymbol{\beta}_A, \boldsymbol{\alpha}_A\Big] = X_{Ai}^{'}\hat{\boldsymbol{\alpha}}_A$$

from Section 2.2 plotted on the z axis, where

$$H_i^{(A)}(\hat{\boldsymbol{\alpha}}_B) = \begin{cases} T_i^R q_{a_j} + H_i^{(B)}(\hat{\boldsymbol{\alpha}}_B), & \text{if } A_i = a_j,\ R_i = 1 \\ T_i q_{a_j}, & \text{if } A_i = a_j,\ R_i = 0 \end{cases}$$

$$H_i^{(B)}(\hat{\boldsymbol{\alpha}}_B) = \begin{cases} (T_i - T_i^R)q_{b_k} + \left|\bar{X}_{Bi}^{'}\hat{\boldsymbol{\alpha}}_B\right|, & \text{if } B_i = b_k,\ \hat{B}_i^{opt} \neq b_k \\ (T_i - T_i^R)q_{b_k}, & \text{if } B_i = b_k,\ \hat{B}_i^{opt} = b_k. \end{cases}$$

The line where $\hat{\gamma}_A = 0$ is the estimated threshold at which the prospective patient is indifferent when choosing between a_1 and a_2.

2.2.3 Inference

Robins (2004) [26], Chakraborty et al. (2009) [6], and Laber et al. (2014) [20] are quick to point out that the estimators derived from Q-learning have non-regular limiting distributions, because the estimated stage 1 pseudo data (and hence the estimated stage 1 model parameters) are a non-smooth (non-differentiable at $\bar{X}'_{Bi}\hat{\boldsymbol{\alpha}}_B=0$) function of $\hat{\boldsymbol{\alpha}}_B$. This motivated Chakraborty et al. (2013) [4] to discuss the m-out-of-n bootstrap in the context of DTRs, in place of standard large-sample inference methods. The m-out-of-n bootstrap technique essentially smooths the empirical distribution function, with more smoothing corresponding to smaller values of m, the resample size, by allowing the empirical distribution function to tend to its limiting distribution at a faster rate than the bootstrap empirical distribution tends to the empirical distribution. We use this technique to create confidence regions in the threshold utility analysis, identifying regions of indifference and strong acceptance when choosing between stage 1 and stage 2 treatments. While Chakraborty et al. (2013) [4] provide several data driven methods for determining the smaller resample size m, we find a suitable m through simulation and apply this same m in the analysis of real data.

2.3 SIMULATION STUDY

In this section we conduct a simulation experiment to evaluate the optimization of dynamic treatment regimes for quality adjusted lifetime described in Section 2.2. Similar to the COG study A3891 that will be presented later in Section 2.4, we consider a 2-stage SMART design.

We generated 5,000 simulations with sample size n=1000. Patients are randomized to one of two induction therapies with probability one-half, and the probability of non-response for each induction therapy is the same, 0.55. Those who respond to induction therapy are further re-randomized with probability one-half to one of two maintenance therapies. Sojourn times to response and/or death were generated from various exponential distributions.

Table 2.1 shows the coverage probabilities over the 5,000 simulations for 90% point-wise bootstrap confidence intervals for the estimated difference in mean quality adjusted lifetime between Stage 1 and Stage 2 treatments (γ_A and γ_B, respectively) when searching for the optimal treatment regime using the simple weighted estimating equations from Section 2.1.1. The 5^{th} and 95^{th} percentiles of the bootstrapped sampling distributions are used to create the confidence intervals. Stage 1 coverage probabilities are estimated at $q_{b_1} = 0.8$ and $q_{b_2} = 0.6$. A variety of re-sample sizes were considered for the stage 1 m-out-of-n bootstrap, and m=850 produced confidence intervals maintaining the nominal coverage probability. The coverage probabilities for the 90% confidence intervals are close to the nominal level for utility weights that are away from zero. This makes sense, as a value of q close to zero greatly reduces the variability in the data, making it difficult to estimate the respective quantities. For some combinations of q_{b_1} and q_{b_2} the estimated stage 2 coverage probabilities are below the nominal level. Although no irregularity issues exist for the stage 2 estimates, the m-out-of-n bootstrap was still employed to improve the coverage probabilities, with m=800. Using the m-out-of-n bootstrap, the stage 1 coverage probabilities for the difference in mean quality adjusted lifetime are well maintained. Similar simulations were performed with a sample size of n=300 and survival times close to that of the COG study A3891. This gives us an idea of an appropriate choice of m. We found that m=240 and m=255 worked well for maintaining the nominal coverage probabilities for stage 2 and stage 1, respectively.

Table 2.1: Coverage probabilities of 90% point-wise bootstrap confidence intervals (500 bootstrap samples), from simulated data with 5000 replicates of n=1000, stage 2 m=800, stage 1 m=850.

$A = a_1 : B = b_1$ vs $B = b_2$

$q_{b_1} \backslash q_{b_2}$	0.0	0.2	0.4	0.6	0.8	1.0
0.00		0.822	0.822	0.822	0.822	0.822
0.20	0.904	0.916	0.831	0.817	0.812	0.813
0.40	0.904	0.927	0.916	0.872	0.831	0.819
0.60	0.904	0.919	0.930	0.916	0.886	0.855
0.80	0.904	0.915	0.927	0.928	0.916	0.894
1.00	0.904	0.913	0.920	0.928	0.927	0.916

$A = a_2 : B = b_1$ vs $B = b_2$

$q_{b_1} \backslash q_{b_2}$	0.0	0.2	0.4	0.6	0.8	1.0
0.00		0.929	0.929	0.929	0.929	0.929
0.20	0.919	0.901	0.919	0.928	0.929	0.930
0.40	0.919	0.909	0.901	0.912	0.919	0.923
0.60	0.919	0.913	0.903	0.901	0.904	0.916
0.80	0.919	0.914	0.909	0.902	0.901	0.903
1.00	0.919	0.916	0.910	0.907	0.901	0.901

$A = a_1$ vs $A = a_2$

$q_{a_1} \backslash q_{a_2}$	0.0	0.2	0.4	0.6	0.8	1.0
0.00	0.913	0.913	0.915	915	0.916	0.911
0.20	0.912	0.912	0.913	0.913	0.916	0.916
0.40	0.910	0.911	0.912	0.914	0.913	0.917
0.60	0.907	0.909	0.912	0.914	0.917	0.917
0.80	0.909	0.909	0.912	0.914	0.915	0.918
1.00	0.907	0.907	0.911	0.913	0.916	0.918

Table 2.2: Coverage probabilities of 90% point-wise bootstrap confidence intervals (500 bootstrap samples), from simulated data with 5000 replicates of n=2000, stage 2 m=1600, stage 1 m=1700.

$A = a_1 : B = b_1$ vs $B = b_2$

$q_{b_1} \backslash q_{b_2}$	0.0	0.2	0.4	0.6	0.8	1.0
0.00		0.702	0.702	0.702	0.702	0.702
0.20	0.859	0.924	0.734	0.690	0.687	0.687
0.40	0.859	0.914	0.924	0.813	0.734	0.702
0.60	0.859	0.897	0.927	0.924	0.857	0.780
0.80	0.859	0.884	0.914	0.932	0.924	0.878
1.00	0.859	0.878	0.904	0.923	0.934	0.924

$A = a_2 : B = b_1$ vs $B = b_2$

$q_{b_1} \backslash q_{b_2}$	0.0	0.2	0.4	0.6	0.8	1.0
0.00		0.933	0.933	0.933	0.933	0.933
0.20	0.914	0.877	0.915	0.927	0.930	0.933
0.40	0.914	0.896	0.877	0.901	0.915	0.923
0.60	0.914	0.901	0.890	0.877	0.891	0.910
0.80	0.914	0.907	0.896	0.885	0.877	0.886
1.00	0.914	0.907	0.889	0.892	0.884	0.877

$A = a_1$ vs $A = a_2$

$q_{a_1} \backslash q_{a_2}$	0.0	0.2	0.4	0.6	0.8	1.0
0.00	0.915	0.918	0.918	0.915	0.912	0.908
0.20	0.910	0.913	0.916	0.917	0.915	0.911
0.40	0.905	0.910	0.911	0.914	0.914	0.915
0.60	0.905	0.908	0.911	0.911	0.916	0.914
0.80	0.903	0.904	0.908	0.912	0.913	0.915
1.00	0.900	0.901	0.905	0.910	0.912	0.914

Figure 2.1 shows the true (left column) and estimated (right column) threshold utility planes for the simulated scenario with n=300. The estimated threshold utility planes are for a single simulated data set. For each combination of q_{b_1} and q_{b_2}, or q_{a_1} and q_{a_2}, the estimated difference in mean quality adjusted lifetime is plotted. The yellow and green represent the region of strong acceptance for choosing between b_1 and b_2, or a_1 and a_2, respectively. The purple and red near the center of the plane have 90% point-wise bootstrap confidence intervals that cover zero and represent the region of indifference when choosing between b_1 and b_2, or a_1 and a_2. We see that for the estimated threshold utility planes, the estimated line of indifference does not correspond exactly with the true line of indifference, yet the 90% confidence region does contain the true line. These threshold utility planes allow us to visualize how the optimal regime changes depending on the values of q_{b_1}, q_{b_2}, q_{a_1}, and q_{a_2}. For example, assume that the threshold utility planes presented on the right panel of Figure 2.1 are the planes computed from the observed data. If for these treatments q_{a_1}=0.8, q_{a_2}=0.5, q_{b_1}=0.7, and q_{b_2}=0.5, then the estimated optimal regime is $d(A = a_1; B = b_1)$. However, if q_{a_1}=0.3, q_{a_2}=0.8, q_{b_1}=0.4, and q_{b_2}=0.6, the estimated optimal regime would be $d(A = a_2; B = b_2)$.

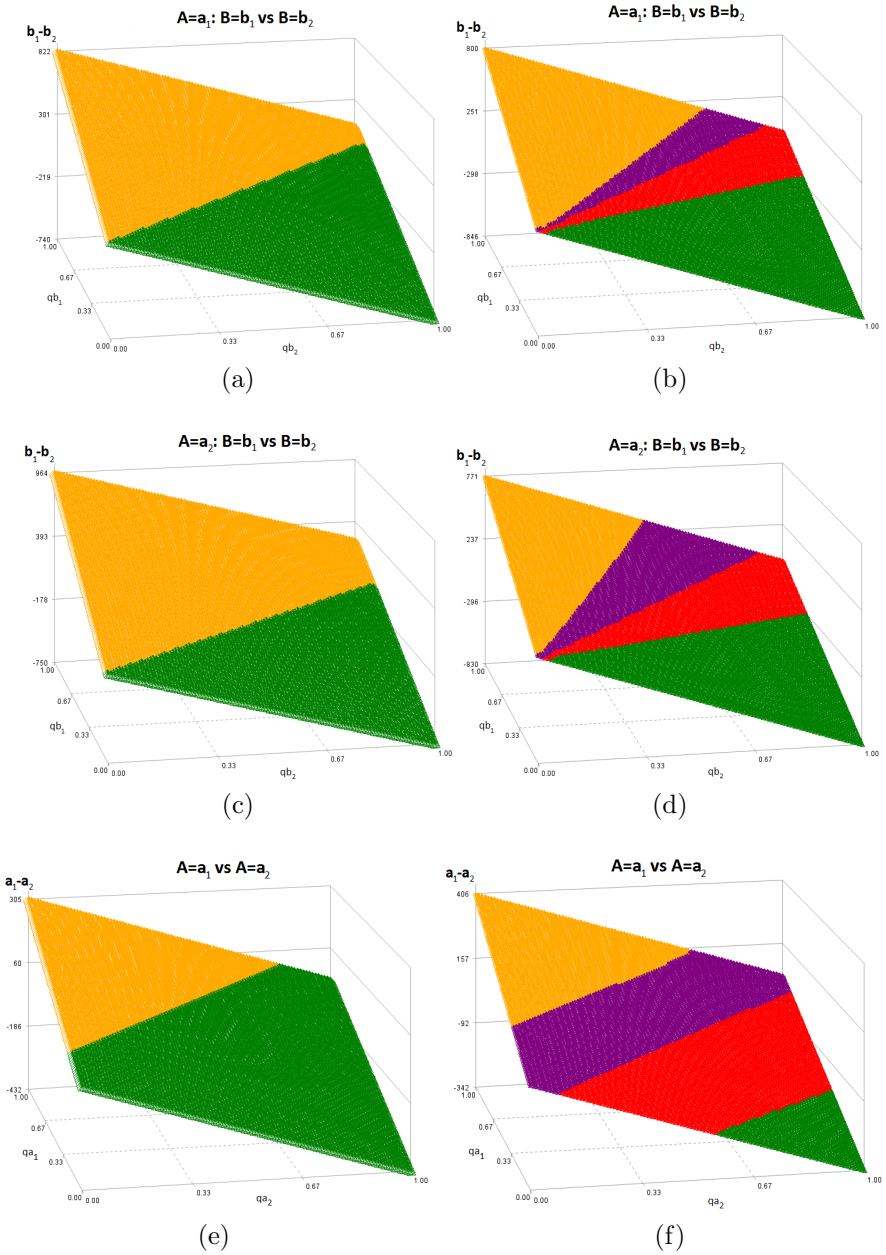

Figure 2.1: True (left column) and estimated (right column) threshold utility planes for the simulated scenario.

2.4 APPLICATION WITH THRESHOLD UTILITY ANALYSIS

In this section we apply the optimization methods discussed previously to the COG study A3891 concerning 379 children ages 6-months to 17 years old receiving treatment for high-risk neuroblastoma. All 379 patients were to receive five cycles of chemotherapy before beginning their induction treatment. Of these, 189 patients were randomized to receive continued chemotherapy (three additional cycles), $A = a_1$, and the remaining 190 were randomized to receive bone marrow transplantation, $A = a_2$. After completing the induction therapy, 203 patients were deemed responders (those for whom the disease did not progress) and consented to further randomization to receive six cycles of 13-cis-retinoic acid (160 mg per square meter per day for 14 consecutive days), $B = b_1$, or no further therapy, $B = b_2$. Survival time was truncated to 2452 days, since this was the largest observed death time in the study.

In what follows we assume the role of the prospective patient, considering only quality of life as affected by toxicity of treatment when choosing between treatments. Each of the therapies in this study comes with its own side effects. Following a cohort of lung cancer patients undergoing chemotherapy, Winter et al. (2013) [36] measured quality of life using the EORTC QLQ-C30 questionaire [1] as the patients completed multiple courses of chemotherapy. In the analysis by Winter et al. (2013) [36], the highest average global quality of life measure (ranging 0 to 100) over multiple courses of chemotherapy was 57. We rescaled these scores between 0 and 1 to have the quality of life weight of those undergoing chemotherapy vary between 0.5 to 0.6. In the case of bone marrow transplant, Felder et al. (2006) [9] analyze the health related quality of life of 68 pediatric patients aged 4 to 18 years old receiving allogeneic bone marrow or stem cell transplantation in a 5-year prospective study using The Pediatric Quality of Life Inventory(PedsQL) and The Health Utilities Index Mark2 + 3(HUI2/3). It is reasonable to interpret these scores as quality weights, indicating that those undergoing bone marrow transplantation have a quality of life near 0.7. Hong et al. (1986) [16] studied the use of 13-cis-retinoic acid in 44 patients with oral leukoplakia, and found

that cheilitis, erythema, and dry skin were most common. Based on the symptoms, mean survival time for patients on 13-cis-retinoic acid could reasonably be quality adjusted by 0.9.

Figure 2.2 (top row) shows the estimated stage 2 threshold utility planes - the estimated mean survival time for those on 13-cis-retinoic acid minus the estimated mean survival time for those on no further treatment. The yellow and green represent the region of strong acceptance for 13-cis-retinoic acid and no further therapy, respectively. The purple and red near the center of the plane give point estimates that favor 13-cis-retinoic acid and no further therapy, respectively, but the 90% point-wise bootstrap confidence intervals cover zero and represent the region of indifference when choosing between 13-cis-retinoic acid and no further therapy.

When the survival times for stage 2 treatments are both given a weight of 1 (no quality adjustment), those who received no further therapy had larger survival times than those who received 13-cis-retinoic acid, following continued chemotherapy; following bone marrow transplant, those who received 13-cis-retinoic acid had, on average, larger survival times than those who received no further therapy. It should be noted, though, that both of these point estimates fall within the m-out-of-n bootstrap indifference regions (the red and purple shaded areas), suggesting there is no statistically significant difference between the stage 2 treatments following either stage 1 treatment.

As one would begin to lower either q_{b_1} or q_{b_2} towards 0, while holding the other fixed, we see that the estimated difference in mean quality adjusted survival time falls in the region of statistical significance, where one stage 2 treatment truly out performs the other, given the stage 1 treatment. For q_{b_1}=0.9 and q_{b_2}=1, the stage 2 quality of life weights considered earlier for this study, the point estimate for the optimal stage 2 treatment falls in the same region as that for q_{b_1}=1 and q_{b_2}=1 described above and yields 13-cis-retinoic acid for those following bone marrow transplantation, and no further therapy for those following continued chemotherapy. If q_{b_1} is lower than 0.9, the optimal stage 2 treatment would be no further therapy for both induction therapies.

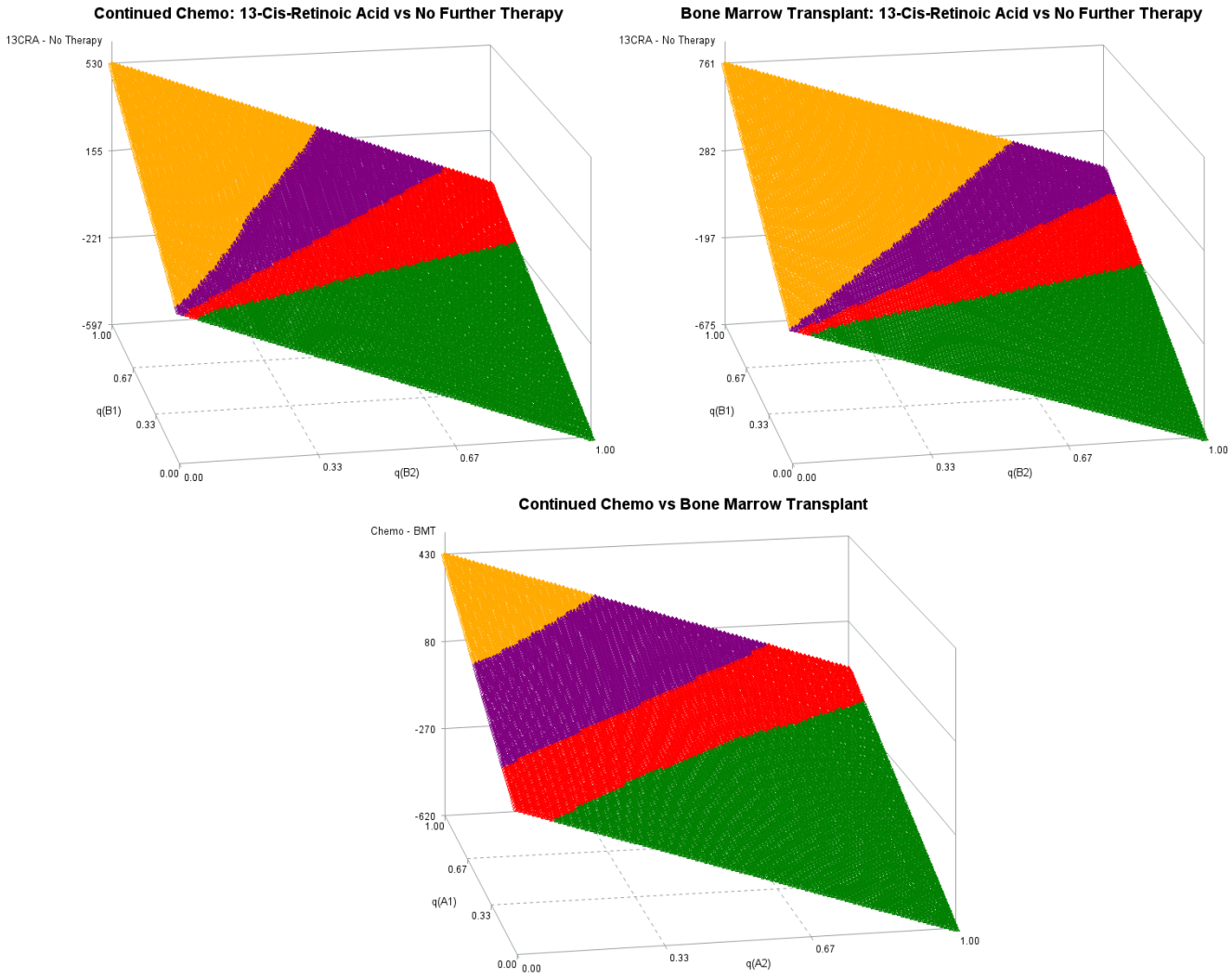

Figure 2.2: Estimated stage 2 (top row) and stage 1 (bottom row) threshold utility planes for COG study A3891.

Figure 2.2 (bottom row) also shows the estimated stage 1 threshold utility plane - the estimated mean survival time for those on continuation chemotherapy minus the estimated mean survival time for those who received a bone marrow transplant. This figure is generated using pseudo data where responders at stage 1 are assumed to take their optimal stage 2 treatment, and their remaining survival time is estimated using the methods from Section

2.2, with q_{b_1}=0.9 for 13-cis-retinoic acid and q_{b_2}=1 for no further treatment. At q_{a_1}=0.5 and q_{a_2}=0.7 the optimal stage 1 treatment is bone marrow transplant, and the point estimate falls within the strong acceptance region, meaning the 90% point-wise bootstrap confidence interval for the difference in mean survival time between continued chemotherapy and bone marrow transplant does not contain zero. Therefore, with q_{a_1}=0.5, q_{a_2}=0.7, q_{b_1}=0.9, and q_{b_2}=1, the optimal regime is to first treat with bone marrow transplantation and, if a response is observed, treat with 13-cis-retinoic acid.

2.5 GENERALIZATION TO OTHER OUTCOMES

Our exploration of Q-learning to optimize a dynamic treatment regime on quality adjusted lifetime leads one to consider $\mathcal{Q}$-functions that weight the expected utility at each stage for any continuous outcome, not just survival time. For a 2-stage SMART design depicted earlier with a primary outcome Y at the end of the second stage, one can use the $\mathcal{Q}$-functions

$$\mathcal{Q}_B\Big(A_i = a_j, \bar{X}_{Bi}, B_i = b_k\Big) = q_{b_k} E\Big[Y_i^{(B)}\Big| A_i = a_j, \bar{X}_{Bi}, B_i = b_k\Big], \tag{2.2}$$

$$\mathcal{Q}_A\Big(X_{Ai}, A_i = a_j\Big) = E\Big[H_i^{(A)}\Big| X_{Ai}, A_i = a_j\Big], \tag{2.3}$$

where

$$H_i^{(A)} = \begin{cases} Y_i^{(A)} q_{a_j} + \underset{b_k}{max}\ \mathcal{Q}_B\Big(A_i = a_j, \bar{X}_{Bi}, B_i = b_k,\Big), & \text{if } A_i = a_j,\ R_i = 1 \\ Y_i^{(A)} q_{a_j}, & \text{if } A_i = a_j,\ R_i = 0, \end{cases} \tag{2.4}$$

and where $Y_i^{(A)}$ and $Y_i^{(B)}$ are the outcomes at the first and second stages, respectively, with $Y_i^{(A)} + Y_i^{(B)} = Y_i$. The law of total expectation can be used to improve computational efficiency when performing a threshold utility analysis. Most authors fit a single regression model for $E\Big[H_i^{(A)}\Big| X_{Ai}, A_i = a_j\Big]$; however, a Q-learning model for stage 1 could be built

using

$$E\Big[H_i^{(A)}\Big|X_{Ai}, A_i = a_j\Big] = P(R_i = 1|X_{A_i}, A_i = a_j)$$
$$\times\Big\{q_{a_j}E\Big[Y_i^{(A)}\Big|X_{Ai}, A_i = a_j\Big] + E\Big[\underset{b_k}{max}\ \mathcal{Q}_B\Big(A_i = a_j, \bar{X}_{Bi}, B_i = b_k\Big)\Big|X_{Ai}, A_i = a_j\Big]\Big\}$$
$$+P(R_i = 0|X_{A_i}, A_i = a_j)\Big\{q_{a_j}E\Big[Y_i^{(A)}\Big|X_{Ai}, A_i = a_j,\Big]\Big\}. \tag{2.5}$$

Written this way, it is clear how the utility weights factor out of the expectations and create what we call quality adjusted Q-learning, for any continuous outcome. This could easily be generalized to SMARTs with an arbitrary number of stages. Modeling $E\Big[H_i^{(A)}\Big|X_{Ai}, A_i = a_j\Big]$ in this way improves computational efficiency since each of the component models only needs to be fit once before varying the utility weights and producing a threshold utility analysis. Other authors, including Song et al. (2011) [31], consider $\mathcal{Q}$-functions that have a single utility weight q, regardless of stage or treatment, that is multiplied to every nested expectation (except the first), creating an effect similar to the autoregressive working correlation structure from generalized linear models. Most authors interpret this single q as a utility weight that, when compounded over the nested expectations, diminishes the expected utility of each subsequent stage. The idea being that the prospective patient might not complete every stage of the DTR, and the optimal regime should give more importance to earlier treatments. However, even with this approach, most authors ignore the utility weight by setting it equal to 1. As we showed above, we propose assigning a separate utility weight to each treatment of each stage, representing the prospective patient's aversion to each treatment based on discomfort, side effects, monetary cost, ethical and/or religious beliefs, ability to complete the treatment schedule, and a host of other unmeasurable factors that might vary from one prospective patient to another. This allows a threshold utility analysis as described in Section 2.2.2 for any continuous outcome Y, and shows us the decision making process of the prospective patient.

2.6 CONCLUDING REMARKS

Quality adjusted lifetime is a natural endpoint for deciding among treatments that prolong survival time, permitting the prospective patient to factor toxicity and financial burden of treatment, among other factors, into the decision. This is particularly useful in the realm of DTRs, allowing the optimal regime to depend not only on patient level characteristics, but also on treatment characteristics. We have demonstrated how threshold utility analysis can be combined with the standard optimization algorithm to produce optimal regimes accounting for patient and treatment level information. For simplicity, our methods did not include any covariate information other than response status, but additional patient characteristics such as age, race, or sex could be included in the optimization algorithm, producing a separate set of utility planes for, say, males and females, or young children and older children. Patient information could also be used to improve efficiency by using the semiparametric estimating equations in Wang & Zhao (2007) [35].

3.0 CONDITIONAL STRUCTURAL MEAN MODELS AND VARIABLE SELECTION FOR OPTIMIZING DYNAMIC TREATMENT REGIMES

3.1 LINEAR REGRESSION AND VARIABLE SELECTION

3.1.1 Background

For observations $\boldsymbol{Y} = [Y_1, Y_2, ..., Y_n]^T$, a least squares linear regression model takes the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{\epsilon}$ is a n-dimensional mean zero random vector of errors, $\boldsymbol{\beta}$ is a $(p+1)$-dimensional vector of parameters, and $\boldsymbol{X}$ is an $n \times (p+1)$ design matrix of covariates, with a vector of 1's corresponding to an intercept. The parameter estimates are obtained by minimizing the sum of squared errors

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \tag{3.2}$$

where $\hat{Y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$. If the total number of covariates is large, one would naturally want to select the best subset of variables that explain the mean outcome. Not only is SSE used to find the line of best fit for a given model, it is also useful for goodness-of-fit when examining competing models and possible interaction effects. The model with smaller sum of squared errors fits the data best.

Other goodness-of-fit information criteria, such as AIC, BIC, and Mallow's C_p statistic, use SSE for model comparison while incorporating a penalty for increasing the number of model parameters. Allen (1974) [2] introduced the leave-one-out cross-validation estimate

of prediction error, CV, in the context of linear regression. Many other cross-validation criteria have been proposed since. The model with the smallest value of such a criteria should provide a good fit for another independent sample of data, that is, the model will have good out-of-sample prediction. It should also be noted that the p-values resulting from F or χ^2 tests of model parameters may also be used when choosing between competing models.

When using any of the above criteria for model selection, it may be infeasible to systematically search for the best subset of variables and interactions, simply because of the sheer number of variables available. In these cases there are several traditional and penalized variable selection methods that are computationally more efficient, though not guaranteed to produce the best subset. Such discrete methods include forward, backward, and stepwise variable selection. Other continuous variable selection methods include the least absolute shrinkage and selection operator (LASSO) and its derivatives. The model comparison criteria and variable selection methods above are not limited to least squares linear models. They work equally well for other regression models such as generalized linear models and regression splines.

3.1.2 Quantitative vs Qualitative Interactions and Variable Selection

These variable selection methods help to find the covariates and their interactions that are predictive when estimating the expected value of Y_i, but do not clearly identify for which values of the covariates the choice of optimal treatment changes, where we take the optimal treatment to be the one with the largest expected outcome. Gunter et al. (2011) [13] define *predictive* variables as those used to reduce the variability and increase the accuracy of the estimator, whereas variables that help prescribe the optimal treatment for a given patient are called *prescriptive* [15]. When estimating the mean outcome, it is best to collect as many predictive variables as possible; however, only those predictive variables that are also prescriptive are needed when deciding between treatments (Figure 3.1). In order for a variable to be prescriptive, it must qualitatively interact with treatment. A variable X is said to qualitatively interact with the treatment Z if there exists at least two distinct

non-empty sets within the space of X for which the optimal treatment is different. That is, there exists disjoint, non-empty sets $S_1, S_2 \subset space(X)$ for which

$$\underset{Z}{argmax} E[Y|X = x_1, Z = z] \neq \underset{Z}{argmax} E[Y|X = x_2, Z = z],$$

for all $x_1 \in S_1$ and $x_2 \in S_2$.

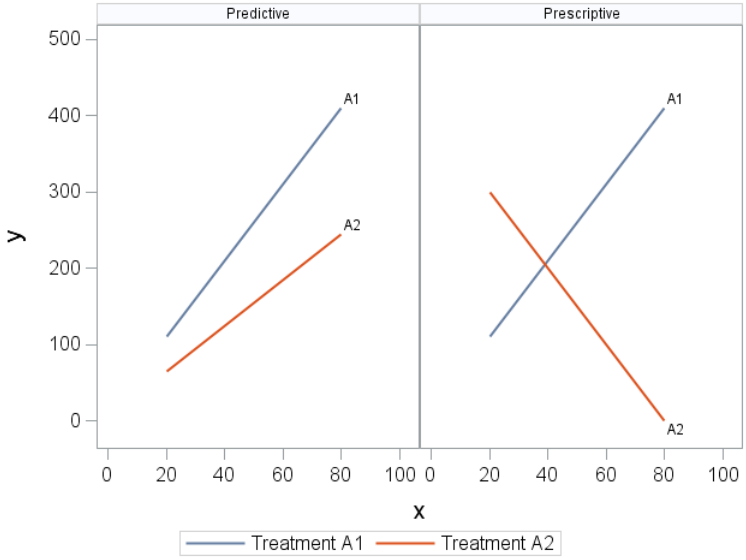

Figure 3.1: Predictive vs Prescriptive Interactions

Working in the single-stage setting using backwards induction, Gunter et al. (2011) [13] propose two different ranking methods to sort variables according to how likely they will qualitatively interact with treatment, and provide a four step algorithm involving LASSO regression on nested subsets of covariates for selecting important predictive variables. Following their work, Zhang (2014) [37] generalizes from the least squares regression model and considers models of the form

$$E[Y_i|X_i, Z_i] = h(X_i, \boldsymbol{\beta}) + Z_i \times u(X_i, \boldsymbol{\alpha}), \tag{3.3}$$

where $Z_i \in \{1, -1\}$ is a treatment assignment indicator, and $u(X_i, \boldsymbol{\alpha})$ are interaction effects with parameters $\boldsymbol{\alpha}$. They offer a simpler, more effective two step method: (i) Multivariate

adaptive regression spline (MARS) models are used to fit a nonparametric model on the outcome of interest and simultaneously select predictive (and prescriptive) variables from a larger subset of variables, and (ii) the sign of the interaction contrasts is used to create a binary variable indicating the optimal regime for each subject, and penalized logistic regression with LASSO (L_1-logistic regression) is used to identify which of the significant interactions are not only predictive, but also prescriptive. The use of a MARS model in the first step is warranted if we do not care about interpretability of model parameters, but are primarily interested in the predicted outcome given the covariates.

As will be seen in Section 3.4 in the context of survival analysis in a two stage sequential multiple assignment randomized trial (SMART) design, a similar two step method can be used to identify important qualitative interactions for structural mean models. We consider a specific two stage setting, similar to that used in our application, but the methods described easily extend to other DTR setups.

3.2 DYNAMIC TREATMENT REGIMES AND CORRESPONDING TERMINOLOGY

Consider a two-stage sequential multiple assignment randomized trial (SMART) design where patients are randomized to one of four induction therapies, $\mathcal{A} = \{a_1, a_2, a_3, a_4\}$. A patient could die, the disease could become resistant to the initial treatment, the patient could respond (complete remission), or he/she could experience disease progression after complete remission. For each of the induction therapies, if treatment resistance or progression following complete remission is observed, patients are further randomized to one of two salvage treatments, $\mathcal{B} = \{b_1, b_2\}$. This design allows for inference on sixteen DTRs that might be carried out in clinical practice, namely, $d(A_i = a_j; B_{1i} = b_k, B_{2i} = b_l),\ j = 1, ..., 4,\ k = 1, 2,\ l = 1, 2$ where $d(A_i; B_{1i}, B_{2i})$ stands for "Treat with A_i; if the patient is resistant to A_i, treat with B_{1i}, if the patient responds to A_i (complete remission) but later experiences disease progres-

sion, treat with B_{2i}." Our goal is to find the optimal treatment regime among these that maximizes expected survival time.

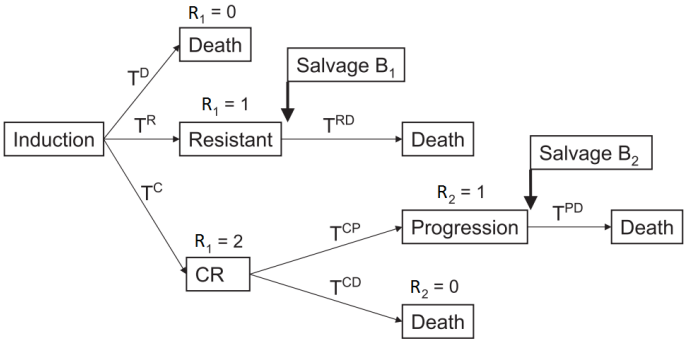

Figure 3.2: Possible pathways, transition times, and salvage therapy following induction treatment.

Let T_i^D, T_i^R, T_i^{RD}, T_i^C, T_i^{CP}, and T_i^{PD}, T_i^{CD}, respectively denote the observed time to death if neither remission nor resistance was observed, the observed time to resistance and the observed time from resistance to death if resistance is observed, the observed time to complete remission, the observed time from complete remission to disease progression, the observed time from progression to death, and the observed time from complete remission to death if complete remission is observed. Using the above sojourn times, each patient's survival time can be expressed as

$$T_i = \begin{cases} T_i^D, & R_{1i} = 0 \\ T_i^R + T_i^{RD}, & R_{1i} = 1 \\ T_i^C + T_i^{CP} + T_i^{PD}, & R_{1i} = 2,\ R_{2i} = 1 \\ T_i^C + T_i^{CD} & R_{1i} = 2,\ R_{2i} = 0, \end{cases}$$

where R_{1i} indicates whether a patient fails, is resistant, or experiences complete remission, and R_{2i} indicates whether or not those that experienced complete remission later experience disease progression. R_{1i} and R_{2i} index the paths of each treatment regime.

In the presence of non-informative right censoring, one might consider the restricted survival time where total follow-up time is limited to L, where L is some value less than the maximum survival time for all patients. Therefore, the survival time for all patients will be truncated at L, T^L =min(T, L). For ease of notation, we will drop the superscript and simply use T. We will denote the i^{th} patient's censoring time by C_i, and the survival distribution of C_i by $K(t) = P(C_i > t)$. Define U_i =min(T_i, C_i) and $\Delta_i = I(T_i \leq C_i)$, respectively, to be the observed time to event (death or censoring), and the death indicator. It is possible that $C_i < T_i$, so that for a single patient some of the sojourn times are censored while others are observed. Therefore, U_i can be expressed as

$$U_i = \begin{cases} U_i^D, & R_{1i} = 0 \\ T_i^R + U_i^{RD}, & R_{1i} = 1 \\ T_i^C + T_i^{CP} + U_i^{PD}, & R_{1i} = 2,\ R_{2i} = 1 \\ T_i^C + U_i^{CD} & R_{1i} = 2,\ R_{2i} = 0, \end{cases}$$

where R_{1i}=0 if a patient fails or is censored prior to observing R_{1i}; R_{2i}=0 if a patient dies after complete remission or is censored after complete remission prior to observing R_{2i}; U_i^D = min(T_i^D, C_i) and $\Delta_i^D = I(T_i^D \leq C_i)$; U_i^{RD} = min$(T_i^{RD}, C_i - T_i^R)$ and $\Delta_i^{RD} = I(T_i^{RD} \leq C_i - T_i^R)$; U_i^{PD} = min$(T_i^{PD}, C_i - T_i^{CP} - T_i^C)$ and $\Delta_i^{PD} = I(T_i^{PD} \leq C_i - T_i^{CP} - T_i^C)$; U_i^{CD} = min$(T_i^{CD}, C_i - T_i^C)$ and $\Delta_i^{CD} = I(T_i^{CD} \leq C_i - T_i^C)$. Then, introducing further indicators for first and second stage treatment, the observed data for the i^{th} patient in the presence of censoring is written as

$$\begin{aligned} D_i^{\delta} = \Big(& Z_{ji}^{(A)},\ R_{1i},\ I\{R_{1i} = 2\}R_{2i},\ I\{R_{1i} = 0\}U_i^D,\ I\{R_{1i} = 1\}T_i^R, \\ & I\{R_{1i} = 1\}Z_{ki}^{(B_1)},\ I\{R_{1i} = 1\}U_i^{RD},\ I\{R_{1i} = 2\}T_i^C, \\ & I\{R_{1i} = 2\}I\{R_{2i} = 1\}T_i^{CP}, I\{R_{1i} = 2\}I\{R_{2i} = 1\}Z_{li}^{(B_2)}, \\ & I\{R_{1i} = 2\}I\{R_{2i} = 1\}U_i^{PD}, I\{R_{1i} = 2\}I\{R_{2i} = 0\}U_i^{CD}, \\ & U_i,\ \Delta_i,\ \Delta_i^D,\ \Delta_i^{RD},\ \Delta_i^{PD},\ \Delta_i^{CD},\ G_i^H(U_i)\Big), \\ & j = 1, 2, 3, 4,\ k, l = 1, 2, \end{aligned}$$

where $Z_{ji}^{(A)}$=$I\{A_i = a_j\}$ equals 1 if patient i received the j^{th} induction therapy, $Z_{ji}^{(A)}$ equals

0 otherwise, $Z_{ki}^{(B_1)}=I\{B_{1i} = b_k\}$ and $Z_{li}^{(B_2)}=I\{B_{2i} = b_l\}$ denote the salvage treatment assignment indicators, defined only if R_{1i}=1 or 2, respectively, and $G_i^H(t)$ denotes information collected on patient i prior to time t. Using the observed data, one can create treatment regime indicators as $d_i(a_j; b_k, b_l) = Z_{ji}^{(A)}\Big(I\{R_{1i} = 0\} + I\{R_{1i} = 1\}Z_{ki}^{(B_1)} + I\{R_{1i} = 2\}I\{R_{2i} = 1\}Z_{li}^{(B_2)} + I\{R_{1i} = 2\}I\{R_{2i} = 0\}\Big)$.

By design, treatments are assigned independently of prognosis or any observed data measured prior to the second stage. Therefore,

$$P\Big(Z_{ji}^{(A)} = 1\Big) = \pi_j^{(A)}, \tag{3.4}$$

$$P\Big(Z_{ki}^{(B_1)} = 1\Big) = \pi_k^{(B_1)}, \tag{3.5}$$

$$P\Big(Z_{li}^{(B_2)} = 1\Big) = \pi_l^{(B_2)}, \tag{3.6}$$

where $\pi_j^{(A)}$, $\pi_k^{(B_1)}$, and $\pi_l^{(B_2)}$ are known randomization probabilities. These three conditions are often referred to as no unmeasured confounders or sequential randomization assumption. This 'no unmeasured confounders' condition holds even if the second-stage randomization probabilities depend on the first-stage treatment assignments.

3.3 STRUCTURAL MEAN MODELS FOR DYNAMIC TREATMENT REGIMES

3.3.1 Structural Mean Models Conditional on Baseline Information

To estimate the mean of each dynamic treatment regime, one can use structural mean models and then compare the means to determine the optimal regime. Inverse-probability-of-treatment weighting (IPTW) and g-computation are two methods. Wahed and Tsiatis (2004) [34] provide a nice discussion of the first method in the context of survival analysis with no adjustment for covariates.

The focus will be to estimate $\mu_{jkl} = E[T_i|d_i(a_j; b_k, b_l) = 1]$, $j = 1, 2, 3, 4$, $k, l = 1, 2$, the mean survival time for those following a given regime. Since our SMART design allows us to confidently assume no unmeasured confounders, each regime mean is representative of the expected outcome had the entire sample of patients followed that regime. Recall that patients following $d(a_j; b_k, b_l)$ are a mixture of four groups. We can use data from these patients to infer about μ_{jkl}, accounting for the two stages of randomization. If there was no randomization, and if everyone in the sample was treated using the same DTR, we would have used the sample average $\frac{1}{n}\sum_{i=1}^{n} T_i$ to estimate μ. If there was only one stage of randomization, we would have considered using $\frac{\sum_{i=1}^{n} Z_{ji}^{(A)} T_i}{\sum_{i=1}^{n} Z_{ji}^{(A)}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_{ji}^{(A)}}{\hat{\pi}_j^{(A)}} T_i \approx \frac{1}{n}\sum_{i=1}^{n} \frac{Z_{ji}^{(A)}}{\pi_j^{(A)}} T_i$. To account for the two stages of randomization we consider the quantity $W_{jkli} = \frac{Z_{ji}^{(A)}}{\pi_j^{(A)}} \Big(I\{R_{1i} = 0\} + I\{R_{1i} = 1\} \frac{Z_{ki}^{(B_1)}}{\pi_k^{(B_1)}} + I\{R_{1i} = 2\} I\{R_{2i} = 1\} \frac{Z_{li}^{(B_2)}}{\pi_l^{(B_2)}} + I\{R_{1i} = 2\} I\{R_{2i} = 0\} \Big)$. Note that $W_{jkli} T_i$ is non-zero only for patients who are treated according to $d(a_j; b_k, b_l)$, and based on the assumptions in Section 3.2 $W_{jkli} T_i$ has expectation equal to μ_{jkl}, which implies that to find an unbiased estimator of μ_{jkl} one need only turn to the empirical average, $\frac{1}{n}\sum_{i=1}^{n} W_{jkli} T_i$. When censoring is present, the above result should be modified slightly. Using the observed data in (3.4), the estimator for μ_{jkl} becomes

$$\hat{\mu}_{jkl}^{IPTWcen} = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} W_{jkli} U_i, \tag{3.7}$$

where $\hat{K}(t)$ is the Kaplan-Meier estimator or any other consistent estimator of the censoring survival distribution.

In a basic randomized clinical trial, the mean outcome for each treatment group is estimated and compared to see which treatment has the largest expected outcome, assuming larger outcomes are better. Similarly, the marginal estimators above are useful for comparing the mean outcomes across treatment regimes to identify which treatment regime has the largest expected outcome. As in a basic randomized clinical trial, a subgroup analysis can be performed to see if the marginal results hold throughout, or if the optimal treatment regime depends on patient characteristics. Following Robins et al. (2008) [28], Orellana & Rotnitzky (2010) [23], and Wang & Zhao (2007) [35], the estimator for mean survival time

in the presence of censoring can be extended to the regression setting to adjust for baseline covariates using an accelerated failure time (AFT) model via the estimating equation

$$\mathcal{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{4}\sum_{k=1}^{2}\sum_{l=1}^{2}\frac{\Delta_i}{\hat{K}(U_i)}W_{jkli}\left\{\frac{\partial m}{\partial \boldsymbol{\theta}}\right\}^T\left[\log U_i - m\Big(X_i, \boldsymbol{d}_i, \boldsymbol{\theta}\Big)\right] = 0, \tag{3.8}$$

where $\{\}^T$ is the transpose operator, X_i is a vector of baseline covariates from $G_i^H(0)$, $\boldsymbol{d}_i$=$\big[d_i(a_1;b_1,b_1)$, ...,$d_i(a_4;b_2,b_2)\big]^T$, $m\Big(X_i, \boldsymbol{d}_i, \boldsymbol{\theta}\Big) = E\Big[\log T_i | X_i, \boldsymbol{d}_i\Big]$ the mean function, and $\mu\Big(X_i, \boldsymbol{d}_i, \boldsymbol{\theta}\Big) \equiv \exp\Big\{m\Big(X_i, \boldsymbol{d}_i, \boldsymbol{\theta}\Big)\Big\} \approx E\Big[T_i | X_i, \boldsymbol{d}_i\Big]$. For example $m\Big(X_i, \boldsymbol{d}_i, \boldsymbol{\theta}\Big)$ could be modeled as

$$\begin{aligned} m\Big(X_i, \boldsymbol{d}_i, \boldsymbol{\theta}\Big) = X_i^T\boldsymbol{\beta} \;&+\; d_i(a_1;b_1,b_1)X_i^T\boldsymbol{\alpha}_{111} \\ &+\; d_i(a_1;b_1,b_2)X_i^T\boldsymbol{\alpha}_{112} \\ &+\; d_i(a_1;b_2,b_1)X_i^T\boldsymbol{\alpha}_{121} \\ &\;\vdots \\ &+\; d_i(a_4;b_2,b_2)X_i^T\boldsymbol{\alpha}_{422}, \end{aligned} \tag{3.9}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}^T, \boldsymbol{\alpha}_{111}^T, \boldsymbol{\alpha}_{112}^T, \cdots, \boldsymbol{\alpha}_{422}^T\}^T$, and X_i contains an element equal to 1 corresponding to an intercept term. The preliminary optimal treatment regime, the one with the largest expected outcome, is given by

$$d^{opt}(X_i) = \{d(a_{j^*};b_{k^*},b_{l^*}),\ a_{j^*},b_{k^*},b_{l^*} = \underset{a_j,b_k,b_l}{argmax}\,\mu(X_i,\boldsymbol{d}_i,\boldsymbol{\theta})\}. \tag{3.10}$$

We use the term 'preliminary' when referring to an optimal regime that is conditional on baseline information, but marginalized over stage 2 information. The optimal frontline treatment is given by $A^{opt}(X_i)$=$\underset{a_j}{argmax}\,\{\underset{b_k,b_l}{max}\,\mu(X_i,\boldsymbol{d}_i,\boldsymbol{\theta})\}$.

To implement this estimating equation, one would create sixteen copies of the analysis data set, each with a distinct value of $\frac{\Delta_i}{\hat{K}(U_i)}W_{jkli}$. The indicators $d_i(a_{j'};b_{k'},b_{l'})$, where $j' \neq j$ or $k' \neq k$ or $l' \neq l$, would be artificially set to zero so that the observations with non-zero weights in a given copy of the data set belong to only one regime. This effectively replicates the observations that are consistent with more than one regime (Chakraborty and Murphy,

2014) [5]. These sixteen data sets would then be stacked one on top another and submitted to a software package for a weighted regression. Treating $\hat{K}(U_i)$ as known, the empirical sandwich estimator of the covariance matrix for the parameter estimates can be used to draw inference when comparing regime means. When treatment assignment is not random, as will be the case in the Sections 3.5 and 3.6, the treatment assignment probabilities can be modeled using logistic regression. This is important in order to maintain the no unmeasured confounders assumption.

It should be noted that the above regression model only incorporated baseline information, yet patient information is available throughout the trial. Using the law of total expectation, the mean survival time under a regime of interest that is conditional on all possible patient information is given by

$$
\begin{aligned}
&E[T_i|X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P, d_i(A; B_1, B_2) = 1] = \\
&\quad P(R_{1i} = 0|A_i, X_i)E[T_i^D|A_i, X_i, R_{1i} = 0] \\
&\quad +P(R_{1i} = 1|A_i, X_i)\Big\{E[T_i^R|A_i, X_i, R_{1i} = 1] + E[T_i^{RD}|A_i, B_{1i}, \bar{X}_i^R, R_{1i} = 1]\Big\} \\
&\quad +P(R_{1i} = 2|A_i, X_i)P(R_{2i} = 1|R_{1i} = 2, A_i, \bar{X}_i^C)\Big\{E[T_i^C|A_i, X_i, R_{1i} = 2] \\
&\qquad +E[T_i^{CP}|A_i, \bar{X}_i^C, R_{1i} = 2, R_{2i} = 1] + E[T_i^{PD}|A_i, B_{2i}, \bar{X}_i^P, R_{1i} = 2, R_{2i} = 1]\Big\} \\
&\quad +P(R_{1i} = 2|A_i, X_i)P(R_{2i} = 0|A_i, \bar{X}_i^C)\Big\{E[T_i^C|A_i, X_i, R_{1i} = 2] \\
&\qquad +E[T_i^{CD}|A_i, \bar{X}_i^C, R_{1i} = 2, R_{2i} = 0]\Big\}, \qquad (3.11)
\end{aligned}
$$

where $\bar{X}_i^R$, $\bar{X}_i^C$, and $\bar{X}_i^P$ are vectors of covariates from $G_i^H(T_i^R)$, $G_i^H(T_i^C)$, and $G_i^H(T_i^C + T_i^{CP})$, respectively. Because we have no unmeasured confounders, it is as though we can peek into alternate universes and see all of the potential outcomes a prospective patient would have for each of the different response groups, given his/her information at each stage. We gather all of that patient information together at once and combine it into a composite score in Equation (3.11). In practice Equation (3.11) is not very useful unless we are willing to consider specific patient information for every intermediate outcome. Nevertheless, one can

set estimating equations, for example

$$\sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} I\{R_{1i}=1\}\left\{\frac{\partial m^{RD}}{\partial \boldsymbol{\theta}^{RD}}\right\}^T \left\{\log U_i^{RD} - m^{RD}\big(\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}\big)\right\} = 0 \qquad (3.12)$$

with mean model $m^{RD}\big(\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}\big)$ and $\mu^{RD}\big(\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}\big)$
$\equiv \exp\Big\{m^{RD}\big(\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}\big)\Big\} \approx E[T_i^{RD}|\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}]$ for those with $R_{1i}=1$, to model the sojourn times of each path. Then a g-computation estimator for
$E[T_i|X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P, d_i(A; B_1, B_2)=1]$ that is conditional on baseline and follow up information can be created using

$$\begin{aligned}
&\mu\Big(X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P, d_i(A; B_1, B_2)=1, \boldsymbol{\theta}, \boldsymbol{\psi}\Big) = \\
&\quad P(R_{1i}=0|A_i, X_i, \boldsymbol{\psi}_1)\Big\{\mu^D(A_i, X_i, \boldsymbol{\theta}^D)\Big\} \\
&\quad +P(R_{1i}=1|A_i, X_i, \boldsymbol{\psi}_1)\Big\{\mu^R(A_i, X_i, \boldsymbol{\theta}^R) + \mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big\} \\
&\quad +P(R_{1i}=2|A_i, X_i, \boldsymbol{\psi}_1)P(R_{2i}=1|A_i, \bar{X}_i^C, \boldsymbol{\psi}_2)\Big\{\mu^C(A_i, X_i, \boldsymbol{\theta}^C) \\
&\qquad\qquad +\mu^{CP}(A_i, \bar{X}_i^C, \boldsymbol{\theta}^{CP}) + \mu^{PD}(A_i, B_{2i}, \bar{X}_i^P, \boldsymbol{\theta}^{PD})\Big\} \\
&\quad +P(R_{1i}=2|A_i, X_i, \boldsymbol{\psi}_1)P(R_{2i}=0|A_i, \bar{X}_i^C, \boldsymbol{\psi}_2)\Big\{\mu^C(A_i, X_i, \boldsymbol{\theta}^C) \\
&\qquad\qquad +\mu^{CD}(A_i, \bar{X}_i^C, \boldsymbol{\theta}^{CD})\Big\}, \qquad (3.13)
\end{aligned}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}^D, \boldsymbol{\theta}^R, \boldsymbol{\theta}^{RD}, \boldsymbol{\theta}^C, \boldsymbol{\theta}^{CP}, \boldsymbol{\theta}^{PD}, \boldsymbol{\theta}^{CD}]$ and $\boldsymbol{\psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2]$. $P(R_{1i}=r|A_i, X_i, \boldsymbol{\psi}_1)$ and $P(R_{2i}=s|A_i, X_i, X_i^C, \boldsymbol{\psi}_2)$ can be modeled through logistic regression. The optimal regime, the one with the largest expected outcome, is given by

$$\begin{aligned}
d^{opt}(X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P) \;=\; &\Big\{d(a_{j^*}; b_{k^*}, b_{l^*}),\ a_{j^*}, b_{k^*}, b_{l^*} = \\
&\underset{a_j, b_k, b_l}{argmax}\ \mu\Big(X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P, d_i(a_j; b_k, b_l)=1, \boldsymbol{\theta}, \boldsymbol{\psi}\Big)\Big\}, \qquad (3.14)
\end{aligned}$$

and the optimal frontline treatment is given by
$A^{opt}(X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P) = \underset{a_j}{argmax}\ \Big\{\underset{b_k, b_l}{max}\ \mu\Big(X_i, \bar{X}_i^R, \bar{X}_i^C, \bar{X}_i^P, d_i(a_j; b_k, b_l)=1, \boldsymbol{\theta}, \boldsymbol{\psi}\Big)\Big\}$.

Equation (3.13) represents a weighted average of patient outcomes between different response groups. The prospective patient has not yet experienced his/her sample path, and in a sense has missing data for all stages after baseline. To estimate the mean outcome for the prospective patient under a regime of interest requires us to integrate (3.13) over the probability measure of the covariates that are missing for this patient, the covariates collected after baseline. This leaves us with an estimated mean under the regime of interest, given the baseline information we have on the prospective patient. Most authors comparing dynamic treatment regimes using structural models, such as IPTW and g-computation estimators, routinely integrate over all patient information, except treatment assignment, performing a marginal comparison of regimes. In our approach we integrate (3.13) over all stage 2 information, except for stage 2 treatment assignment, facilitating a comparison of treatment regimes conditional on baseline information, similar to (3.8) and (3.9). Integrating (3.13) produces a preliminary optimal regime $d^{opt}(X_i)$ in (3.14), and $A^{opt}(X_i) = \underset{a_j}{argmax} \left\{ \underset{b_k, b_l}{max}\, \mu\Big(X_i, d_i(a_j; b_k, b_l) = 1, \boldsymbol{\theta}, \boldsymbol{\psi}\Big) \right\}$.

If one is willing to assume that, conditional on patient information, the response proportions are independent of the corresponding mean sojourn times (with respect to the covariates), then the integration can be performed piece-wise for each component. To operationalize this, one would first fit the component model for response or sojourn time with all of the significant terms through stage 2. The predicted values of this model would then be regressed on the same covariates as before, except for any stage 2 covariates. This effectively averages the model over the stage two covariates, leaving a model that is conditional on baseline information only. These integrated component models can then be combined to form (3.13). When treatment assignment is not random, as will be the case in the Sections 3.5 and 3.6, all variables that are confounded with treatment assignment should be included in the sojourn time models. This is important in order to maintain the no unmeasured confounders assumption.

3.3.2 Tailoring the Salvage Therapy

Regardless of whether g-computation (3.13) or IPTW (3.8) and (3.9) is used, to tailor the stage 2 treatment prescribed by the preliminary optimal regime, the mean models for the stage 2 sojourn times, i.e. $E[T_i^{RD}|A_i, B_{1i}, \bar{X}_i^R, R_{1i} = 1]$ and $E[T_i^{PD}|A_i, B_{2i}, \bar{X}_i^P, R_{1i} = 2, R_{2i} = 1]$, can be examined using the estimating equations

$$\sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} I\{R_{1i} = 1\} \Big\{ \frac{\partial m^{RD}}{\partial \boldsymbol{\theta}^{RD}} \Big\}^T \Big\{ \log U_i^{RD} - m^{RD}\big(\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}\big) \Big\} = 0 \tag{3.15}$$

and

$$\sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} I\{R_{1i} = 2\} I\{R_{2i} = 1\} \Big\{ \frac{\partial m^{PD}}{\partial \boldsymbol{\theta}^{PD}} \Big\}^T \Big\{ \log U_i^{PD} - m^{PD}\big(\bar{X}_i^P, A_i, B_{2i}, \boldsymbol{\theta}^{PD}\big) \Big\} = 0. \tag{3.16}$$

By evaluating $\mu^{RD}\big(\bar{X}_i^R, A_i, B_{1i}, \boldsymbol{\theta}^{RD}\big)$ and $\mu^{PD}\big(\bar{X}_i^P, A_i, B_{2i}, \boldsymbol{\theta}^{PD}\big)$ at $A_i = A^{opt}(X_i)$, the optimal stage 2 treatment given optimal stage 1 treatment can be identified using

$$B_1^{opt}(\bar{X}_i^R) = \underset{b_k}{argmax}\ \mu^{RD}\big(\bar{X}_i^R, A_i = A^{opt}(X_i), B_{1i} = b_k, \boldsymbol{\theta}^{RD}\big) \tag{3.17}$$

and

$$B_2^{opt}(\bar{X}_i^P) = \underset{b_l}{argmax}\ \mu^{PD}\big(\bar{X}_i^P, A_i = A^{opt}(X_i), B_{2i} = b_l, \boldsymbol{\theta}^{PD}\big) \tag{3.18}$$

for $R_{1i} = 1$, and $R_{1i} = 2$ and $R_{2i} = 1$, respectively. The optimal treatment regime using conditional structural mean models can then be constructed as "Treat with $A^{opt}(X_i)$; if resistance is observed, treat with $B_1^{opt}(\bar{X}_i^R)$; if disease progression after complete remission is observed, treat with $B_2^{opt}(\bar{X}_i^P)$." The beauty of constructing optimal dynamic treatment regimes in this way is that if additional stage 2 patient information is not available, a salvage treatment based on baseline information can still be prescribed using $d^{opt}(X_i)$. Although we have demonstrated this technique for optimizing a dynamic treatment regime on a specific two stage SMART design, the methods are easily generalized to other SMART designs with an arbitrary number of stages. The IPTW or g-computation estimator is used at each stage to estimate the preliminary optimal treatment regime given patient information up to the current stage and prior treatment assignment. Essentially this tailors the optimal treatment assignment at the current stage, and provides an optimal strategy for the remaining stages

given the information currently available. The IPTW and g-computation estimators reduce to a simple regression model for the final stage. All authors we have encountered who use conditional structural models (IPTW) do so using only baseline information, prescribing the optimal treatment regime using $d^{opt}(X_i)$, but naturally it is best to re-evaluate the strategy as more information becomes available. This is what we propose.

3.3.3 Comparison with Q-learning

From the reinforcement learning literature in the field of DTRs, the Bellman equations [3] identify the optimal treatment at each stage and lead to the Q-functions that comprise Q-learning. For our two stage SMART design, assuming no unmeasured confounders, these would be

$$\begin{aligned}
\mathcal{Q}_{B_1}\Big(A_i, \bar{X}_i^R, B_{1i} = b_k\Big) &= E[T_i^{RD} | A_i, B_{1i} = b_k, \bar{X}_i^R, R_{1i} = 1], \\
\mathcal{Q}_{B_2}\Big(A_i, \bar{X}_i^P, B_{2i} = b_l\Big) &= E[T_i^{PD} | A_i, B_{2i} = b_l, \bar{X}_i^P, R_{1i} = 2, R_{2i} = 1], \\
\mathcal{Q}_A\Big(X_i, A_i = a_j\Big) &= E\Big[H_i^{(A)} | X_i, A_i = a_j\Big],
\end{aligned}$$

where

$$H_i^{(A)} = \begin{cases}
T_i^D, & \text{if } R_{1i} = 0 \\
T_i^R + \underset{b_k}{max}\ \mathcal{Q}_{B_1}\Big(A_i, \bar{X}_i^R, B_{1i} = b_k\Big), & \text{if } R_{1i} = 1 \\
T_i^C + T_i^{CP} + \underset{b_l}{max}\ \mathcal{Q}_{B_2}\Big(A_i, \bar{X}_i^P, B_{2i} = b_l\Big), & \text{if } R_{1i} = 2,\ R_{2i} = 1 \\
T_i^C + T_i^{CD}, & \text{if } R_{1i} = 2,\ R_{2i} = 0,
\end{cases}$$

with $A^{opt}(X_i) \equiv \underset{a_j}{argmax}\ \mathcal{Q}_A\Big(X_i, A_i = a_j\Big)$, $B_1^{opt}(\bar{X}_i^R) \equiv \underset{b_k}{argmax}\ \mathcal{Q}_{B_1}\Big(A_i = A^{opt}(X_i), \bar{X}_i^R, B_{1i} = b_k\Big)$, and $B_2^{opt}(\bar{X}_i^P) \equiv \underset{b_l}{argmax}\ \mathcal{Q}_{B_2}\Big(A_i = A^{opt}(X_i), \bar{X}_i^P, B_{2i} = b_l\Big)$. The similarity between Q-learning and g-computation is striking, except that Q-learning averages over stage 2 information and stage 2 treatment assignment, whereas g-computation averages over stage 2 information while holding stage 2 treatment assignment fixed when searching for $A^{opt}(X_i)$. To see this, compare the expected value of $H_i^{(A)}$ using the law of total expectation with Equation (3.11). Regardless of what estimation method is used (structural or nested mod-

els), the optimal choice of frontline treatment depends on what salvage treatment is taken. To identify $A^{opt}(X_i)$, Q-learning (through the use of pseudo data $H_i^{(A)}$) assumes that those who move to stage 2 take their optimal salvage therapy. The optimization of A_i is marginalized over all $B_1^{opt}(\bar{X}_i^R)$ and $B_2^{opt}(\bar{X}_i^P)$. Q-learning estimates all of the best strategies over later stages, combines them into an average best strategy, and assigns the optimal frontline treatment based on this average best strategy given baseline information. Conditional structural mean models consider an average patient over later stages, they identify the single best strategy for the average patient, and assign the optimal frontline treatment based on this best strategy given baseline information. Interestingly, all of the same steps are applied in Q-learning and g-computation, the only difference being their order. G-computation first finds the sojourn means conditional on response status, combines them using the law of total expectation, integrates over stage 2 information, and then applies the max and argmax operators to identify $A^{opt}(X_i)$. On the other hand, Q-learning first finds the sojourn means conditional on response status, applies the max operators to the stage 2 sojourn means, integrates over stage 2 information, combines them using the law of total expectation, and then applies the argmax operator to identify $A^{opt}(X_i)$.

Colloquially, both methods are said to estimate the optimal dynamic treatment regime. The question then becomes, "Under what conditions, if any, are the two methods equivalent, and is one method preferable over the other?" Notwithstanding the differences just described, Q-learning and g-computation are also different estimators because g-computation, as constructed in Equation (3.11), includes additional patient information, $\bar{X}^C$, between complete remission and disease progression for the outcome models and response proportions. To make a fair comparison, we momentarily consider the g-computation model where the R_{2i} response proportion models and intermediate outcome models for T_i^{CP} and T_i^{CD} depend only on baseline information. Since Q-learning and g-computation (as we propose) use the same method for identifying the optimal salvage treatment given the estimated optimal frontline treatment, it remains to be shown whether both methods choose the same frontline treatment in all samples or at least under certain conditions. Modeling, as before, the intermediate outcomes with least squares models and the response proprtions with logistic regression

models, the integrated g-computation estimator that is conditional on baseline information only, maximized over stage 2, can be written as

$$
\begin{aligned}
&\max_{b_k,b_l}\ \mu\Big(X_i = x, d_i(a_j; b_k, b_l) = 1, \boldsymbol{\theta}, \boldsymbol{\psi}\Big) = \\
&P(R_{1i} = 0|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)\Big\{\mu^D(A_i = a_j, X_i = x, \boldsymbol{\theta}^D)\Big\} \\
&+P(R_{1i} = 1|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)\Big\{\mu^R(A_i = a_j, X_i = x, \boldsymbol{\theta}^R) \\
&\qquad\qquad +\max_{b_k}\ E\Big[\mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j, B_{1i} = b_k\Big]\Big\} \\
&+P(R_{1i} = 2|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)P(R_{2i} = 1|A_i = a_j, X_i = x, \boldsymbol{\psi}_2)\Big\{\mu^C(A_i = a_j, X_i = x, \boldsymbol{\theta}^C) \\
&+\mu^{CP}(A_i = a_j, X_i = x, \boldsymbol{\theta}^{CP}) + \max_{b_l}\ E\Big[\mu^{PD}(A_i, B_{2i}, \bar{X}_i^P, \boldsymbol{\theta}^{PD})\Big|X_i = x, A_i = a_j, B_{2i} = b_l\Big]\Big\} \\
&+P(R_{1i} = 2|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)P(R_{2i} = 0|A_i = a_j, X_i = x, \boldsymbol{\psi}_2)\Big\{\mu^C(A_i = a_j, X_i = x, \boldsymbol{\theta}^C) \\
&\qquad\qquad +\mu^{CD}(A_i = a_j, X_i = x, \boldsymbol{\theta}^{CD})\Big\}, \qquad (3.19)
\end{aligned}
$$

where $E\Big[\mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j, B_{1i} = b_k\Big]$ denotes the integration of $\mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})$ over the distribution of covariates from stage 2 in $\bar{X}_i^R$ that are not available at baseline in $X_i = x$, which we will denote as X_i^R=$\{s \in \bar{X}_i^R | s \notin X_i\}$. Similarly for $E\Big[\mu^{PD}(A_i, B_{2i}, \bar{X}_i^P, \boldsymbol{\theta}^{PD})\Big|X_i = x, A_i = a_j, B_{2i} = b_l\Big]$, with X_i^P=$\{r \in \bar{X}_i^P | r \notin X_i\}$. Most authors who implement Q-learning use a single regression model on $H_i^{(A)}$ for estimating the mean outcome across stage 1 treatments; however, if the same component models from g-computation are used to construct an estimator for $\mathcal{Q}_A\Big(X_i = x, A_i = a_j\Big) = E\Big[H_i^{(A)}|X_i = x, A_i = a_j\Big]$ using the law of total expectation, a Q-learning model for stage 1

could be written as

$$
\begin{aligned}
&E\Big[H_i^{(A)}|X_i = x, A_i = a_j\Big] = \\
&P(R_{1i} = 0|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)\Big\{\mu^D(A_i = a_j, X_i = x, \boldsymbol{\theta}^D)\Big\} \\
&+P(R_{1i} = 1|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)\Big\{\mu^R(A_i = a_j, X_i = x, \boldsymbol{\theta}^R) \\
&\qquad\qquad +E\Big[\underset{b_k}{max}\ \mu^{RD}(A_i, B_{1i} = b_k, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j\Big]\Big\} \\
&+P(R_{1i} = 2|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)P(R_{2i} = 1|A_i = a_j, X_i = x, \boldsymbol{\psi}_2)\Big\{\mu^C(A_i = a_j, X_i = x, \boldsymbol{\theta}^C) \\
&\qquad +\mu^{CP}(A_i = a_j, X_i = x, \boldsymbol{\theta}^{CP}) + E\Big[\underset{b_l}{max}\ \mu^{PD}(A_i, B_{2i} = b_l, \bar{X}_i^P, \boldsymbol{\theta}^{PD})\Big|X_i = x, A_i = a_j\Big]\Big\} \\
&+P(R_{1i} = 2|A_i = a_j, X_i = x, \boldsymbol{\psi}_1)P(R_{2i} = 0|A_i = a_j, X_i = x, \boldsymbol{\psi}_2)\Big\{\mu^C(A_i = a_j, X_i = x, \boldsymbol{\theta}^C) \\
&\qquad\qquad +\mu^{CD}(A_i = a_j, X_i = x, \boldsymbol{\theta}^{CD})\Big\}. \qquad (3.20)
\end{aligned}
$$

Written this way, the similarity between Q-learning and g-computation is even more striking. Even if they are not equivalent under all circumstances, when viewed this way a strong case is made to tailor the salvage therapy when using conditional structural mean models since they are tailored in Q-learning and the g-computation model for frontline treatment closely resembles the Q-learning model. To compare $A^{opt}(X_i = x) = \underset{a_j}{argmax}\ E\Big[H_i^{(A)}|X_i = x, A_i = a_j\Big]$ using (3.20) vs $A^{opt}(X_i = x) = \underset{a_j}{argmax}\Big\{\underset{b_k,b_l}{max}\ \mu\big(X_i = x, d_i(a_j; b_k, b_l) = 1, \boldsymbol{\theta}, \boldsymbol{\psi}\big)\Big\}$ using (3.19), all that remains is to examine whether

$$E\Big[\underset{b_k}{max}\ \mu^{RD}(A_i, B_{1i} = b_k, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j\Big] \qquad (3.21)$$

is equal to

$$\underset{b_k}{max}\ E\Big[\mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j, B_{1i} = b_k\Big] \qquad (3.22)$$

and

$$E\Big[\underset{b_l}{max}\ \mu^{PD}(A_i, B_{2i} = b_l, \bar{X}_i^P, \boldsymbol{\theta}^{PD})\Big|X_i = x, A_i = a_j\Big] \qquad (3.23)$$

is equal to

$$\underset{b_l}{max}\ E\Big[\mu^{PD}(A_i, B_{2i}, \bar{X}_i^P, \boldsymbol{\theta}^{PD})\Big|X_i = x, A_i = a_j, B_{2i} = b_l\Big]. \qquad (3.24)$$

The answers depend on the distributions of $\mu^{RD}(A_i = a_j, B_{1i} = b_k, \bar{X}_i^R, \boldsymbol{\theta}^{RD})$ and $\mu^{PD}(A_i = a_j, B_{2i} = b_l, \bar{X}_i^P, \boldsymbol{\theta}^{PD})$ over X_i^R and X_i^P, respectively. When the stochastic inequality of $\mu^{RD}(A_i = a_j, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})$ for $B_{1i} = b_1$ and $B_{1i} = b_2$ over X_i^R for a given $A_i = a_j$ is large,

$$E\Big[\underset{b_k}{max}\ \mu^{RD}(A_i, B_{1i} = b_k, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j\Big]$$

$$\approx \underset{b_k}{max}\ E\Big[\mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j, B_{1i} = b_k\Big],$$

with the approximation approaching equality as the stochastic inequality grows. This corresponds to a scenario where one stage 2 treatment significantly out performs the other over all of X_i^R. When the stochastic inequality is small and the variances are equal, the same approximation holds when the correlation is large, with the approximation approaching equality as the correlation reaches 1. This corresponds to a scenario where one stage 2 treatment is incrementally better than the other, and the treatment effect is the same over all of X_i^R. In both cases, g-computation yields the same or nearly the same result for $A^{opt}(X_i = x)$ as Q-learning. When the stochastic inequality is small and either i) the correlation departs from 1, ii) the variances are unequal, or iii) both i and ii, then

$$E\Big[\underset{b_k}{max}\ \mu^{RD}(A_i, B_{1i} = b_k, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j\Big]$$

in Equation (3.20) grows larger, putting more emphasis on identifying $A_i = a_j$ as optimal, while

$$\underset{b_k}{max}\ E\Big[\mu^{RD}(A_i, B_{1i}, \bar{X}_i^R, \boldsymbol{\theta}^{RD})\Big|X_i = x, A_i = a_j, B_{1i} = b_k\Big]$$

in Equation (3.19) does not. This corresponds to a scenario where one treatment is incrementally better than the other, on average, and the treatment effect varies over X_i^R. The same arguments hold for (3.23) and (3.24). Therefore, g-computation and Q-learning, even when constructed using the same component models, are not guaranteed to yield the same optimal treatment at each stage. For Q-learning, the use of additional distributional fea-

tures of stage 2 intermediate outcomes can be seen as both a blessing and a curse. While the typical patient may not see much difference in expected utility between the salvage treatments, small segments of the population might and Q-learning raises the importance of the corresponding frontline treatment based on these small segments. Rather than discarding conditional structural mean models as suboptimal, we prefer to view them as robust to extreme stage 2 observations. Additionally, structural mean models do not suffer from nonregulariy issues associated with nested mean models (since the maximum operator is outside of the expectation), facilitating standard large sample theory and the bootstrap for constructing confidence intervals and performing hypothesis tests [6].

The IPTW and g-computation estimators above allow us to identify the optimal treatment regime given patient information. What may not be immediately clear from these estimators is the functional dependence they outline between the patient information and the optimal treatment regime. This is especially true for the g-computation estimator. The variable selection method presented next allows us to identify which of the covariates in these structural mean models are prescriptive, and to describe the functional dependence between these prescriptive variables and the optimal treatment regime.

3.4 PRESCRIPTIVE VARIABLE SELECTION FOR CONDITIONAL STRUCTURAL MEAN MODELS

Expanding on the ideas in Zhang (2014) [37], we propose a two-step approach for prescriptive variable selection in structural mean models. For models of the form in equations (3.8) and (3.9), the first step implements a variable selection method from Section 3.1.1 to identify significant interaction effects between baseline covariate information and treatment regimes when estimating mean survival time. Equation (3.10) is used to create a categorical variable indicating the estimated preliminary optimal treatment regime for each subject, given his/her baseline covariate information. In the second step we use the estimated preliminary optimal treatment regime as the outcome in a classification method such as multinomial lo-

gistic regression, using significant baseline covariates from (3.9) in step 1 as predictors. Any baseline effects deemed significant in the second step are prescriptive variables that qualitatively interact with treatment regime when estimating mean survival time and prescribe the preliminary optimal treatment regime. An analogous two step method can be used for models of the form in equations (3.12) and (3.13).

At this point the reader might be left wondering what the purpose is of the classification method in the second step. After all, we do have everything we need from the first step to assign the preliminary optimal treatment regime conditional on baseline information. If we were estimating marginal means for each treatment regime, we would directly compare the 16 means and identify the largest one with, say, a forest plot. By conditioning on baseline information, there may be more than one optimal regime. We could create a separate forest plot for each combination of baseline covariates, but as the number of baseline covariates increases this becomes tedious, and this may not suggest a clear functional relationship between the baseline covariates and the optimal regime. The importance of the second step is two fold: i) if the variable selection process from the first step included many baseline covariates, the second step narrows our focus to those baseline covariates that are prescriptive ii) once we have narrowed our focus, the second step allows us to see the functional dependence between the preliminary optimal treatment regime and these prescriptive covariates. Both of these points are especially important when using the g-computation estimator.

In a sense we are modeling our model, using a classification method to model the argmax of the g-computation or IPTW estimator. We admit that this extra layer of modeling may introduce additional misclassification than using the g-computation or IPTW estimator alone, but it allows us to clearly and succinctly describe the prescription of our g-computation or IPTW model. It should also be noted that only the prescriptive variables need be collected to prescribe according to the classification model, saving hospital and patient resources. If additional accuracy is desired, the g-computation or IPTW estimator can be used to prescribe the optimal treatment regimes, and the classifier in step two can be used to describe the prescription mechanism.

Regardless of whether g-computation (3.13) or IPTW (3.8) and (3.9) is used, the same two step method can be used when tailoring the stage 2 treatment prescribed by the estimated preliminary optimal regime. The first step implements a quantitative variable selection method from Section 3.1.1 to identify significant interaction effects when estimating the mean sojourn time from stage 2 to death. Equations (3.17) and (3.18) are used to create a categorical variable indicating the estimated optimal stage 2 treatment given the estimated optimal stage 1 treatment and patient information. In the second step we use the estimated optimal stage 2 treatment as the outcome in a classification method such as logistic regression, using information up to stage 2 as predictors. Any effects deemed significant in the second step are prescriptive variables that qualitatively interact with stage 2 treatment when estimating the mean sojourn time from stage 2 to death. Scatter and box plots overlaid with the estimated treatment means, that are grouped and paneled by the prescriptive variables, are used to confirm and report the results.

3.5 SIMULATION

We conducted a simulation experiment to evaluate the optimization of frontline treatment, salvage treatment, and treatment regime given patient information using the methods described in Sections 3.3 and 3.4. Identical to the acute myelogenous leukemia or myelodysplastic syndrome (AML-MDS) trial design presented in Section 3.2, and later in Section 3.6, we consider a 2-stage SMART design.

The scenario was generated to closely mimic the AML-MDS data described in Section 3.6. Subjects were randomly assigned to one of four induction therapies, $\mathcal{A}$={(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA}. The simulated population experienced one of three possible cytogenetic abnormalties with equal probability, and age was distributed using a Weibull distribution, truncated between 20 and 90 years. Response sta-

tus R_{1i} depended on frontline treatment, age, and cytogenentic abnormality, while response status R_{2i} depended on frontline treatment only. If $R_{1i} = 1$, assignment to follow-up therapy $\mathcal{B}$={(0)Other treatment, (1)HDAC} depended on age, while if $R_{2i} = 1$, assignment to follow-up therapy depended on $\log T_i^{CP}$. Sojourn times $I\{R_{1i} = 0\}\log T_i^{D}$, $I\{R_{1i} = 1\}\log T_i^{R}$, $I\{R_{1i} = 1\}\log T_i^{RD}$, $I\{R_{1i} = 2\}\log T_i^{C}$, $I\{R_{1i} = 2\}I\{R_{2i} = 1\}\log T_i^{CP}$, $I\{R_{1i} = 2\}I\{R_{2i} = 1\}\log T_i^{PD}$, $I\{R_{1i} = 2\}I\{R_{2i} = 0\}\log T_i^{CD}$ followed various Weibull distributions, with means depending on frontline treatment, age, cytogenic abnormality, and where appropriate, earlier sojourn times and follow-up therapy.

In this scenario n=1000, n=2000, and n=4000 observations (training data) were simulated, and the g-computation and IPTW regression estimators were fit. For each subject, the estimated preliminary optimal regime conditional on baseline information was identified, and logistic regression (written as logistic^{gcomp} and logistic^{IPTW}) was used to identify the functional dependence between the preliminary optimal regime and baseline covariates. For those subjects whom $R_{1i} = 1$ or $R_{2i} = 1$, the sojourn models for $\log T_i^{RD}$ and $\log T_i^{PD}$, respectively, were evaluated at $\hat{A}_i^{opt}(X_i)$, and $\hat{B}_{1i}^{opt}(\bar{X}_i^R)$ and $\hat{B}_{2i}^{opt}(\bar{X}_i^P)$ were estimated. Logistic regression (also written logistic^{gcomp} and logistic^{IPTW}) was used to identify the functional dependence between the estimated optimal salvage treatment and patient information up to stage 2. The g-computation, IPTW, and classification models were then applied to a new set of n=100,000 observations (test data) to determine how well the models correctly classify subjects to their optimal frontline treatment, salvage treatment, and treatment regime, and to determine how well the models agree with one another. The classification rate is calculated on the n=100,000 subjects. This process is replicated 5,000 times, and the average classification rates are reported in Tables 3.1 through 3.6.

In row 1 of each table, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment and (b) salvage treatment. The remaining rows are interpreted similarly. The results in Tables 3.1 through 3.3 are under correct model specification for the g-computation model, and nearly-

correct model specification for the IPTW model. We say 'nearly' since the data generative process followed the g-computation estimator, hence the IPTW model will not be exactly correct. Correct model specification is to say there was no model selection in either step of the two step method. The correct models were known and fit. This is to give us an idea of how the g-computation, IPTW, logistic^{gcomp}, and logistic^{IPTW} models work under the most ideal circumstances in the given scenario. As expected, the IPTW model and its associated logistic^{IPTW} model were in perfect agreement with one another over 99% of the time when identifying the optimal frontline treatment. This is not surprising since the IPTW estimating equations replicate the observations belonging to multiple regimes, and the mean model fits a separate slope/intercept for every regime. Any covariates that interact with treatment regime can be perfectly captured in the logistic^{IPTW} model. On the other hand, the g-computation model fits separate parameters for the covariates across the mean sojourn times, and the associated logistic^{gcomp} model can not perfectly capture these relationships. Nevertheless, the logistic^{gcomp} model did agree with the g-computation model over 99% of the time as well, on average. The results in Tables 3.4 through 3.6 incorporate backward variable selection using AIC in step one and backward variable selection using significance level in step two of the proposed variable selection method for each of the 5,000 replications. This is to give us a sense of how the g-computation, IPTW, logistic^{gcomp}, and logistic^{IPTW} models work under usual model building circumstances in the given scenario. Backward selection was chosen because there were relatively few covariates at each stage to choose from. Methods such as LASSO work particularly well when there are many candidate variables.

Table 3.1: Agreement rates (se) under correct model specification. 5,000 simulations of n=1,000.

	(a) Frontline treatment	(b) Tailored Salvage treatment
g-comp vs logistic^{gcomp}	99.9% (0.01%)	94.2% (0.11%)
g-comp vs Truth	81.2% (0.19%)	74.3% (0.18%)
logistic^{gcomp} vs Truth	81.2% (0.19%)	74.7% (0.20%)
IPTW vs logistic^{IPTW}	99.8% (0.03%)	90.5% (0.14%)
IPTW vs Truth	73.7% (0.19%)	71.7% (0.17%)
logistic^{IPTW} vs Truth	73.6% (0.19%)	72.8% (0.22%)
g-comp vs IPTW	76.9% (0.18%)	89.1% (0.12%)
IPTW vs logistic^{gcomp}	76.9% (0.18%)	84.3% (0.15%)
g-comp vs logistic IPTW	76.8% (0.18%)	81.2% (0.18%)
logistic^{gcomp} vs logistic^{IPTW}	76.8% (0.18%)	83.0% (0.18%)

In row 1, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment, (b) tailored salvage treatment, and (c) tailored treatment regime. The remaining rows are interpreted similarly.

Table 3.2: Agreement rates (se) under correct model specification. 5,000 simulations of n=2,000.

	(a) Frontline treatment	(b) Tailored Salvage treatment
g-comp vs logistic^{gcomp}	99.9% (0.01%)	95.4% (0.09%)
g-comp vs Truth	86.5% (0.10%)	78.7% (0.15%)
logistic^{gcomp} vs Truth	86.5% (0.10%)	79.3% (0.17%)
IPTW vs logistic^{IPTW}	99.9% (0.02%)	91.9% (0.13%)
IPTW vs Truth	78.4% (0.17%)	75.5% (0.15%)
logistic^{IPTW} vs Truth	78.4% (0.17%)	76.5% (0.19%)
g-comp vs IPTW	80.4% (0.15%)	90.8% (0.10%)
IPTW vs logistic^{gcomp}	80.4% (0.16%)	86.8% (0.12%)
g-comp vs logistic IPTW	80.4% (0.16%)	84.0% (0.16%)
logistic^{gcomp} vs logistic^{IPTW}	80.4% (0.16%)	85.8% (0.16%)

In row 1, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment, (b) tailored salvage treatment, and (c) tailored treatment regime. The remaining rows are interpreted similarly.

Table 3.3: Agreement rates (se) under correct model specification. 5,000 simulations of n=4,000.

	(a) Frontline treatment	(b) Tailored Salvage treatment
g-comp vs logistic^{gcomp}	99.9% (0.01%)	96.3% (0.07%)
g-comp vs Truth	89.0% (0.07%)	81.9% (0.12%)
logistic^{gcomp} vs Truth	89.0% (0.07%)	82.4% (0.13%)
IPTW vs logistic^{IPTW}	99.9% (0.02%)	93.4% (0.10%)
IPTW vs Truth	82.8% (0.14%)	79.4% (0.12%)
logistic^{IPTW} vs Truth	82.7% (0.14%)	80.2% (0.15%)
g-comp vs IPTW	83.9% (0.13%)	92.4% (0.08%)
IPTW vs logistic^{gcomp}	83.9% (0.13%)	89.1% (0.10%)
g-comp vs logistic IPTW	83.9% (0.13%)	86.9% (0.12%)
logistic^{gcomp} vs logistic^{IPTW}	83.9% (0.13%)	88.6% (0.12%)

In row 1, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment and (b) tailored salvage treatment. The remaining rows are interpreted similarly.

Table 3.4: Agreement rates (se) using backward selection for model building. 5,000 simulations of n=1,000.

	(a) Frontline treatment	(b) Tailored Salvage treatment
g-comp vs logistic^{gcomp}	86.4% (0.16%)	94.4% (0.12%)
g-comp vs Truth	80.9% (0.19%)	73.9% (0.20%)
logistic^{gcomp} vs Truth	74.6% (0.17%)	74.5% (0.22%)
IPTW vs logistic^{IPTW}	95.1% (0.17%)	94.7% (0.12%)
IPTW vs Truth	72.7% (0.18%)	70.9% (0.19%)
logistic^{IPTW} vs Truth	72.5% (0.19%)	71.5% (0.21%)
g-comp vs IPTW	75.8% (0.17%)	88.4% (0.13%)
IPTW vs logistic^{gcomp}	71.3% (0.20%)	83.7% (0.16%)
gcomp vs logistic^{IPTW}	75.2% (0.19%)	83.9% (0.16%)
logistic^{gcomp} vs logistic^{IPTW}	72.6% (0.21%)	86.1% (0.17%)

In row 1, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment and (b) tailored salvage treatment. The remaining rows are interpreted similarly.

Table 3.5: Agreement rates (se) using backward selection for model building. 5,000 simulations of n=2,000.

	(a) Frontline treatment	(b) Tailored Salvage treatment
g-comp vs logistic^{gcomp}	89.7% (0.17%)	97.1% (0.07%)
g-comp vs Truth	83.3% (0.11%)	76.9% (0.17%)
logistic^{gcomp} vs Truth	78.5% (0.12%)	77.4% (0.17%)
IPTW vs logistic^{IPTW}	98.2% (0.10%)	97.0% (0.09%)
IPTW vs Truth	78.5% (0.16%)	74.7% (0.17%)
logistic^{IPTW} vs Truth	78.4% (0.16%)	75.1% (0.18%)
g-comp vs IPTW	82.8% (0.14%)	92.0% (0.09%)
IPTW vs logistic^{gcomp}	77.5% (0.17%)	89.5% (0.11%)
gcomp vs logistic^{IPTW}	82.3% (0.14%)	89.5% (0.12%)
logistic^{gcomp} vs logistic^{IPTW}	77.6% (0.17%)	89.9% (0.12%)

In row 1, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment and (b) tailored salvage treatment. The remaining rows are interpreted similarly.

Table 3.6: Agreement rates (se) using backward selection for model building. 5,000 simulations of n=4,000.

	(a) Frontline treatment	(b) Tailored Salvage treatment
g-comp vs logistic^{gcomp}	91.5% (0.15%)	99.1% (0.04%)
g-comp vs Truth	86.3% (0.09%)	80.4% (0.14%)
logistic^{gcomp} vs Truth	80.7% (0.11%)	80.5% (0.14%)
IPTW vs logistic^{IPTW}	99.5% (0.05%)	99.1% (0.05%)
IPTW vs Truth	83.6% (0.14%)	78.7% (0.15%)
logistic^{IPTW} vs Truth	83.5% (0.14%)	78.7% (0.15%)
g-comp vs IPTW	85.5% (0.11%)	93.4% (0.07%)
IPTW vs logistic^{gcomp}	80.6% (0.14%)	92.7% (0.08%)
g-comp vs logistic IPTW	85.4% (0.11%)	92.6% (0.08%)
logistic^{gcomp} vs logistic^{IPTW}	80.6% (0.14%)	92.7% (0.09%)

In row 1, g-computation vs logistic^{gcomp} compares the proportion of times the argmax of the g-computation estimator produces the same result as the classification model of the argmax of the g-computation estimator for (a) frontline treatment and (b) tailored salvage treatment. The remaining rows are interpreted similarly.

3.6 APPLICATION

In this section we apply the methods discussed previously to the AML-MDS trial concerning 210 patients with leukemia [33]. The data set arose from a randomized trial of four combination chemotherapies given as frontline treatments to patients with poor prognosis

acute myelogenous leukemia (AML) or myelo-dysplastic syndrome (MDS). Chemotherapy of AML or MDS proceeds in stages. A remission induction chemotherapy combination is given first, with the aim of achieving a complete remission (CR), which is defined as the patient having less than 5% blast cells, a platelet count greater than 105 mm3 and white blood cell count greater than 103 mm3, based on a bone marrow biopsy. If the induction chemotherapy does not achieve a CR, or a CR is achieved but the patient suffers a relapse, then salvage chemotherapy usually is given in a second attempt to achieve a CR. The AML-MDS trial used a 2×2 factorial design with chemotherapy combinations fludarabine plus cytosine arabinoside plus idarubicin (FAI), FAI plus all-trans-retinoic acid (FAI+ATRA), FAI plus granulocyte colony stimulating factor (FAI+G) and FAI plus all-trans-retinoic acid plus granulocyte colony stimulating factor (FAI+G+ATRA). The primary aim was to assess the effects of adding ATRA, G or both to FAI on the probability of success, which was defined as the patient being alive and in CR at 6 months.

Table 3.7: Initial outcomes following frontline treatment

Group				Total, N
	Death	Resistant Disease	Complete Remission	
	N(%)	N(%)	N(%)	
All patients	69(33)	39(19)	102(48)	210
FAI	17(31)	17(31)	20(37)	54
FAI+ATRA	15(28)	13(24)	26(48)	54
FAI+G	20(38)	4(8)	28(54)	52
FAI+G+ATRA	17(34)	5(10)	28(56)	50

Because there were many different salvage treatments, we classified salvage as either containing high dose arabinoside cytosine (HDAC) or not (Other treatment). In the AML-MDS trial, patients were randomized between the four induction combinations, whereas the salvage treatments B_1 and B_2 were chosen subjectively by the attending physicians, patient by patient. Consequently, considering the multicourse structure of the patients actual therapy,

the data are observational because salvage treatments were not chosen by randomization. By modeling the stage 2 treatment assignment probabilities, incorporating all covariates that explain treatment assignment, the IPTW regression estimator remains consistent. Similarly, by incorporating all confounders of stage 2 treatment assignment into the stage 2 intermediate outcome component models in Equation (3.13), the g-computation estimator also remains consistent. This no unmeasured confounders assumption is important for our causal inference interpretation of counterfactual/potential outcomes, allowing us to consistently estimate the mean outcome under a regime of interest for the entire sample of patients. We found that assignment to B_1 treatments was associated with age, while assignment to B_2 treatments was associated with $\log T^{CP}$. Tables 3.7 and 3.8 summarize the counts for the seven possible events illustrated in Figure 3.2 for the leukemia data. These include the three induction therapy outcomes (indexed by R_{1i}) for each treatment arm and the four possible subsequent outcomes.

Table 3.8: Outcomes following CR or Resistant Disease

Group	Resistant Disease	Complete Remission		
	Death	Death	Progression	Death after Progression
	N(%)	N(%)	N(%)	N(%)
All patients	37(95)	9(9)	93(91)	83(93)
HDAC	25(93)	-	-	47(89)
Other treatment	12(100)	-	-	36(90)

It is well known that age and type of cytogenetic abnormality are highly reliable predictors of the probability of CR and survival time in AML or MDS. In particular, cytogenetic abnormality, characterized by missing portions of the fifth and seventh chromosomes (denoted by (-5,-7)), and older age are strongly associated with a lower probability of CR and shorter survival time. Because this trial's entry criteria required patients to have at least one unfavorable prognostic characteristic, the distributions of age and cytogenetic abnormality

were different from those seen in the population of newly diagnosed AMLMDS patients. For example, only four patients had the comparatively favorable cytogenetic abnormality with an inversion of the 16th chromosome, or T(8,21), a translocation between chromosomes 8 and 21. Consequently, to take advantage of cytogenetic abnormality as a prognostic variable in our regression analyses, we grouped it into three categories: poor {(5,7)}; intermediate {diploid, Y, or insufficient metaphases to classify}; good {+8, 11Q, INV16, T(8,21), MISC}.

To ensure stability of the model fits, six of the seven component models were fitted by restricting the time to the particular event to a fixed upper limit, with the limits set by first examining the observed distribution of each event time. Specifically, the variables U^D, T^C, U^{RD}, U^{CD}, T^{CP}, and U^{PD} were restricted to 100, 110, 1408, 692, 1326 and 2274 days respectively. The parameter estimates and standard errors for the mean sojourn time models are presented in Tables 3.9 and 3.10. Backward variable selection, using AIC as the criterion for optimality, was used to determine the significant covariates and their possible two-way interactions in each model. Frontline and salvage treatment were forced into each model.

Unfortunately, many AML patients undergoing chemotherapy to induce CR die during this process, before either CR is achieved or it can be determined that the patients disease is resistant to the induction chemotherapy. Although such deaths may be attributed to either the leukemia or the chemotherapy, so-called 'regimen-related death', because both the disease and the treatment cause low white blood cell counts and other adverse events it often is very difficult to identify a sole cause of death. The patients in this study were especially susceptible to induction death due to their poor prognosis at entry to the trial, with overall rate of death during induction chemotherapy 33% (69/210), varying from 28% to 38% across the four induction regimens (p-value, 0.70; generalized Fisher exact test). In the fitted model for the three induction event times (Table 3.9), no baseline covariate was significantly associated with $\log T^D$. There did not appear to be any significant difference between the induction treatment effects on $\log T^D$, although ATRA may have had a slightly deleterious effect in that, among the 69 patients who died during induction, the patients in the two ATRA arms died a few days sooner, on average.

Table 3.9: Models for sojourn time to death, time to resistance, and time to complete remission.

	$\log T^D$	$\log T^R$	$\log T^C$
Frontline Treatment	✓	✓	✓
Age			✓
Age × Frontline treatment		✓	

Resistance to induction treatment occurred in 39 (18.6%) patients, relatively more frequently among patients receiving FAI and FAI+ATRA (31% and 24% respectively) compared with those who received FAI+G or FAI+G+ATRA (7.8% and 10% respectively). The times to treatment resistance were similar across the four induction treatments, but with greater variability in the FAI+G arm (Table 3.9). Among the 39 patients who were resistant to front-line treatment, 27 were given HDAC as salvage treatment. Two patients in this cohort were censored before observing death. Using backward variable selection with AIC as the criteria of optimality, factors that were associated with time from induction treatment resistance to death are presented in Table 3.10. About half (48.6%) of the 210 patients achieved CR, with CR rates of 37%, 48%, 53% and 56% in the FAI, FAI+ATRA, FAI+G, and FAI+G+ATRA arms respectively. Of the 102 patients who achieved CR, 93 (91%) had disease progression before death or being lost to follow-up. Among these, 53 (57%) received HDAC as salvage treatment. Since only nine patients died in CR, an intercept-only model was used for modeling T^{CD}.

Table 3.10: Models for sojourn time from resistance to death, complete remission to disease progression, and from progression to death.

	$\log T^{RD}$	$\log T^{CP}$	$\log T^{PD}$
Frontline treatment×Age		✓	
Frontline×Salvage	✓		✓
Age	✓	✓	✓
Cytogenetic group	✓	✓	✓
$\log T^{R}$×age	✓		
Frontline×Cytogenetic			✓
Age× Cytogenetic group		✓	✓
$\log T^{C}$×Frontline		✓	
$\log T^{C}$×Cytogenetic group		✓	
$\log T^{C}$×Salvage			✓
$\log T^{CP}$×Frontline			✓
$\log T^{CP}$×Salvage			✓

Interaction effects imply lower order terms, i.e. there were no nested effects.

3.6.1 Strategy effects

Figure 3.3 shows the results of the classification model for the argmax of the g-computation model using the proposed two step prescriptive variable selection method. Figure 3.3 depicts the proportion of patients, or estimated probability of, having a particular preliminary optimal treatment regime, given baseline information. Using the integrated form of (3.13), Equation (3.14) was the outcome in a logistic regression. The significant baseline covariates from the g-computation model were the candidate variables in another variable selection process in the classification model to determine which covariates are prescriptive. Both cytogenetic abnormality and age are prescriptive when determining a patient's preliminary optimal treatment regime using baseline information. For all of those with poor and intermediate cytogenetic abnormalities, and mostly all of those with good cytogenetic abnormalities, the preliminary regime $d\big((2)\text{FAI+ATRA};\ (0)\text{Other treatment},\ (1)\text{HDAC}\big)$ was optimal. For

those with good cytogenetic abnormality and who are over 70 years of age, the preliminary optimal treatment regime is $d\big((3)\text{FAI+G; }(1)\text{HDAC, }(0)\text{Other treatment}\big)$. This is consistent with the marginal results found in Wahed and Thall (2013), who determined that the optimal dynamic treatment regime, marginalized over baseline information, was $d\big((2)\text{FAI+ATRA; }(0)\text{Other treatment, }(1)\text{HDAC}\big)$. Our IPTW regression model found that $d\big((2)\text{FAI+ATRA; }(0)\text{Other treatment, }(1)\text{HDAC}\big)$ was the preliminary optimal treatment regime for all, regardless of baseline information. This is not surprising since the IPTW and g-computation estimators are very different from one another. For each of the preliminary optimal treatment regimes prescribed using the two-step variable selection method, Figure 3.4 plots the estimated mean survival time from the g-computation model with 90% point-wise bootstrap confidence intervals using the 5^{th} and 95^{th} percentiles of the bootstrapped sampling distribution of the mean (500 bootstrap re-samples). Though the confidence interval is wide, 20 year old patients with intermediate cytogenetic abnormalities are estimated to live over 8 years on average from commencement of regime $d((2)\text{FAI+ATRA; }(0)\text{Other, }(1)\text{HDAC})$, compared to 3 years or less for other ages and cytogenetic groups.

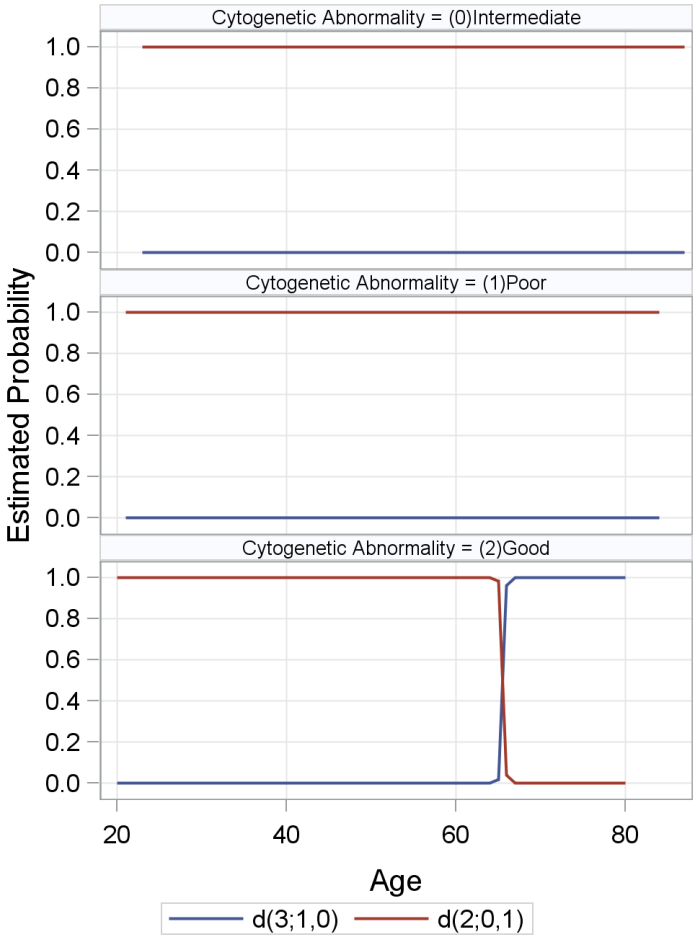

Figure 3.3: Classification model for argmax of g-computation model using the proposed two step prescriptive variable selection method.

Regimes $d(A; B_1, B_2)$, where A=(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA; B_1,B_2=(0)Other treatment, (1)HDAC.

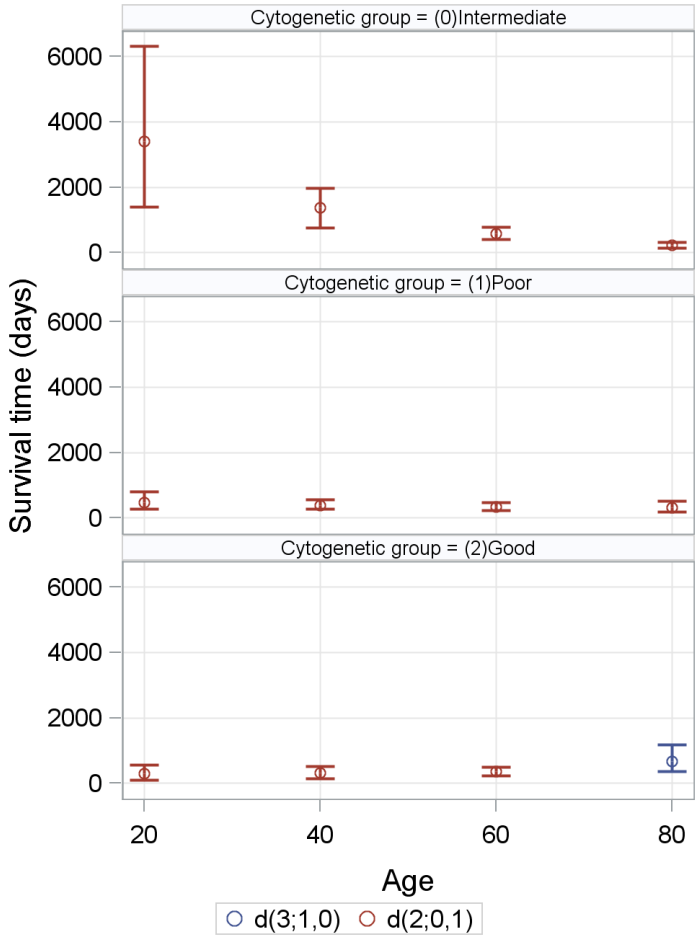

Figure 3.4: Forest plot of g-computation model with 90% point-wise bootstrap confidence intervals.

Regimes $d(A; B_1, B_2)$, where A=(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA; B_1,B_2=(0)Other treatment, (1)HDAC.

For those who experienced disease progression after complete remission, the optimal salvage therapy can be tailored according to Figure 3.5, which shows the results of the classification model for the argmax of the $\log T^{PD}$ model using the proposed two step prescriptive variable selection method. As indicated in Table 3.10, the model for $\log T^{PD}$ depends on age, cytogenetic group, $\log T^{C}$, $\log T^{CP}$, and several interaction effects, but Figure 3.5 shows that only $\log T^{CP}$is needed to prescribe the optimal salvage therapy. Figure 3.5 depicts the

proportion of patients, or estimated probability of, having a particular optimal salvage treatment, given patient information up to stage 2. For those who took (2)FAI+ATRA as their frontline treatment and experienced disease progression after complete remission, (1)HDAC remained their optimal salvage therapy so long as their T^{CP} was greater than 6 logs. However, for patients with logT^{CP} less than 6 following treatment with (2)FAI+ATRA, most had (0)Other treatment as their optimal salvage therapy. A similar result holds for those treated with frontline therapy (3)FAI+G, except that the decision point to alter the salvage treatment occurs near logT^{CP}=5. Since the IPTW regression model identified $d\big($(2)FAI+ATRA; (0)Other treatment, (1)HDAC$\big)$ as the preliminary optimal treatment regime, regardless of baseline information, its corresponding graph for tailoring the salvage treatment when $R_{2i} = 1$ is the top panel of Figure 3.5. For those who experienced resistant disease, no further tailoring of the optimal salvage treatment was possible, since the optimal salvage therapy was Other treatment for everyone experiencing resistant disease. It should be noted that the logistic curves in Figure 3.5 are more smooth than the near-discontinuous curves in Figure 3.3. This indicates that there are other covariates that aide in the prescription of the optimal salvage therapy, but were not deemed significant by the backward variable selection process of the classification model. When age is included in the classification model for the argmax of the logT^{PD} model, the logistic curves in Figure 3.5 become sharper when paneled by age, though the same functional dependence and overall prescription remain the same. For each of the optimal salvage therapies, Figure 3.6 plots the estimated mean survival time from disease progression for the g-computation model with 90% point-wise bootstrap confidence intervals using the 5^{th} and 95^{th} percentiles of the bootstrapped sampling distribution of the mean (500 bootstrap re-samples).

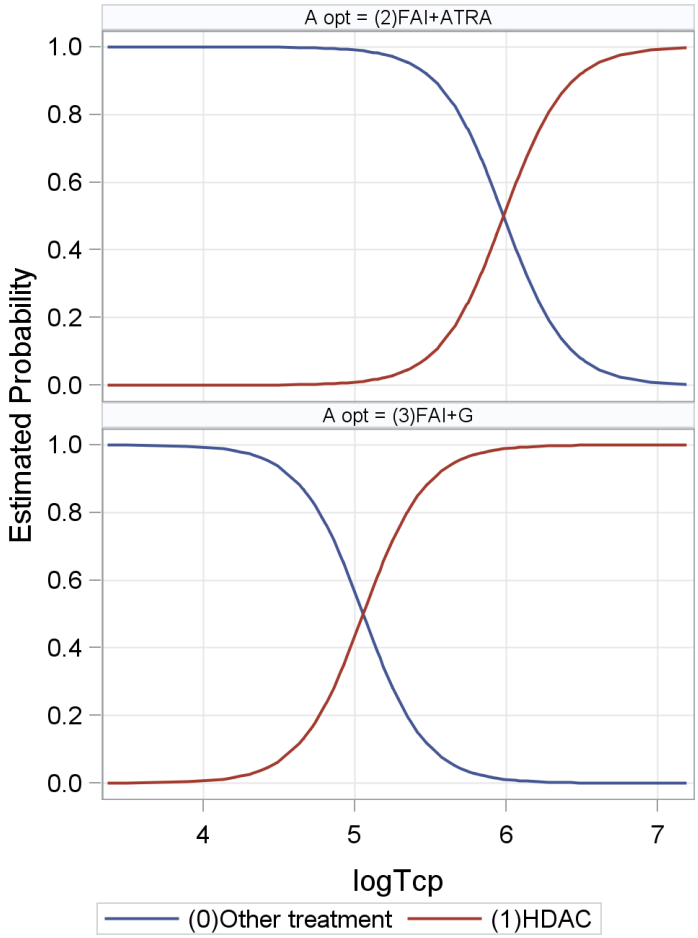

Figure 3.5: Classification model for argmax of $\log T^{PD}$ model using the proposed two step prescriptive variable selection method.

B_2=(0)Other treatment, (1)HDAC.

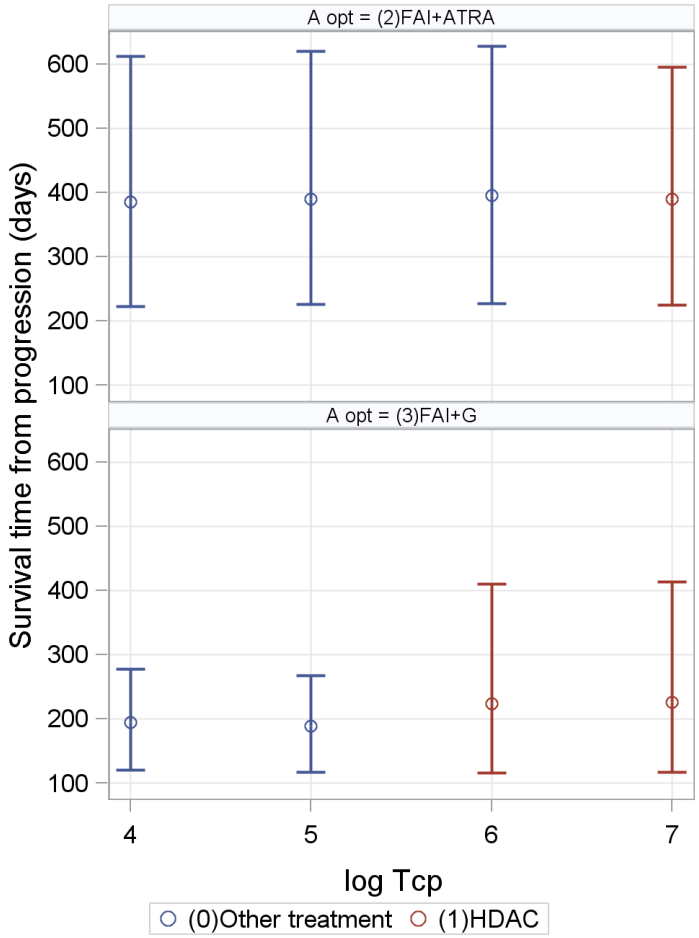

Figure 3.6: Forest plot for $\log T^{PD}$ model with 90% point-wise bootstrap confidence intervals. B_2=(0)Other treatment, (1)HDAC.

To further showcase the proposed methods for optimizing a dynamic treatment regime, we next analyze a single simulated data set of n=1,000 patients. Figure 3.7 shows the results of the classification model for the argmax of the g-computation model using the proposed two step prescriptive variable selection method on the n=1,000 simulated patients. Figure 3.7 depicts the proportion of patients, or estimated probability of, having a particular preliminary optimal treatment regime, given baseline information. Both cytogenetic abnormality and age are prescriptive when determining a patient's preliminary optimal treatment regime using baseline information. With five different optimal treatment regimes, each depending

differently on baseline information, the proposed two step variable selection method is indispensable when identifying the functional dependence. Figure 3.9 shows the results of the classification model for the argmax of the IPTW model using the proposed two step prescriptive variable selection method on the n=1,000 simulated patients. Figure 3.9 depicts the proportion of patients, or estimated probability of, having a particular preliminary optimal treatment regime, given baseline information. At first glance it may seem that the prescription offered by IPTW is noticeably different from that offered by g-computation. This is not surprising since each model offers a different functional dependence between covariate information and mean survival time. However, both methods prescribe nearly the same optimal frontline treatment given baseline information. Regardless of whether IPTW or g-computaion is used, when many of the preliminary optimal regimes share the same frontline treatment, one can sum the estimated proportions from the classification method in step 2 according to frontline treatment and report the optimal frontline treatment; however, it is important to classify on the preliminary optimal treatment regime, not the optimal frontline treatment, in order to capture the functional dependence with the covariates (Figures 3.8 and 3.10). For example, if two preliminary optimal treatment regimes shared the same frontline treatment, one for young patients and the other for old patients, while a third preliminary treatment regime (with a different frontline treatment) was optimal for middle aged patients, classifying on optimal frontline treatment using a linear term for age would report an average proportion across age. This of course could be remedied by considering higher ordered terms for age, but when the number of covariates increases this becomes cumbersome.

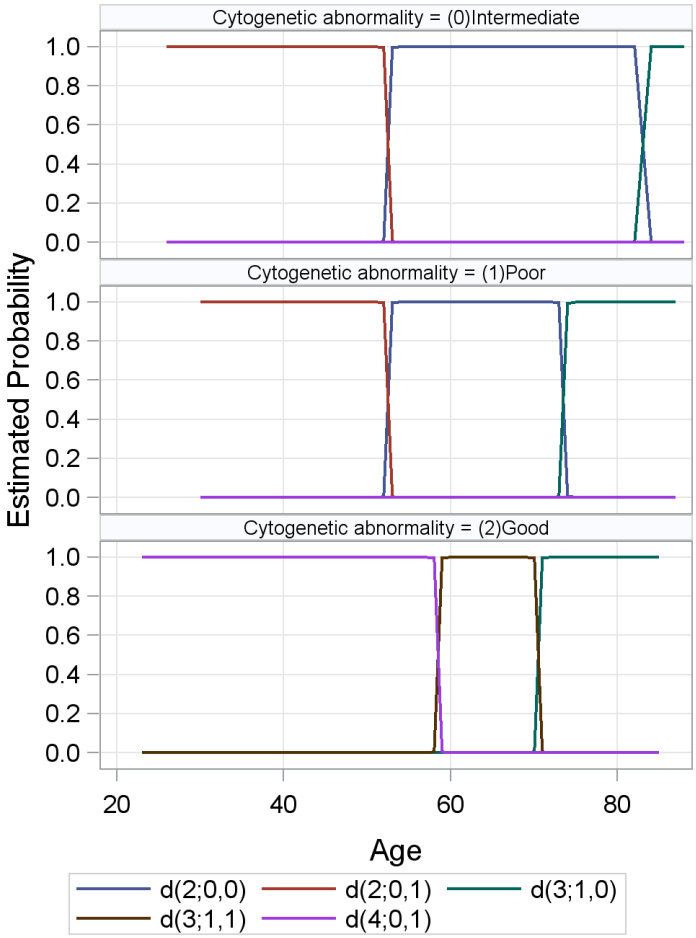

Figure 3.7: Classification model for argmax of g-computation model using the proposed two step prescriptive variable selection method on n=1,000 simulated patients.

Regimes $d(A; B_1, B_2)$, where A=(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA; B_1,B_2=(0)Other treatment, (1)HDAC.

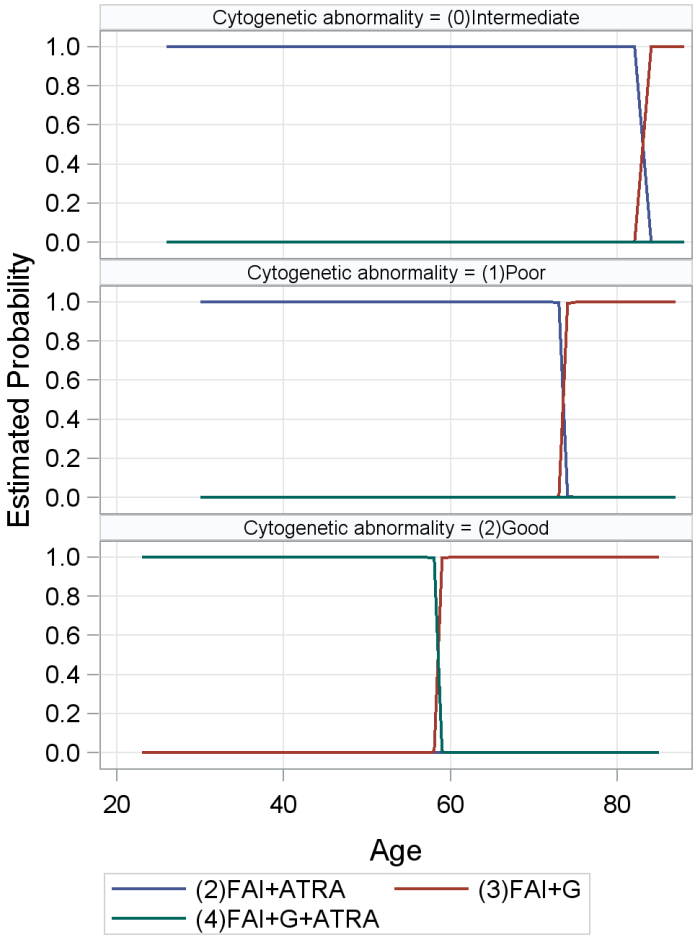

Figure 3.8: Classification model for argmax of g-computation model using the proposed two step prescriptive variable selection method on n=1,000 simulated patients.

$A^{opt}(X_i)$=(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA.

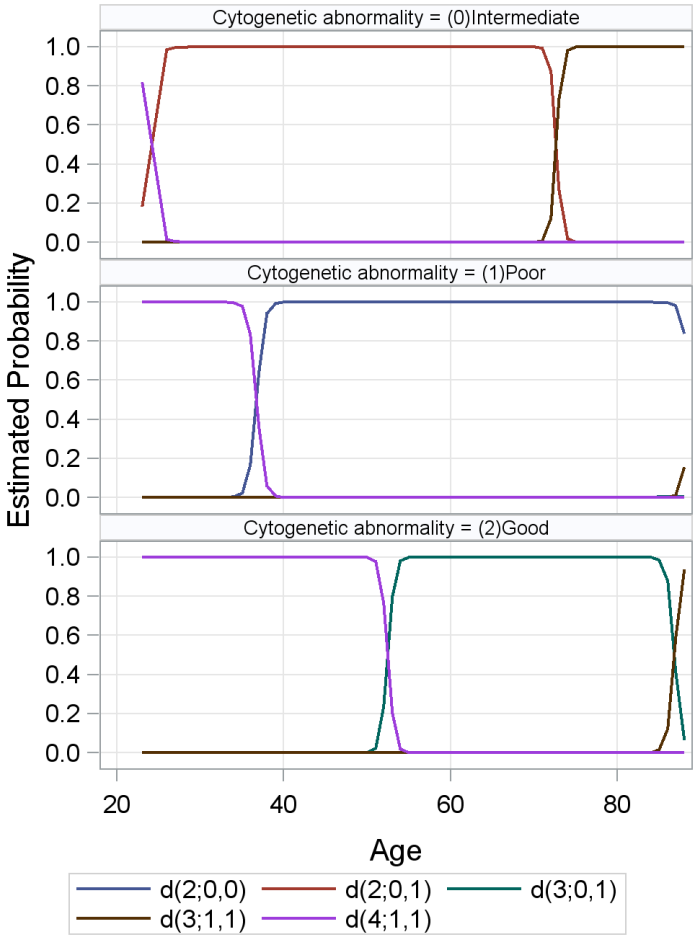

Figure 3.9: Classification model for argmax of IPTW model using the proposed two step prescriptive variable selection method on n=1,000 simulated patients.

Regimes $d(A; B_1, B_2)$, where A=(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA; B_1,B_2=(0)Other treatment, (1)HDAC.

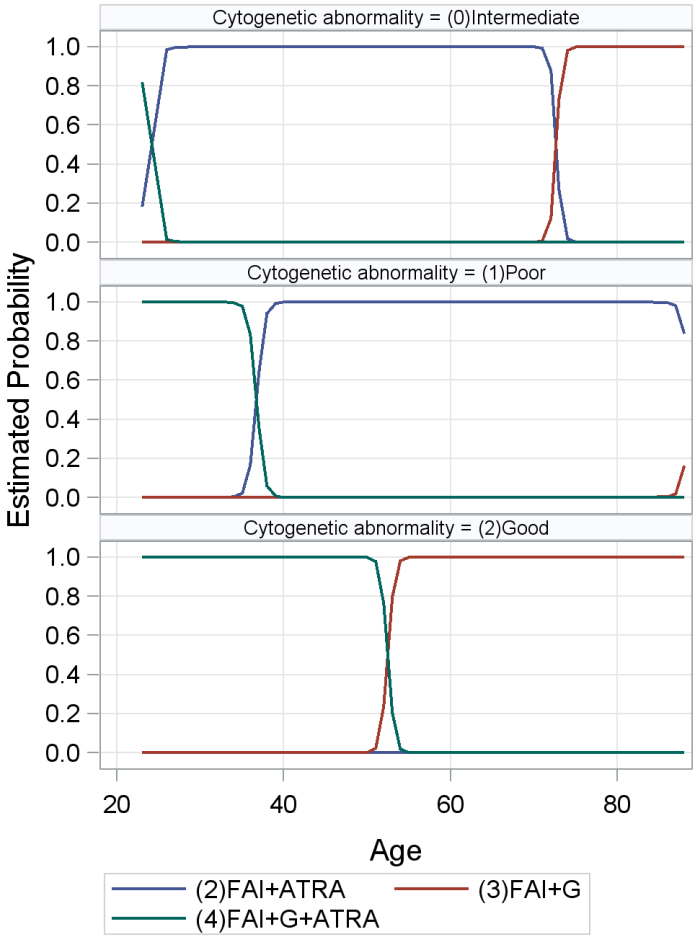

Figure 3.10: Classification model for argmax of IPTW model using the proposed two step prescriptive variable selection method on n=1,000 simulated patients.

A$^{opt}(X_i)$=(1)FAI, (2)FAI+ATRA, (3)FAI+G, or (4)FAI+G+ATRA.

For those simulated patients who experienced disease progression after complete remission, the optimal salvage therapy can be tailored according to Figure 3.11, which shows the results of the classification model for the argmax of the logT^{PD} model using the proposed two step prescriptive variable selection method, with $A^{opt}(X_i)$ estimated using g-computation. Figure 3.11 depicts the proportion of patients, or estimated probability of, having a particular optimal salvage treatment, given patient information up to stage 2. It is quite similar to what was found in the AML-MDS data, except that here (1)HDAC is preferred at smaller logT^{CP}.

Figure 3.12 shows the results of the classification model for the argmax of the logT^{PD} model using the proposed two step prescriptive variable selection method, with $A^{opt}(X_i)$ estimated using IPTW. Figure 3.12 depicts the proportion of patients, or estimated probability of, having a particular optimal salvage treatment, given patient information up to stage 2. In this case, the backward variable selection process of the classification model included age, which further tailors the prescription of optimal salvage treatment by adjusting the decision point along logT^{CP}.

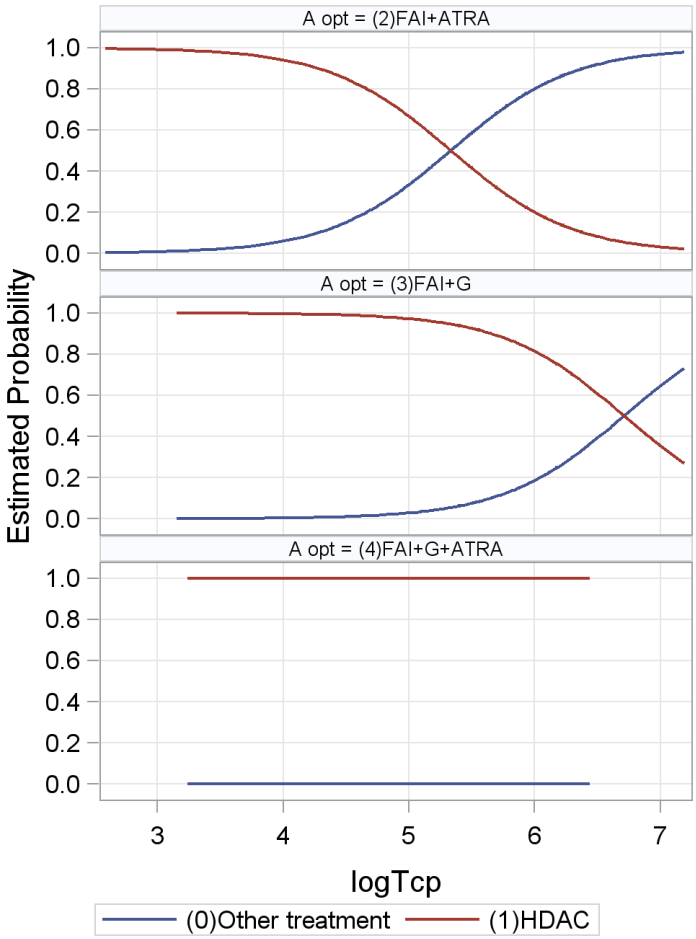

Figure 3.11: Classification model for argmax of logT^{PD} model using the proposed two step prescriptive variable selection method on n=1,000 simulated patients.

$A^{opt}(X_i)$ estimated using g-computation. B_2=(0)Other treatment, (1)HDAC.

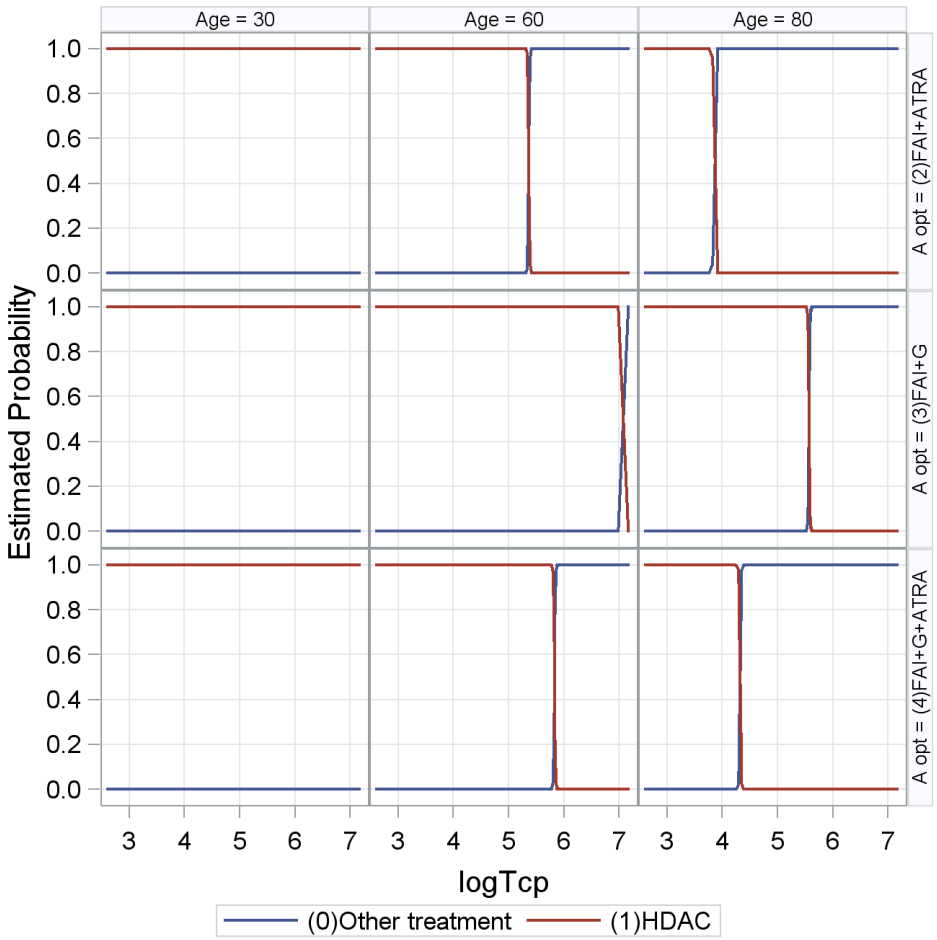

Figure 3.12: Classification model for argmax of $\log T^{PD}$ model using the proposed two step prescriptive variable selection method on n=1,000 simulated patients.

$A^{opt}(X_i)$ estimated using IPTW. B_2=(0)Other treatment, (1)HDA

3.7 CLOSING REMARKS

The methods for optimizing a DTR described herein, though for a specific two stage SMART design, can be easily generalized to other SMART designs. For SMARTs with an arbitrary number of stages the same framework holds. Conditional on baseline covariates, a prelimi-

nary optimal regime is estimated using the g-computation or IPTW estimator. Conditional on information up to stage two (including frontline treatment assignment and response status prior to stage 2), the g-computation or IPTW estimator is used again to estimate the mean outcome for each treatment regime over the remaining stages (stage 2 and onwards). This process continues until the last stage, where the g-computation or IPTW estimator reduces to a simple regression comparing the last stage treatment assignment. Each successive g-computation or IPTW estimator tailors the optimal treatment assignment at the current stage and provides a strategy for the remaining stages, given past treatment assignment and patient data. One might be inclined to compare structural mean models to the Bellman equations and declare them suboptimal. We prefer to view them as an alternative method that is robust to extreme observations at later stages when choosing the optimal treatment at the current stage. They are particularly useful in the event that no future patient information becomes available. The proposed variable selection method is easily applied at every stage of the SMART design.

With structural mean models like g-computation or IPTW estimators, each treatment regime must be directly compared to determine the optimal one. When this comparison is also conditional on patient information, this technique for optimizing dynamic treatment regimes becomes overwhelming. The proposed two step prescriptive variable selection procedure supports the tailored optimization of dynamic treatment regimes using conditional structural mean models by eliminating from consideration any suboptimal treatment regimes and sifting out the covariates that prescribe the optimal treatment regimes. The weighting techniques of the g-computation and IPTW estimators allow an appropriate comparison of the treatment regimes, while avoiding the non-regularity issues of pseudo data associated with backwards induction techniques. This facilitates standard large sample theory and the bootstrap for constructing confidence intervals and performing hypothesis tests.

4.0 FUTURE WORK: CONDITIONAL STRUCTURAL COX MODELS FOR OPTIMIZING DTRS

4.1 COX PROPORTIONAL HAZARDS MODEL AND VARIABLE SELECTION

For observations $\{(U_i, \Delta_i, X_i),\ i = 1, 2, ..., n\}$, U_i=min(T_i, C_i) is the observed event time where T_i is the time to death, C_i is the time to censoring, and $\Delta_i = I\{T_i < C_i\}$ is the death indicator. Define $N_i(u)$=$I(U_i \leq u, \Delta_i = 1)$, $N(u)$=$\sum_{i=1}^{n} N_i(u)$ is the number of deaths up to time u, $Y_i(u)$=$I(U_i \geq u)$, and $Y(u) = \sum_{i=1}^{n} Y_i(u)$ is the number at risk at time u. Then, the Nelson-Aalen estimator for the cumulative hazard of death, $\Lambda(t) = \int_0^t \lambda(t)$, can be written as

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\sum_{i=1}^{n} Y_i(u)},$$

where

$$\lambda(t) = \lim_{h \to 0^+} \frac{P[t \leq T_i < t + h | T_i \geq t]}{h}.$$

T_i and C_i are assumed to be conditionally independent, so that the cause-specific hazard is equal to the hazard for T_i. Adjusting for covariates under the proportional hazards assumption, the Cox model [7] takes the form

$$\lambda(t|X_i, \boldsymbol{\beta}) = \lambda_0(t)\exp(X_i^T \boldsymbol{\beta}),$$

where X_i is a p-dimensional vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of parameters.

To construct the likelihood for $\boldsymbol{\beta}$ under the Cox model while treating time as continuous, we divide the time axis from $[0, t]$ into m intervals so that t_0=0, t_1=t/m, t_2=$2t/m$,...,t_m=t,

and also let $\Delta N_i(t_j)$=$I(U_i \in [t_j, t_{j+1}), \Delta_i = 1)$ and $\Delta(t_j)$=$t_{j+1} - t_j$. We consider intervals that are arbitrarily small so that $\Delta N(t_j)$=$\sum_{i=1}^{n} \Delta N_i(t_j)$ is at most 1. Further define $\mathcal{F}(t_j)$=$\sigma\{I\{U_i \leq u, \Delta_i = 1\}, I\{U_i \leq u, \Delta_i = 0\}, X_i; u \leq t_j, i = 1, 2, ..., n\}$ as the filtration containing all of the survival and censoring information up to time t_j, as well as patient information X_i. Conditional on $\mathcal{F}(t_j)$ we know who died or was censored prior to t_j and their event times, and we know the individuals at risk right after t_j, which we denote as $Y(t_j^+)$. What we do not know is $\Delta N_i(t_j)$=$I(U_i \in [t_j, t_{j+1}), \Delta_i = 1)$, whether or not the i^{th}subject will die in $[t_j, t_{j+1})$, which is distributed as Bernoulli$\big(Y_i(t_j^+)\lambda(t_j|X_i, \boldsymbol{\beta})\Delta(t_j)\big)$. By further conditioning on $\Delta N(t_j)$ (to help us eliminate the nuisance baseline hazard from the resulting likelihood), the distribution of $\Delta N_i(t_j)|\Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)$ can be written as

$$P\big[\Delta N_i(t_j) = r \big| \Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)\big]$$

$$= \begin{cases} \left[\dfrac{\lambda(t_j|X_i, \boldsymbol{\beta})\Delta(t_j)}{\sum_{i=1}^{n} Y_i(t_j^+)\lambda(t_j|X_i, \boldsymbol{\beta})\Delta(t_j)}\right]^r \\ \times \left[1 - \dfrac{\lambda(t_j|X_i, \boldsymbol{\beta})\Delta(t_j)}{\sum_{i=1}^{n} Y_i(t_j^+)\lambda(t_j|X_i, \boldsymbol{\beta})\Delta(t_j)}\right]^{1-r}, & \text{for } r = 0, 1; \text{ if } \Delta N(t_j) = 1 \\ \\ 0^r 1^{1-r}, & \text{for } r = 0, 1; \text{ if } \Delta N(t_j) = 0, \end{cases}$$

and the joint distribution, $P\big[\Delta N_i(t_j) = r, \Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)\big]$, is equal to

$$= P\big[\Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)\big] \times P\big[\Delta N_i(t_j) = r \big| \Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)\big].$$

Therefore, viewing the collection $\big\{\big(\Delta N_i(t_j)|\Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)\big); i = 1, 2, ..., n\big\}$ as a generalized Bernoulli random variable, the likelihood for $\boldsymbol{\beta}$ over all m time intervals is equal to $L(\boldsymbol{\beta})$

$$= \prod_{j=1}^{m-1} P\big[\mathcal{F}(t_j) = f(t_j), \Delta N(t_j) = k\big] \prod_{i=1}^{n} P\big[\Delta N_i(t_j) = 1 \big| \Delta N(t_j) = k, \mathcal{F}(t_j) = f(t_j)\big]^{\Delta N_i(t_j)},$$

and the partial likelihood is

$$\begin{aligned}
PL(\boldsymbol{\beta}) &= \prod_{i=1}^{n}\prod_{j=1}^{m-1} P\big[\Delta N_i(t_j)=1 \big| \Delta N(t_j)=k, \mathcal{F}(t_j)=f(t_j)\big]^{\Delta N_i(t_j)} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m-1} \left[\frac{\lambda(t_j|X_i,\boldsymbol{\beta})\Delta(t_j)}{\sum_{i=1}^{n} Y_i(t_j^+)\lambda(t_j|X_i,\boldsymbol{\beta})\Delta(t_j)}\right]^{\Delta N_i(t_j)\Delta N(t_j)} \times \left[0\right]^{\Delta N_i(t_j)\{1-\Delta N(t_j)\}} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m-1} \left[\frac{\lambda(t_j|X_i,\boldsymbol{\beta})\Delta(t_j)}{\sum_{i=1}^{n} Y_i(t_j^+)\lambda(t_j|X_i,\boldsymbol{\beta})\Delta(t_j)}\right]^{\Delta N_i(t_j)} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m-1} \left[\frac{\lambda_0(t_j)\exp(X_i^T\boldsymbol{\beta})\Delta(t_j)}{\sum_{i=1}^{n} Y_i(t_j^+)\lambda_0(t_j)\exp(X_i^T\boldsymbol{\beta})\Delta(t_j)}\right]^{\Delta N_i(t_j)} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m-1} \left[\frac{\exp(X_i^T\boldsymbol{\beta})}{\sum_{i=1}^{n} Y_i(t_j^+)\exp(X_i^T\boldsymbol{\beta})}\right]^{\Delta N_i(t_j)}.
\end{aligned}$$

Taking the log of $PL(\boldsymbol{\beta})$ and differentiating with respect to $\boldsymbol{\beta}$ yields the score function

$$\begin{aligned}
\mathcal{U}_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[X_i - \frac{\sum_{i=1}^{n} X_i Y_i(t_j^+)\exp(X_i^T\boldsymbol{\beta})}{\sum_{i=1}^{n} Y_i(t_j^+)\exp(X_i^T\boldsymbol{\beta})}\right]\Delta N_i(t_j) \\
&= \sum_{i=1}^{n}\int_0^L \left[X_i - \frac{\sum_{i=1}^{n} X_i Y_i(t)\exp(X_i^T\boldsymbol{\beta})}{\sum_{i=1}^{n} Y_i(t)\exp(X_i^T\boldsymbol{\beta})}\right] dN_i(t) \\
&= 0.
\end{aligned}$$

Lastly, the Breslow estimator of the baseline cumulative hazard can be written as

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n}\int_0^t \frac{dN_i(u)}{\sum_{i=1}^{n} Y_i(u)\exp(X_i^T\hat{\boldsymbol{\beta}})},$$

where $\hat{\boldsymbol{\beta}}$ maximizes $PL(\boldsymbol{\beta})$, and $\hat{\Lambda}(t|X_i,\boldsymbol{\beta}) = \hat{\Lambda}_0(t)\exp(X_i^T\hat{\boldsymbol{\beta}})$.

Much of the discussion in Section 3.1.1 regarding variable selection applies to the Cox model as well. The same goodness-of-fit information criteria, such as AIC and BIC can be used for model comparison, though each of these would rely on the likelihood instead of the SSE. Just as before, the p-values resulting from χ^2 tests of model parameters may also be used when choosing between competing models. When using any of the above criteria for model selection, it may be infeasible to systematically search for the best subset of variables

and interactions, simply because of the sheer number of variables available. As mentioned previously, discrete methods for variable selection include forward, backward, and stepwise methods. Other continuous variable selection methods include the least absolute shrinkage and selection operator (LASSO) and its derivatives. In the case of the Cox model, the LASSO based methods rely on a penalized likelihood instead of a penalized SSE. In the remainder of this final chapter we provide the framework for extending the conditional structural mean models in Chapter 3 to the Cox proportional hazards model. The same two step method for prescriptive variable selection as developed in Section 3.4 can also be applied. We consider a specific two stage setting, but the methods described easily extend to other DTR setups.

4.2 STRUCTURAL COX MODELS FOR DYNAMIC TREATMENT REGIMES

4.2.1 Structural Cox Models Conditional on Baseline Information

Considering the same 2-stage SMART design in Section 3.2, the focus will be to estimate $\Lambda_{jkl}(t) = \int_0^t \lambda_{jkl}(u)du = \int_0^t \lambda(u|d_i(a_j;b_k,b_l=1))du$, $j=1,2,3,4$, $k=1,2$, the cumulative hazard of death under a regime of interest. Since our SMART design allows us to confidently assume no unmeasured confounders, each regime cumulative hazard is representative of the cumulative hazard had the entire sample of patients followed that regime. Recall that patients following $d(a_j;b_k,b_l)$ are a mixture of four groups. We can use the data from these patients to infer about $\Lambda_{jkl}(t)$, accounting for the two stages of randomization. If there was no randomization, and if everyone in the sample was treated using the same DTR, we would have used the Nelson-Aalen estimator $\sum_{i=1}^n \int_0^t \frac{dN_i(u)}{Y(u)}$ to estimate $\Lambda(t)$. If there was only one stage of randomization, we would have considered using $\sum_{i=1}^n \int_0^t \frac{Z_{ji}^{(A)} dN_i(u)}{\sum_{i=1}^n Z_{ji}^{(A)} Y_i(u)} = \sum_{i=1}^n \int_0^t \frac{Z_{ji}^{(A)} dN_i(u)/\pi_j^{(A)}}{\sum_{i=1}^n Z_{ji}^{(A)} Y_i(u)/\pi_j^{(A)}}$. To account for the two stages of randomization we consider, just as before, the quantity $W_{jkli} = \frac{Z_{ji}^{(A)}}{\pi_j^{(A)}}\Big(I\{R_{1i}=0\} + I\{R_{1i}=1\}\frac{Z_{ki}^{(B_1)}}{\pi_k^{(B_1)}} + I\{R_{1i}=2\}I\{R_{2i}=1\}\frac{Z_{li}^{(B_2)}}{\pi_l^{(B_2)}} + I\{R_{1i}=2\}I\{R_{2i}=0\}\Big)$. Note that $W_{jkli}dN_i(u)$

can be non-zero only for patients who are treated according to $d(a_j; b_k, b_l)$, and conditional on $\mathcal{F}_{jkl}(u)$, $E\big[\sum_{i=1}^{n} W_{jkli} dN_i(u)\big] = \lambda_{jkl}(u)du \sum_{i=1}^{n} Y_{jkli}(u) = \lambda_{jkl}(u)du E\big[\sum_{i=1}^{n} W_{jkli} Y_i(u)\big]$, where $\sum_{i=1}^{n} Y_{jkli}(u)$ is the number at risk and $\mathcal{F}_{jkl}(u)$ is the filtration at time u for those following regime $d(a_j; b_k, b_l)$. This implies that to find a consistent estimator of $\Lambda_{jkl}(t)$ one need only turn to the weighted Nelson-Aalen estimator

$$\hat{\Lambda}_{jkl}(t) = \sum_{i=1}^{n} \int_0^t \frac{W_{jkli} dN_i(u)}{\sum_{i=1}^{n} W_{jkli} Y_i(u)}.$$

In a basic randomized clinical trial, the cumulative hazard for each treatment group is estimated and compared to see which treatment has the smallest hazard of death. Similarly, the marginal estimator above is useful for comparing the cumulative hazards across treatment regimes to identify which treatment regime has the smallest hazard of death. As in a basic randomized clinical trial, a subgroup analysis can be performed to see if the marginal results hold throughout, or if the optimal treatment regime depends on patient characteristics. Following Tang & Wahed (2013) [32], the estimator for the cumulative hazard can be extended to the regression setting under a stratified proportional hazards assumption to adjust for baseline covariates using a Cox model via the estimating equation

$$\begin{aligned} \mathcal{U}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{j=1}^{4} \sum_{k=1}^{2} \sum_{l=1}^{2} \int_0^L \left[X_i - \frac{\sum_{i=1}^{n} X_i W_{jkli} Y_i(t) \exp(X_i^T \boldsymbol{\beta}_{jkl})}{\sum_{i=1}^{n} W_{jkli} Y_i(t) \exp(X_i^T \boldsymbol{\beta}_{jkl})} \right] W_{jkli} dN_i(t) \\ &= 0, \end{aligned} \tag{4.1}$$

where X_i is a vector of baseline covariates from $G_i^H(0)$ and some or all of the coefficients in $\boldsymbol{\theta} = [\boldsymbol{\beta}_{111}^T, \boldsymbol{\beta}_{112}^T, ..., \boldsymbol{\beta}_{422}^T]^T$ are allowed to differ by treatment regime. This corresponds to the model

$$\lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big) = \lambda_{jkl0}(t) \exp(X_i^T \boldsymbol{\beta}_{jkl}),$$

where the hazard is proportional within each regime, and the effect of the baseline covariates on the hazard within a regime can be quantified by the log hazard ratio parameter $\boldsymbol{\beta}_{jkl}$. Under this model the Breslow estimator of the baseline cumulative hazard within a regime

is

$$\hat{\Lambda}_{jkl0}(t) = \sum_{i=1}^{n} \int_0^t \frac{W_{jkli} dN_i(u)}{\sum_{i=1}^{n} W_{jkli} Y_i(u) \exp(X_i^T \hat{\boldsymbol{\beta}}_{jkl})}, \tag{4.2}$$

where $\hat{\boldsymbol{\beta}}_{jkl}$ solves 4.1, and $\hat{\Lambda}\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big) = \hat{\Lambda}_{jkl0}(t)\exp(X_i^T \hat{\boldsymbol{\beta}}_{jkl})$. Then a comparison of treatment regimes can be carried out using the hazard ratio

$$\begin{aligned} \gamma_{jklj'k'l'}(t|X_i, \boldsymbol{\theta}) &= \frac{\Lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big)}{\Lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_{j'}; b_{k'}, b_{l'}) = 1\big)} \\ &= \frac{\Lambda_{jkl0}(t)\exp(X_i^T \boldsymbol{\beta}_{jkl})}{\Lambda_{j'k'l'0}(t)\exp(X_i^T \boldsymbol{\beta}_{j'k'l'})}, \end{aligned}$$

which is a function of t and baseline patient information X_i. The preliminary optimal treatment regime, the one with the smallest cumulative hazard at time t, is given by

$$d^{opt}(t, X_i) = \{d(a_{j^*}; b_{k^*}, b_{l^*}),\ a_{j^*}, b_{k^*}, b_{l^*} = \underset{a_j, b_k, b_l}{argmin}\ \Lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big)\}. \tag{4.3}$$

We use the term 'preliminary' when referring to an optimal regime that is conditional on baseline information, but marginalized over stage 2 information. The optimal frontline treatment is given by $A^{opt}(t, X_i) = \underset{a_j}{argmin}\ \{\underset{b_k, b_l}{min}\ \Lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big)\}$.

To implement this estimating equation when all of the parameters in $\boldsymbol{\beta}_{jkl}$ differ across each of treatment regimes, one would create sixteen copies of the analysis data set, each wtih a distinct value of W_{jkli}. The indicators $d_i(a_{j'}; b_{k'}, b_{l'})$, where $j' \neq j$ or $k' \neq k$ or $l' \neq l$, would be artificially set to zero so that the observations with non-zero weights in a given copy of the data set belong to only one regime. This effectively replicates the observations that are consistent with more than one regime [5]. These sixteen data sets would then be stacked one on top another and submitted to a software package for sixteen different weighted regressions stratified by treatment regime using $d_i(a_j; b_k, b_l)$, $j = 1, 2, 3, 4$, $k, l = 1, 2$. When treatment assignment is not random, as was the case in Sections 3.5 and 3.6, the treatment assignment probabilities can be modeled using logistic regression. This is important in order to maintain the no unmeasured confounders assumption.

Alternatively, using the law of total probability the hazard of death for a regime of interest that is conditional on baseline information is given by

$$\begin{aligned}
&\Lambda\big(t|X_i, d(A;B_1,B_2)=1\big) = \\
&\quad P(R_{1i}=0|A_i,X_i)\Big\{\Lambda(t|A_i,X_i,R_{1i}=0)\Big\} \\
&\quad +P(R_{1i}=1|A_i,X_i)\Big\{\Lambda(t|A_i,B_{1i},X_i,R_{1i}=1)\Big\} \\
&\quad +P(R_{1i}=2|A_i,X_i)P(R_{2i}=1|R_{1i}=2,A_i,X_i)\Big\{\Lambda(t|A_i,B_{2i},X_i,R_{1i}=2,R_{2i}=1)\Big\} \\
&\quad +P(R_{1i}=2|A_i,X_i)P(R_{2i}=0|A_i,X_i)\Big\{\Lambda(t|A_i,X_i,R_{1i}=2,R_{2i}=0)\Big\}. \qquad (4.4)
\end{aligned}$$

Following the framework in Section 3.3 we aim to construct a model of the cumulative hazard of death for a regime of interest given all patient information. Such a model might resemble

$$\begin{aligned}
&\Lambda\big(t|X_i,\bar{X}_i^R,\bar{X}_i^C,\bar{X}_i^P,d(A;B_1,B_2)=1\big) = \\
&P(R_{1i}=0|A_i,X_i)\Big\{\Lambda(t|A_i,X_i,R_{1i}=0)\Big\} \\
&+P(R_{1i}=1|A_i,X_i)\iint\limits_{s+u=t}\Big\{\Lambda^R(s|A_i,X_i,R_{1i}=1)\times\Lambda^{RD}(u|A_i,B_{1i},\bar{X}_i^R,R_{1i}=1)\Big\}du\ ds \\
&+P(R_{1i}=2|A_i,X_i)P(R_{2i}=1|R_{1i}=2,A_i,\bar{X}_i^C)\iiint\limits_{s+u+v=t}\Big\{\Lambda^C(s|A_i,X_i,R_{1i}=2) \\
&\qquad\times\Lambda^{CP}(u|A_i,\bar{X}_i^C,R_{1i}=2,R_{2i}=1)\times\Lambda^{PD}(v|A_i,B_{2i},\bar{X}_i^P,R_{1i}=2,R_{2i}=1)\Big\}dv\ du\ ds \\
&+P(R_{1i}=2|A_i,X_i)P(R_{2i}=0|A_i,\bar{X}_i^C)\iint\limits_{s+u=t}\Big\{\Lambda^C(s|A_i,X_i,R_{1i}=2) \\
&\qquad\times\Lambda^{CD}(u|A_i,\bar{X}_i^C,R_{1i}=2,R_{2i}=0)\Big\}du\ ds, \qquad (4.5)
\end{aligned}$$

where for example

$$\Lambda^{RD}(u|A_i,B_{1i},\bar{X}_i^R,R_{1i}=1)=\int_0^u\lim_{h\to0^+}\frac{P[w\le T_i^{RD}<w+h|T_i^{RD}\ge w,A_i,B_{1i},\bar{X}_i^R,R_{1i}=1]}{h}dw.$$

More formally, in 4.5 the integral $\iint\limits_{s+u=t} du\ ds$ stands for $\int_o^t\int_{t-s}^0 du\ ds$, and $\iiint\limits_{s+u+v=t} dv\ du\ ds$ stands for $\int_o^t\int_0^{t-s}\int_{t-u-s}^0 dv\ du\ ds$. If $\bar{X}_i^R$ contains information on T_i^R, then the joint cumulative hazard for T_i^R and T_i^{RD} factors into the product of a marginal and a conditional

cumulative hazard. If $\bar{X}_i^R$ does not contain information on T_i^R, then the cumulative hazards for T_i^R and T_i^{RD} are assumed to be independent. Similarly for $R_{2i} = 0, 1$. When treatment assignment is not random, as was be the case in Sections 3.5 and 3.6, all variables that are confounded with treatment assignment should be included in the sojourn cumulative hazard models. This is important in order to maintain the no unmeasured confounders assumption. One can set proportional hazards estimating equations for each of the component models in 4.4 or 4.5. After integrating over probability measure of the stage 2 covariates, the preliminary optimal treatment regime, the one with the smallest cumulative hazard at time t, is given by

$$d^{opt}(t, X_i) = \{d(a_{j^*}; b_{k^*}, b_{l^*}),\ a_{j^*}, b_{k^*}, b_{l^*} = \underset{a_j, b_k, b_l}{argmin}\ \Lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big)\}, \tag{4.6}$$

and the optimal frontline treatment is given by $A^{opt}(t, X_i) = \underset{a_j}{argmin}\ \{\underset{b_k, b_l}{min}\ \Lambda\big(t|X_i, \boldsymbol{\theta}, d_i(a_j; b_k, b_l) = 1\big)\}$.

4.2.2 Tailoring the Salvage Therapy

Regardless of whether g-computation or IPTW is used, to tailor the stage 2 treatment prescribed by the preliminary optimal regime, the cumulative hazard models for the stage 2 sojourn times, i.e. $\Lambda^{RD}(t|A_i, B_{1i}, \bar{X}_i^R = 1, \boldsymbol{\theta}^{RD})$ and $\Lambda^{PD}(t|A_i, B_{2i}, \bar{X}_i^P, R_{1i} = 2, R_{2i} = 1, \boldsymbol{\theta}^{PD})$, can be examined using the estimating equations

$$\sum_{i=1}^{n} \int_0^L \left[X_i - \frac{\sum_{i=1}^n X_i Y_i^{RD}(t) \exp(X_i^T \boldsymbol{\theta}^{RD})}{\sum_{i=1}^n Y_i^{RD}(t) \exp(X_i^T \boldsymbol{\theta}^{RD})} \right] dN_i^{RD}(t) = 0 \tag{4.7}$$

and

$$\sum_{i=1}^{n} \int_0^L \left[X_i - \frac{\sum_{i=1}^n X_i Y_i^{PD}(t) \exp(X_i^T \boldsymbol{\theta}^{PD})}{\sum_{i=1}^n Y_i^{PD}(t) \exp(X_i^T \boldsymbol{\theta}^{PD})} \right] dN_i^{PD}(t) = 0, \tag{4.8}$$

where $N_i^{RD}(u) = I(U_i^{RD} \leq u, R_{1i} = 1, \Delta_i = 1)$, $Y_i^{RD}(u) = I(U_i^{RD} \geq u, R_{1i} = 1)$, $N_i^{PD}(u) = I(U_i^{PD} \leq u, R_{1i} = 2, R_{2i} = 1, \Delta_i = 1)$, $Y_i^{PD}(u) = I(U_i^{PD} \geq u, R_{1i} = 2, R_{2i} = 1)$. By evaluating $\Lambda^{RD}(t|A_i, B_{1i}, \bar{X}_i^R = 1, \boldsymbol{\theta}^{RD})$ and $\Lambda^{PD}(t|A_i, B_{2i}, \bar{X}_i^P, R_{1i} = 2, R_{2i} = 1, \boldsymbol{\theta}^{PD})$ at

$A_i = A^{opt}(X_i)$, the optimal stage 2 treatment given optimal stage 1 treatment can be identified using

$$B_1^{opt}(t, \bar{X}_i^R) = \underset{b_k}{argmin}\ \Lambda^{RD}\left(t|\bar{X}_i^R, A_i = A^{opt}(X_i), B_{1i} = b_k, \boldsymbol{\theta}^{RD}\right) \tag{4.9}$$

and

$$B_2^{opt}(t, \bar{X}_i^P) = \underset{b_l}{argmin}\ \Lambda^{PD}\left(t|\bar{X}_i^P, A_i = A^{opt}(X_i), B_{2i} = b_l, \boldsymbol{\theta}^{PD}\right) \tag{4.10}$$

for $R_{1i} = 1$, and $R_{1i} = 2$ and $R_{2i} = 1$, respectively. The optimal treatment regime using conditional structural Cox models can then be constructed as "Treat with $A^{opt}(t, X_i)$; if resistance is observed, treat with $B_1^{opt}(t, \bar{X}_i^R)$; if disease progression after complete remission is observed, treat with $B_2^{opt}(t, \bar{X}_i^P)$." The beauty of constructing optimal dynamic treatment regimes in this way is that if additional stage 2 patient information is not available, a salvage treatment based on baseline information can still be prescribed using $d^{opt}(t, X_i)$. Although we have demonstrated this technique for optimizing a dynamic treatment regime on a specific two stage SMART design, the methods are easily generalized to other SMART designs with an arbitrary number of stages. The IPTW or g-computation estimator is used at each stage to estimate the preliminary optimal treatment regime given patient information up to the current stage and prior treatment assignment. Essentially this tailors the optimal treatment assignment at the current stage, and provides an optimal strategy for the remaining stages given the information currently available. The IPTW and g-computation estimators reduce to a simple regression model for the final stage. All authors we have encountered who use conditional structural Cox models (IPTW) do so using only baseline information, prescribing the optimal treatment regime using $d^{opt}(t, X_i)$, but naturally it is best to re-evaluate the strategy as more information becomes available. This is what we propose.

If it is indeed possible to create a g-computation estimator using 4.5, or some similar construction to incorporate all patient information, then we also intend to develop a Q-learning model to identify the optimal treatment at each stage using the cumulative hazard of the stage 1 and stage 2 sojourn times as the criteria of optimality. We will then finish this

work by comparing the g-computation and Q-learning models, before demonstrating these methods along with prescriptive variable selection in a simulation and application.

BIBLIOGRAPHY

[1] N. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N. Duez, A. Filiberti, H. Flechtner, S. Fleishman, J. de Haes, et al. The european organization for research and treatment of cancer qlq-c30: a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the national cancer institute*, 85(5):365–376, 1993.

[2] D. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

[3] R. Bellman. Dynamic programming princeton university press. *Princeton, NJ*, 1957.

[4] B. Chakraborty, E. Laber, and Y. Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723, 2013.

[5] B. Chakraborty and S. Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447, 2014.

[6] B. Chakraborty, S. Murphy, and V. Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research*, 2009.

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

[8] E. Estey, P. Thall, S. Pierce, J. Cortes, M. Beran, H. Kantarjian, M. Keating, M. Andreeff, and E. Freireich. Randomized phase ii study of fludarabine+ cytosine arabinoside+ idarubicin±all-trans retinoic acid±granulocyte colony-stimulating factor in poor prognosis newly diagnosed acute myeloid leukemia and myelodysplastic syndrome. *Blood*, 93(8):2478–2484, 1999.

[9] R. Felder-Puig, A. Di Gallo, M. Waldenmair, P. Norden, A. Winter, H. Gadner, and R. Topf. Health-related quality of life of pediatric patients receiving allogeneic stem cell or bone marrow transplantation: results of a longitudinal, multi-center study. *Bone marrow transplantation*, 38(2):119–126, 2006.

[10] R. Gelber, R. Gelman, and A. Goldhirsch. A quality-of-life-oriented endpoint for comparing therapies. *Biometrics*, 45:781–795, 1989.

[11] P. Glasziou, R. Simes, and R. Gelber. Quality adjusted survival analysis. *Statistics in Medicine*, 9:1259–1276, 1990.

[12] A. Goldhirsch, R. Gelber, R. Simes, P. Glasziou, and A. Coates. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *Journal of Clinical Oncology*, 7:36–44, 1989.

[13] L. Gunter, J. Zhu, and S. Murphy. Variable selection for qualitative interactions. *Statistical methodology*, 8(1):42–55, 2011.

[14] M. Hernán, E. Lanoy, D. Costagliola, and J. Robins. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & clinical pharmacology & toxicology*, 98(3):237–242, 2006.

[15] S. Hollon and A. Beck. *Cognitive and cognitive-behavioral therapies. In: Lambert, MJ., editor. Garfield and Bergins Handbook of Psychotherapy and Behavior Change: An Empirical Analysis*. John Wiley & Sons, New York, 5th edition, 2004.

[16] W. Hong, J. Endicott, L. Itri, W. Doos, J. Batsakis, R. Bell, S. Fofonoff, R. Byers, E. Atkinson, C. Vaughan, et al. 13-cis-retinoic acid in the treatment of oral leukoplakia. *New England Journal of Medicine*, 315(24):1501–1505, 1986.

[17] D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[18] X. Huang, J. Ning, and A. Wahed. Optimization of individualized dynamic treatment regimes for recurrent diseases. *Statistics in medicine*, 33(14):2363–2378, 2014.

[19] E. Korn. On estimating the distribution function for quality of life in cancer clinical trials. *Biometrika*, 80:535–542, 1993.

[20] E. Laber, D. Lizotte, M. Qian, W. Pelham, and S. Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225, 2014.

[21] K. Matthay, J. Villablanca, R. Seeger, D. Stram, R. Harris, N. Ramsay, P. Swift, H. Shimada, C. Black, G. Brodeur, et al. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. *New England Journal of Medicine*, 341(16):1165–1173, 1999.

[22] S. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society B*, 65:331–366, 2003.

[23] L. Orellana, A. Rotnitzky, et al. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *The International Journal of Biostatistics*, 6(2):1–49, 2010.

[24] B. Pradhan and A. Dewanji. On induced dependent censoring for quality adjusted lifetime (qal) data in a simple illness–death model. *Statistics & Probability Letters*, 79(20):2170–2176, 2009.

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

[26] J. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

[27] J. Robins, M. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[28] J. Robins, L. Orellana, and A. Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine*, 27(23):4678–4721, 2008.

[29] J. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, pages 297–331. Springer, 1992.

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

[31] R. Song, W. Wang, D. Zeng, and M. Kosorok. Penalized q-learning for dynamic treatment regimes. *arXiv preprint arXiv:1108.5338*, 2011.

[32] X. Tang and A. Wahed. Cumulative hazard ratio estimation for treatment regimes in sequentially randomized clinical trials. *Statistics in biosciences*, 7(1):1–18, 2013.

[33] A. Wahed and P. Thall. Evaluating joint effects of induction–salvage treatment regimes on overall survival in acute leukaemia. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):67–83, 2013.

[34] A. Wahed and A. Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60(1):124–133, 2004.

[35] H. Wang and H. Zhao. Regression analysis of mean quality-adjusted lifetime with censored data. *Biostatistics*, 8(2):368–382, 2007.

[36] L. Wintner, J. Giesinger, A. Zabernigg, M. Sztankay, V. Meraner, G. Pall, W. Hilbe, and B. Holzner. Quality of life during chemotherapy in lung cancer patients: results across different treatment lines. *British journal of cancer*, 109(9):2301–2308, 2013.

[37] N. Zhang. Variable selection for optimal treatment regimes. PhD dissertation, http://repository.lib.ncsu.edu/ir/bitstream/1840.16/9846/1/etd.pdf, North Carolina State University, 2014.

[38] H. Zhao. and A. Tsiatis. A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*, 84:339–348, 1997.

[39] H. Zhao and A. Tsiatis. Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics*, 55:1101–1107, 1999.

[40] H. Zhao and A. Tsiatis. Estimating mean quality adjusted lifetime with censored data. *Sankhya: The Indian Journal of Statistics*, 62:175–188, 2000.

[41] H. Zhao and A. Tsiatis. Testing equality of survival functions of quality-adjusted lifetime. *Biometrics*, 57:861–867, 2001.