HYPOTHESIS TESTING IN SEQUENTIALLY RANDOMIZED DESIGNS THROUGH ARTIFICIAL RANDOMIZATION

by

Semhar B. Ogbagaber

M.S., Virginia Commonwealth University, 2010

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Semhar B. Ogbagaber

It was defended on

July 24th, 2014

and approved by

Dissertation Advisor:

Abdus S. Wahed, PhD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Committee Members:

Stewart Anderson, PhD

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Yu Cheng, PhD

Associate Professor

Department of Statistics

School of Arts and Sciences

University of Pittsburgh

Jordan Karp, MD

Associate Professor

Department of Medicine

School of Medicine

University of Pittsburgh

HYPOTHESIS TESTING IN SEQUENTIALLY RANDOMIZED DESIGNS THROUGH ARTIFICIAL RANDOMIZATION

Semhar B. Ogbagaber, PhD

University of Pittsburgh, 2014

Abstract

An adaptive treatment strategy (ATS) is an outcome-guided algorithm that allows personalized treatment of complex diseases based on patients' disease status and treatment history. Conditions such as AIDS, depression, and cancer usually require several stages of treatment due to the chronic, multifactorial nature of illness progression and management. Sequential multiple assignment randomized (SMAR) designs permit simultaneous inference about multiple ATSs, where patients are sequentially randomized to treatments at different stages depending upon response status. The purpose of the first part of the dissertation is to develop a sample size formula to ensure adequate power for comparing two or more ATSs. Based on a Wald-type statistic for comparing multiple ATSs with a continuous endpoint, we develop a sample size formula and test it through simulation studies. We show via simulation that the proposed sample size formula maintains the nominal power. The proposed sample size formula is applicable to designs with continuous endpoints and will be useful for practitioners while designing SMAR trials to compare adaptive treatment strategies.

Hypothesis testing to compare adaptive treatment strategies are usually based on inverse weighting and g-estimation. However, regression methods that allow for comparison of treatment strategies that flexibly adjusts for baseline covariates are not as straight-forward using these methods due to the fact that one patient can belong to multiple strategies. This poses a challenge for data analysts as it violates basic assumptions of regression modeling of unique group membership. In the second part of the dissertation, we propose an "artificial randomization" technique to make the data appear that each subject belongs to a specific

ATS. This enables treatment strategy indicators to be inserted as covariates in a regression model. The properties of this method are investigated analytically and through simulation.

Public Health Significance: Chronic diseases such as cancer and mental health problems are becoming a major health care burden that present challenges to caregivers and public health officials. Adaptive treatment strategies are a natural way of treating patients as subjects' conditions change repeatedly over a course of treatment. Finding optimal ATS is therefore vital for the benefit of the patient as well as for society to reduce the health care burden. SMAR (sequential multiple adaptive randomized) trials are convenient methods to compare ATSs. In this dissertation, we provide a sample size formula to help design SMARTs. We also introduce an "artificial randomization" technique that would allow researchers to compare strategies in regression based models. These contributions enhance our understanding of debilitating chronic diseases and will help managing them better.

Keywords: Sample size, power, sequential multiple assignment randomized trial, adaptive treatment strategy, artificial randomization, multiple imputation, ANOVA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to acknowledge my great indebtedness to Dr. Abdus S Wahed for being my doctoral advisor. His help and guidance during the various stages of my dissertation was decisive. Thank you for your patience, kindness, enthusiasm and for inspiring me to appreciate statistics. You have helped me to approach what seems insurmountable task with an open mind. Other than research and statistics, I have always enjoyed talking about issues ranging from current affairs to sports.

I am grateful for the Department of Biostatistics for providing me with a graduate student research (GSR) support throughout the years. I would also like to thank Dr. Anderson for being my GSR mentor, being in my dissertation committee and for providing valuable inputs in my work. Besides, I loved hearing his candid stories and life experiences that he would share during our chats. As part of my GSR assignment with the Western Psychiatric Institute and Clinic, I had the pleasure of collaborating with Dr. Jordan Karp whom I would like to appreciate and thank for being in my dissertation committee.

Many people have influenced my life positively. To all friends and loved ones who believed in me and shaped me to be a better person, I love you all. Finally, many thanks to my parents for their continuous and relentless support. For that, I dedicate this work to you. As my dad would always suggest, I hope to pay it forward.

1.0 INTRODUCTION

In this chapter, we introduce a general review of important concepts and established facts that will be referred to repeatedly throughout the dissertation.

1.1 SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS AND ADAPTIVE TREATMENT STRATEGIES

An adaptive treatment strategy (ATS) is an outcome-guided algorithm that allows personalized treatment [14, 20, 23] of complex diseases based on disease status (response, recurrence, remission, relapse) and intermediate treatment history. ATSs are also called dynamic treatment regimes (DTR) due to time-varying nature of the approach [27, 29]. Complex diseases such as AIDS, depression, and cancer usually involve several stages of treatment due to dynamic disease progression. For instance, a patient with depression may benefit if she initiates treatment with citalopram (CIT). Depending on response, she may remain on CIT or switch to or add another medication or psychosocial treatment during the next phase of treatment. An example of an adaptive treatment strategy from a SMART design of a depression trial (STAR*D) (Rush et al., 2004) [33] is, "Treat with CIT for 6-8 weeks; if response is not achieved with CIT; augment with cognitive behavioral therapy (CBT) for 8 weeks; otherwise, continue with CIT for another 8 weeks". Given first stage, $A_j, j = 1, 2$ and second stage treatments, $B_k, k = 1, 2$ and $C_l, l = 1, 2$, one can define ATS as $A_j B_k C_l$ to indicate "Treat with A_j followed by B_k, if response and C_l, if no response". Learning about ATSs can be realized by studying sequences of treatments in a whole trial.

Sequential multiple assignment randomized trials (SMARTs) are multistage trials whereby subjects are sequentially randomized and re-randomized to available treatment options in several stages after enrollment. Randomization of treatments at each stage is dependent on covariate history (such as patient preference) and previous treatment response. Although previous studies have described SMARTs [12, 13], their characterization was formalized by Murphy (2005) [22]. SMARTs allow us to collect data which in turn inform the development of adaptive treatment strategies (ATSs) to guide treatment of future patients. In practice, SMAR designs usually contain two to three stages of treatment.

1.2 WALD TEST AND POWER

A Wald-test based approach for power and sample size computation is useful in a hypothesis testing setting in statistics (Wald, 1943) [41]. Wald (1943) developed the theory for testing multiple parameters when the number of observations is large. From large sample theory, if appropriate assumptions hold, it is established that the maximum likelihood estimator (MLE) $\hat{\Theta}$ of a vector of parameters Θ, has a limiting normal distribution, that is

$$\sqrt{n}\left(\hat{\Theta}-\Theta\right) \xrightarrow{d} N\left(0, I^{-1}\left(\Theta\right)\right)$$

as n $\rightarrow \infty$, where $\xrightarrow{d}$ is convergence in distribution and $I\left(\Theta\right)$ is the information matrix. Because $I\left(\Theta\right)$ is continuous in Θ, then for a sequence of estimators $\hat{\Theta}_n$,

$$I\left(\hat{\Theta}_n\right) \xrightarrow{p} I\left(\Theta\right)$$

as n $\rightarrow \infty$, where $\xrightarrow{p}$ is convergence in probability. By Slutsky's theorem,

$$n\left(\hat{\Theta}_n-\Theta\right)^T I\left(\hat{\Theta}_n\right)\left(\hat{\Theta}_n-\Theta\right) \xrightarrow{d} V^T I\left(\Theta\right) V$$

where r = dim $\left(\Theta\right)$, $V \sim N_r\left(0, I^{-1}\left(\Theta\right)\right)$ and $V^T I\left(\Theta\right) V \sim \chi^2_{r,\alpha}$. To determine the significance of a test, Wald-test statistic is compared with its theoretical frequency distribution, which is

the chi-square distribution with r degrees of freedom under the null-hypothesis (χ^2_r). That is, the Wald test rejects $H_0 : \Theta = \Theta_0$ in favor of $H_1 : \Theta \neq \Theta_0$ where

$$W_n(\Theta_0) = n\left(\hat{\Theta}_n - \Theta_0\right)^T I(\Theta_0)\left(\hat{\Theta}_n - \Theta_0\right) \geq \chi^2_{r,\alpha}$$

has asymptotic level α. Usually, $I(\Theta_0)$ is evaluated by replacing the MLE $\hat{\Theta}_n$ into the Hessian matrix.

We develop a sample size formula based on Wald-type statistic to compare multiple ATSs with a continuous endpoint. The sample size required to test the null hypothesis with specified significance level α and power $1-\beta$ against the alternative is computed by finding the non-centrality parameter λ from a non-central chi-square distribution with r degrees of freedom and then equating it with $n\Theta^T C^T [C\Sigma C^T]^{-1} C\Theta$ where C is a contrast matrix and Σ is the variance-covariance matrix of Θ.

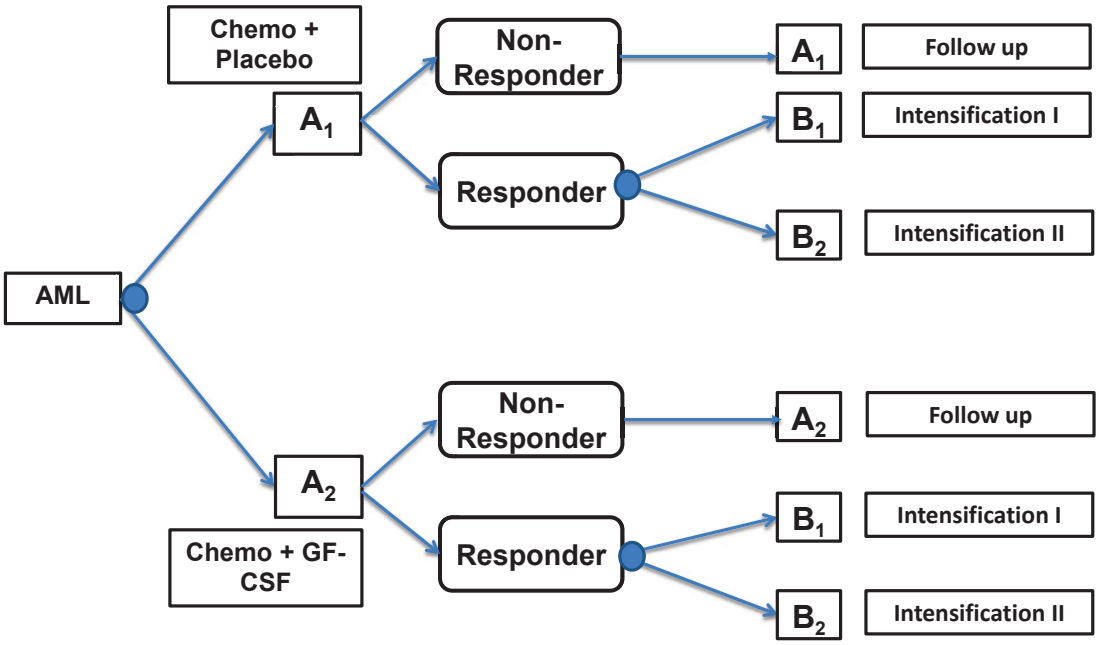
1.3 INVERSE PROBABILITY WEIGHTING

In clinical and biomedical research, missing data is an unavoidable phenomenon. Naive estimation procedures that only consider fully observed data induce bias. Therefore, weighted estimation procedures are suggested. Inverse probability weighting (IPW) for estimating population means in survey sample non-response problems was first described by Horvitz and Thompson (1952) [9]. One such method, namely, the Horvitz-Thompson mean estimator can be written as $\hat{\mu}^{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}$, where $y_i, i = 1, ..., n$ are independent measurements from $n \geq N$ strata with common population mean μ; π_i is the probability that a randomly selected subject belongs to stratum i. A subject weighted by $\frac{1}{\pi_i}$ represents himself/herself as well as missing subjects $\frac{1}{\pi_i} - 1$ in the same stratum. Horvitz-Thompson's mean estimator gives more weight to subjects with more missingness in their strata. A slight variant of Thompson's estimator is the normalized inverse probability weighted estimator, $\hat{\mu}^{IPWN} = \frac{\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}}{\sum_{i=1}^{N}\frac{R_i}{\pi_i}}$, where $R_i = 1$ if y_i is observed and $R_i = 0$ if y_i is missing. Weighting boosts the sample size by expanding the representation of observed data. Applications of IPW has been adequately covered in articles by Robins et al. [28], Murphy et al. [21] and Bembom and van der

Laan [2]. In sequential randomized trials, not everyone in the study follows one of the eight regimes, say $A_jB_kC_l, j, k, l = 1, 2$. Assume a population follows regime $A_1B_1C_1$. Because the estimated strategy mean is biased, inversely weighting responders who were assigned to B_1 and non-responders who were assigned to C_1 by their corresponding probabilities of randomization would correct the bias. Weighting adjusts the mean by creating representation of missing responders (B_2) and non-responders (C_2).

1.4 LEUKEMIA STUDY

Cancer and Leukemia Group B (CALGB) clinical trial (Protocol 8923) was designed for treatment of acute myelogenous leukaemia (AML) among elderly patients of 60 years and older [36]. It was a double-blind, placebo-controlled two-stage trial with two randomizations conducted in the early 1990's. At the first stage, patients were initially randomized to two induction therapies granulocyte-macrophage colony-stimulating factor (GM-CSF) or placebo, following a standard chemotherapy. Response is defined as "complete remission" where complete remission is defined as recovery of morphologically normal bone marrow and blood counts (i.e., neutrophils = 1,500/μl and platelets $<$ 100,000/μl), and no circulating leukemic blasts or evidence of extramedullary leukemia. In the second stage, responders were randomized to two maintenance therapies, namely, intensification therapy I (cytarabine) and intensification therapy II (cytarabine + mitoxantrone). For simplicity we will use the notations, the first stage treatments, A_1: standard chemotherapy + placebo, A_2: standard chemotherapy + GM-CSF; the second stage treatments are: B_1, B_2: intensification treatments I and II. By design, this trial offers second stage randomization to stage 1 treatment responders but not to non-responders. Figure 1 is a schematic of CALGB 8923 study. We use the data from this study to compare the treatment strategies in a regression setting and further generate informative strategies for leukemia treatment.



Patients with AML are randomized to initial treatments A_1 and A_2. If a patient responds to the initial treatment he/she is randomized to either B_1 or B_2, otherwise the patient goes to follow up.

Figure 1: **CALGB 8923 Study.**

1.5 MOTIVATION AND OBJECTIVES

SMAR designs are becoming popular in the context of long term treatment options for chronic diseases. The appeal lies in their naturalistic approach as patient characteristics dictate treatments. Same treatments given to heterogeneous individuals might not have the same effectiveness since, naturally, response is subject-specific (Nahum-Shani et al., 2012) [24]. SMAR trials allow patients to be randomized in multiple stages. Critical decisions are made at each stage of SMAR design when clinically significant event happens during the course of treatment. That is, when patients respond well to sequences of treatments or otherwise.

ATSs are tools that operationalize critical decisions that physicians utilize to manage these complex diseases. ATSs take inputs such as patient characteristics, covariate history, previous treatments and output a treatment choice. Construction of ATSs is made possible via data collected from sequential multiple assignment randomized trials (SMARTs) or observational studies.

The objective of this dissertation is twofold. First, we propose sample size formulas for various SMAR designs when the outcome of interest is continuous. Sample sizes have been developed for survival and continuous outcomes to compare strategies [6, 7, 16]. However, Wallace and Moodie (2014) [42] suggest limiting the number of strategies because of the dangers of creating high dimensional problem and impracticable sample sizes. As such, we also propose a sample size formula for pairwise comparisons that target only desired number of ATSs. Second, we propose the simple multiple artificial randomization tool (SMART) estimator that combines 'artificially randomized' estimators borrowing techniques from multiple imputation methods of Rubin (1977, 1987) [31, 32]. Subsequently, we show how the artificial randomization approach can be used to construct regression models to compare strategies adjusting for baseline covariates, and hence inform further useful strategies.

2.0 DESIGN OF SEQUENTIALLY RANDOMIZED TRIALS FOR TESTING ADAPTIVE TREATMENT STRATEGIES

2.1 INTRODUCTION

An adaptive treatment strategy (ATS) is an outcome-guided algorithm that allows personalized treatment [14, 20, 23] of complex diseases based on disease status (response, recurrence, remission, relapse) and intermediate treatment history. ATS is also called dynamic treatment regimes (DTR) due to time-varying nature of the approach [27, 29]. Complex diseases such as AIDS, depression, and cancer usually involve several stages of treatment due to dynamic disease progression. For instance, a patient with depression may benefit if she initiates treatment with citalopram (CIT). Depending on response, she may remain on CIT or switch to or add another medication or psychosocial treatment during the next phase of treatment [33]. In principle, a clinician monitors a depressed patient and decides on interventions at different time points based on the patient's clinical status. Availability of multiple treatment options at each stage of treatment, various possibilities for the duration of each stage, and various responses that can be achieved through different stages of therapy could lead to a multitude of adaptive treatment strategies. Examples of treatment strategies for a patient with moderate depression include [33]:

1. Treat with CIT for 6-8 weeks; if response is not achieved with CIT, augment with cognitive behavioral therapy (CBT) for 8 weeks, otherwise continue with CIT for another 8 weeks.
2. Treat with CIT for 6-8 weeks; if response is not achieved with CIT, switch to CBT for 8 weeks, otherwise switch to BUS (buspirone) for another 8 weeks.

3. Treat with CIT for 6-8 weeks; if response is not achieved with CIT, switch to CBT for 8 weeks, otherwise switch to BUP-SR (bupropion sustained release) for another 8 weeks.
4. Treat with CIT for 6-8 weeks; if response is not achieved with CIT, switch to SERT (sertraline) for 8 weeks, otherwise switch to CBT for another 8 weeks.

Adaptive treatment strategies are often compared via sequential multiple assignment randomized (SMAR) designs [12, 13, 22]. Even though SMAR trials are useful for comparing ATSs because different ATSs can be tested from the same experimental design and the procedure for inference about ATSs from data arising from such trials are well-established, the design issues (e.g. sample size and power) have not been adequately addressed. This may be due to the challenges posed by the adaptive and sequential nature of SMAR designs. Nevertheless, a few articles have alluded to the development of sample size formula for SMAR designs.

Murphy [22] provides a sample size formula to test the equality of two strategies that do not share same initial sets of treatments, making data from two groups of patients following these strategies statistically independent. Feng and Wahed [6] also constructed a sample size formula for survival outcomes. However, their formula was developed for censored survival times to test equality of point-wise survival probabilities under two ATSs that have the same initial, but different second stage treatments. They also proposed another formula based on weighted log-rank test for the equality of survival curves under two strategies that share different initial treatments [7]. Recently, Li and Murphy [16] presented a sample size formula for survival data to relax the assumptions set forth by Feng and Wahed [6, 7]. Oetting et al. [25] established four sample size formulas, of which only two are relevant to adaptive treatment strategies. One of the formulas (referred to as #3 in their chapter) deals with a hypothesis testing the equality of a pair of strategy means. The other relevant formula (referred to as #4 in their chapter) is developed with the goal of finding the best strategy (as opposed to hypothesis testing comparing multiple strategies). Dawson and Lavori [3] also devised a sample size formula for the nested structure of successive SMAR randomizations when the outcome is continuous. They extended the sample size for the usual t-test to be applicable to SMAR trials. Using a semi-parametric approach, their formula includes stage-specific variance inflation factor (VIF) and marginal outcome variance σ_Y^2. Due to the

presence of between-strategy covariance, one cannot make inference for a pair of strategy means that share the same initial treatments by just pooling the VIF's and marginal outcome variances across the stages. As a remedy, Dawson and Lavori [4] proposed a conservative approach to adjust the sample size formula using the VIF. The caveat with their approach is its difficulty of application. It involves cumbersome computation of all stage-specific VIFs, σ_Y^2 and coefficient of determination by regressing the final outcome on previous states. More recent simulation work by Ko and Wahed [11] considered the power for detecting differences between multiple strategy means for arbitrary sample sizes for a two-stage SMAR design.

Most of the work related to sample sizes in SMAR trials are either confined to two-strategy comparisons [12, 7] or require assumptions about population parameters that are difficult to ascertain (e.g. VIF's and stage-specific variances) in multi-strategy comparison settings. We derive sample size formulas for a variety of SMAR designs in order to test specific alternative hypotheses related to continuous outcomes.

The rest of this chapter is organized as follows. In Section 2.2, we introduce notation and consider three SMAR designs that are being used in various disease areas. The parameters needed to be specified in advance correspond to well-defined subgroups in the patient population and hence are relatively simple to specify. In Section 2.3, we show estimation of strategy means and their corresponding variances and covariances given some fundamental assumptions. Based on a Wald-type statistic, in Section 2.4 we derive an overall sample size formula required to compare multiple strategy means. In Section 2.5, we introduce a t-test based sample size formula to power pairwise comparisons among all or selected ATSs. In Section 2.6, we validate our sample size formulas through simulation experiments. We conclude the chapter with a discussion in Section 2.7.

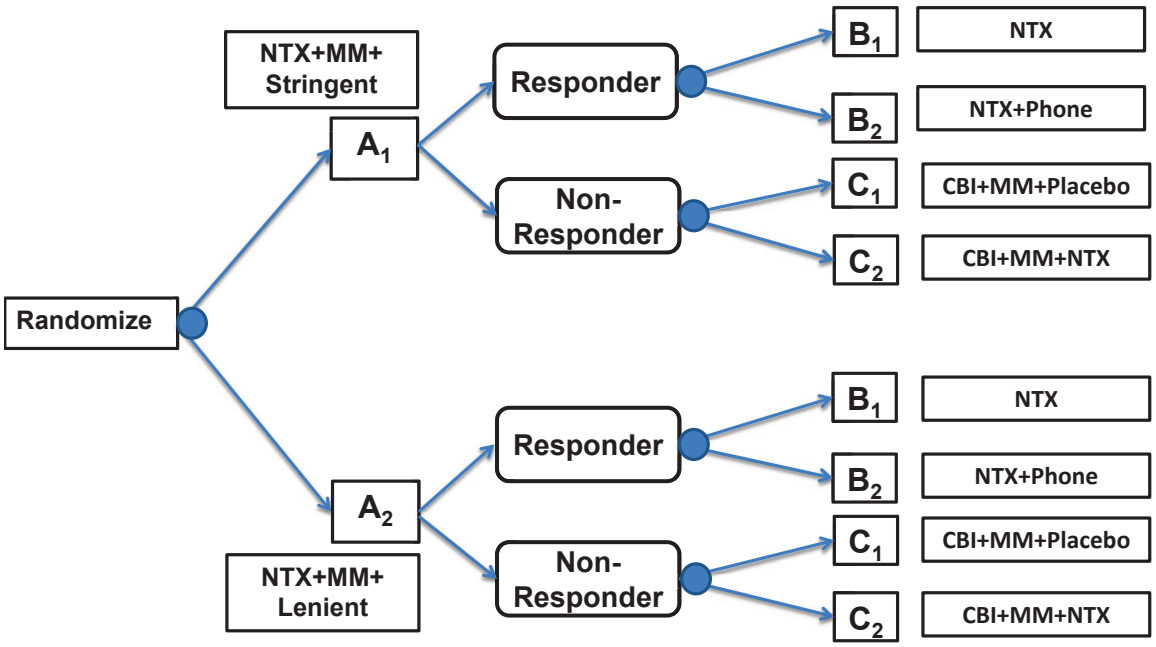
2.2 SET-UP

We consider three two-stage SMAR designs. Figures 2, 3 and 4 display the three SMAR designs. In the first design, n subjects are to be randomized to two induction treatments $A_j, j = 1, 2$. Then maintenance treatments, $B_k, k = 1, 2$, are to be administered randomly

if they responded to induction treatments, or else they are randomized to $C_l, l = 1, 2$. We use the Lei et al. [14] design for alcohol-dependence interventions as an example to explain the first design (Figure 1). All patients are provided with "NTX+MM" as their initial intervention (NTX = naltrexone, MM = medical management). Then patients are randomized to two groups based on how the intermediate response to "NTX+MM" would be ascertained. In one group, referred to as A_1, the response criteria would be stringent (2+ days of heavy drinking), whereas in the other group, referred here forth as A_2, the criterion would be lenient (5+ days of heavy drinking). Following eight weeks of treatment, participants are randomized to the second line treatments depending on their non-response status. Non-responders were re-randomized to either "NTX" (B_1) or "NTX+TDM" (B_2), otherwise, they were re-randomized to two maintenance treatments: "CBI+MM+Placebo" (C_1) or "CBI+MM+NTX" (C_2), where CBI = combined behavioral intervention, and TDM = telephone disease management. At the end of the study, the primary outcome (defined as "percent of heavy drinking days" over the last two months of the study) was obtained.

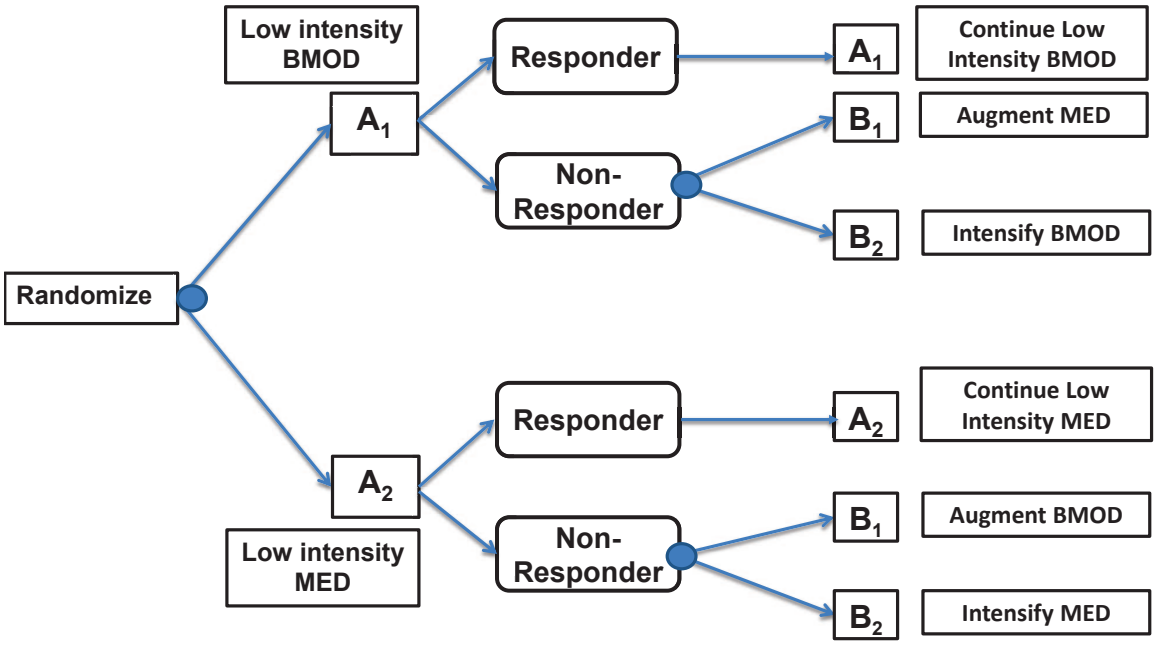
Design 1 allows inference related to eight possible ATSs, namely $A_jB_kC_l, j, k, l = 1, 2$, where $A_jB_kC_l$ stands for "Treat with A_j followed by B_k if they respond, or by C_l if not". For example, one might want to test the equality of all strategy means $H_0 : \mu_{111} = \mu_{112} = \mu_{121} = \mu_{122} = \mu_{211} = \mu_{212} = \mu_{221} = \mu_{222}$, where μ_{ijk} is the mean response under strategy $A_jB_kC_l, j, k, l = 1, 2$ against the alternative of at least one pair being different. Testing equality of any combination of treatment strategies (e.g. pairwise comparisons) may also be of interest. In the sequel we consider the sample sizes required to test varieties of treatment strategy comparisons with adequate statistical power.

The second design was used by Pelham et al. [26] for an Attention Deficit Hyperactivity Disorder (ADHD) clinical trial (Figure 3). This trial involved treating children with ADHD with behavioral and pharmacological interventions during stage 1. In the first stage participants were randomized to low intensity "psychostimulant drug (low intensity MED)" (A_1) or low intensity "behavioral modification (low intensity BMOD)" (A_2). Behavioral modification consists of school-based, weekend and at-home activity sessions. A child's response to the first line treatment is assessed using Impairment Rating Scale (IRS) and an individualized list of target behaviors (ITB). IRS is a comprehensive measure of improvement in



At entry, patients are randomized to initial treatments A_1 and A_2. If a patient responds to the initial treatment she is randomized to either B_1 or B_2, otherwise the patient is randomized to either C_1 or C_2.

Figure 2: **Design 1.**



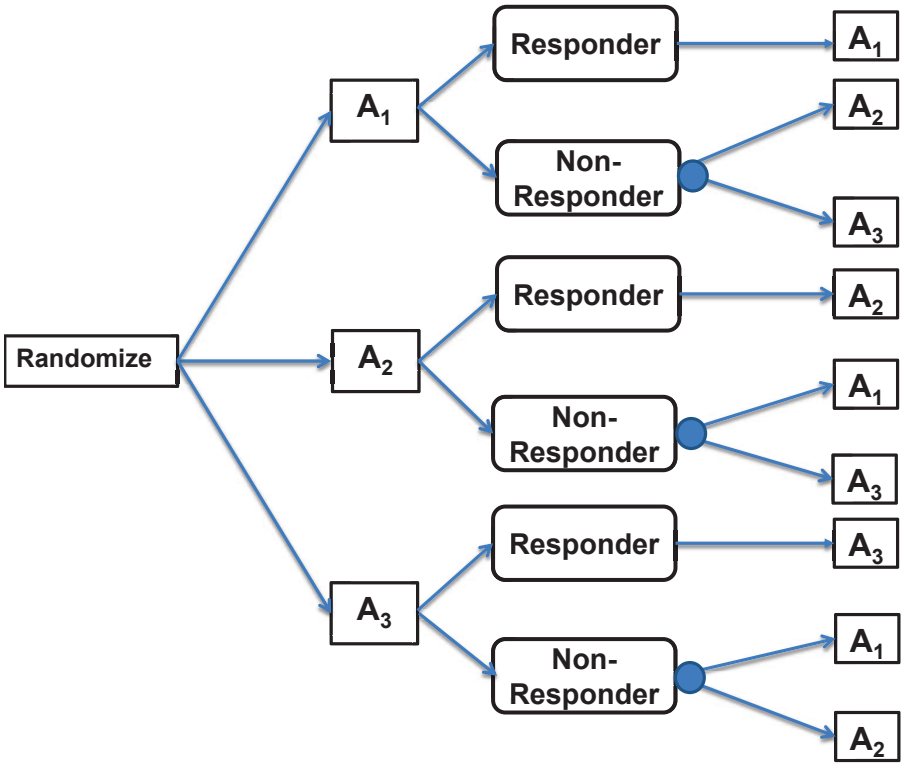
At entry, patients are randomized to initial treatments A_1 and A_2. If a patient responds to the initial treatment she stays on the same initial treatment, otherwise the patient is re-randomized to subsequent treatments: B_1 or B_2 if she does not respond to A_1; similarly, to B_1 or B_2 if she does not respond to A_2.

Figure 3: **Design 2.**

social performance while ITB is a child-specific monitor of social performance. IRS and ITB are "tailoring" variables that determine response status and randomization to the second stage treatments. A response is defined if the child attains an average ITB performance of less than 75% and is reported impaired by teachers on at least one domain in IRS. Based on these tailoring variables, participants who responded to first-stage treatment remained on the same treatment whereas non-responders were re-randomized. Children who did not respond to low intensity BMOD (A_1) were re-randomized to either intensified BMOD (A_1) or BMOD augmented with MED (B_2). Those children who did not respond to low intensity MED (A_2) were re-randomized to either intensified MED (B_1) or MED augmented with BMOD (B_2).

Thus, if a patient responds to A_1 then she stays on A_1 but is randomized to B_1 or B_2 for non-response. Similarly, if a patient responds to A_2 then she stays on A_2; otherwise, she is switched to either B_1 or B_2. Following the notation from the first design, there are 4 possible treatment strategies, namely, $A_1A_1B_1$, $A_1A_1B_2$, $A_2A_2B_1$, or $A_2A_2B_2$. Note that it might seem redundant to use two A_1's or A_2's to define the strategies, but this formulation allows us to cast the problem in the frame of Design 1. It might be of interest to test equality of all 4 strategy means, $H_0 : \mu_{111} = \mu_{112} = \mu_{221} = \mu_{222}$.

The third design considered is described in Thall et al. [39] (Figure 4). Patients received one of three initial treatments A_1, A_2 and A_3 during the first randomization. If a patient initially assigned to A_1 responded, she would remain on A_1 during the second stage; otherwise she would be randomized to A_2 or A_3. Similarly, if a patient responds to initial treatment A_2 then he/she would continue A_2 in the second stage; otherwise would be randomized to A_1 or A_3. Similarly, patients not responding to A_3 would be re-randomized to A_1 or A_2 in the second stage. There are 6 possible regimes, namely, $A_1A_1A_2$, $A_1A_1A_3$, $A_2A_2A_1$, $A_2A_2A_3$, $A_3A_3A_1$, or $A_3A_3A_2$. The null hypothesis of equality of strategy means is, $H_0 : \mu_{112} = \mu_{113} = \mu_{221} = \mu_{223} = \mu_{331} = \mu_{332}$. For all the three designs, we develop a sample size formula to detect meaningful differences between strategy means.



At entry, patients are randomized to initial treatments A_1, A_2 and A_3. If a patient responds to the initial treatment she stays on the same initial treatment, otherwise the patient is re-randomized to subsequent treatments: A_2 or A_3 if she does not respond to A_1; A_1 or A_3 if she does not respond to A_2; A_1 or A_2 if she does not respond to A_3.

Figure 4: **Design 3.**

2.3 COMPARING MULTIPLE TREATMENT STRATEGIES

The goal is to design a sample size formula for a test that detects differences in strategy means from SMAR designs with a continuous endpoint. The formula is based on Design 1 (Figure 2) which is described in Section 2.2. In order to achieve this goal, let us introduce the following notation. Let R_j be the counterfactual response indicator for an individual who responded to $A_j, j = 1, 2$; $Y(A_jB_k)$ is the counterfactual outcome of an individual had he/she received A_j, responded, then took B_k; similarly, $Y(A_jC_l)$ is the counterfactual outcome of an individual had he/she received A_j, did not respond, then took C_l. Based on these three counterfactual outcomes, consider $Y(A_jB_kC_l)$ as the outcome under strategy $A_jB_kC_l$, which can be written as

$$Y(A_jB_kC_l) = R_jY(A_jB_k) + (1 - R_j)Y(A_jC_l), j, k, l = 1, 2. \tag{2.1}$$

To clarify the distinction between the observed and unobserved quantities, for example, for a patient who received A_1, responded, and received B_1, $\{R_2, Y(A_1B_2), Y(A_2B_1), Y(A_2B_2)\}$ are all unobservable. What is observed here is only $Y(A_1B_1)$ (see consistency assumption below). As described in Section 2, we are interested in estimating $\mu_{jkl} = \mathrm{E}\{Y(A_jB_kC_l)\}$. Conditioning on R_j, μ_{jkl} can be expressed as

$$\mu_{jkl} = \pi_j\mu_{A_jB_k} + (1 - \pi_j)\mu_{A_jC_l}, \tag{2.2}$$

where $\mu_{A_jB_k} = \mathrm{E}\{Y(A_jB_k)\}$ is the sub-group mean of the population receiving A_j followed by B_k, $\mu_{A_jC_l} = \mathrm{E}\{Y(A_jC_l)\}$ is sub-group mean of the population receiving A_j followed by C_l. Our development of the sample size formula is based on Wald-type test statistic. Thus, an estimator of the strategy means and corresponding variance and covariance expressions is required. We will rely on the method of normalized inverse probability weighting (IPWN, Ko and Wahed, 2012) to construct unbiased estimator of strategy means. Although we focus on sample size for a continuous endpoint, the formulas developed apply equally for designs with a binary endpoint.

Consider Design 1 (Figure 2) in Section 2.2. Contrary to the counterfactual variables defined above, the observed data for this design consists of i.i.d (independent and identically

distributed) random variables, $(X_{ji}, R_i Z_{ki}, (1-R_i)Z'_{li}, Y_i)$ where $X_{ji} = 1$, if the i^{th} patient is randomized to A_j; 0 otherwise. Y_i is the observed outcome for the i^{th} individual, R_i is the indicator for initial response, $R_i = 1$ if the i^{th} patient responded to initial therapy, 0, otherwise; Z_{ki} is the indicator for receiving B_k, i.e. $Z_{ki} = 1$ if subject i is randomized to receive B_k after responding to the first-stage treatment, 0, otherwise; similarly, Z'_{li} is the indicator for receiving C_l. We make the usual assumptions of causal inference to construct consistent estimators for μ_{jkl} [29]. They are:

A1. Consistency Assumption: A patient's counterfactual outcome under the observed intervention (exposure) and the observed outcome agree. In the SMAR trial considered here,

$$R_i = X_{1i}R_{1i} + (1 - X_{1i})R_{2i}, \tag{2.3}$$

and

$$Y_i = X_{1i}\left[R_{1i}Y_i(A_1B_1) + (1-R_{1i})Y_i(A_1C_1)\right] + (1-X_{1i})\left[R_{2i}Y_i(A_2B_1) + (1-R_{2i})Y_i(A_2C_2)\right], \tag{2.4}$$

where R_{1i} and R_{2i} are indicators for counterfactual response to A_1 and A_2, respectively, defined previously. The consistency assumption (CA) allows us to connect counterfactual and observed data.

A2. Sequential Randomization Assumption: The probability of a particular treatment allocation at stage a at a treatment time k does not depend on the counterfactual outcome given observed data up to but not including stage k randomization. This assumption follows since treatments are assigned randomly at each stage.

A3. Positivity: There is a non-zero probability of receiving any level of intervention for every combination of values of interventions.

Under these assumptions, we define the normalized weighted inverse probability estimator for strategy mean μ_{jkl} is given by

$$\hat{\mu}_{jkl}^{IPWN} = \frac{\sum_{i=1}^{n} W_{jkli} Y_i}{\sum_{i=1}^{n} W_{jkli}}, \tag{2.5}$$

where $W_{jkli} = X_{ji}\left\{\frac{R_i Z_{ki}}{P_k} + \frac{(1-R_i)Z'_{li}}{Q_l}\right\}$, X_{ji} is the assignment indicator for first-stage treatment A_j; P_k and Q_l are probabilities of second treatment assignment for responders and non-responders.

Estimator (2.5) is similar to that in Ko and Wahed (2012) (Section 3.3) except that it treats the group sample sizes in Stage 1 as random rather than being treated as fixed. This is more reasonable because the group sizes in Stage 1 are determined through randomization. The IPWN estimator, $\hat{\mu}_{jkl}^{IPWN}$, defined in Equation (2.5) is consistent and asymptotically normal. This can be shown as follows. We can write,

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{jkl}^{IPWN} - \mu_{jkl}) &= \sqrt{n}\left[\frac{\sum_{i=1}^n W_{jkli}Y_i}{\sum_{i=1}^n W_{jkli}} - \mu_{jkl}\right] \\ &= n^{-1/2}\frac{\sum_{i=1}^n W_{jkli}(Y_i - \mu_{jkl})}{\frac{1}{n}\sum_{i=1}^n W_{jkli}}.\end{aligned}$$

By the weak law of large numbers, $\frac{1}{n}\sum_{i=1}^n W_{jkli} \xrightarrow{p} \frac{1}{\kappa_j}$ where κ_j is the inverse of the randomization probability to A_j (i.e., $\kappa_j = \frac{1}{P(X_{ji}=1)}$). This follows from the fact that W_{jkli}'s are i.i.d random variables with expectation,

$$\begin{aligned}\mathrm{E}\{W_{jkli}\} &= \mathrm{E}\left[\mathrm{E}\left\{\frac{R_i Z_{ki}}{P_k} + \frac{(1-R_i)Z'_{li}}{Q_l}\right\}X_{ji}|X_{ji}, R_i\right] \\ &= \mathrm{E}\left[X_{ji}\mathrm{E}\left\{\frac{R_i Z_{ki}}{P_k} + \frac{(1-R_i)Z'_{li}}{Q_l}\right\}|R_i\right] \\ &= \mathrm{E}\left[X_{ji}\mathrm{E}\{R_i + (1-R_i)\}|R_i\right] \\ &= \mathrm{E}\left[X_{ji}\right] \\ &= P(X_{ji}=1) \\ &= \frac{1}{\kappa_j}.\end{aligned}$$

Also, by the central limit theorem, $n^{-1/2}\sum_{i=1}^n W_{jkli}(Y_i - \mu_{jkl}) \xrightarrow{d} N(0, \sigma_{jkl}^2)$, where σ_{jkl}^2 is given in Equation (2.7) below. Therefore, by Slutsky's theorem, $\sqrt{n}(\hat{\mu}_{jkl}^{IPWN} - \mu_{jkl})$ is asymptotically equivalent in distribution to $n^{-1/2}\kappa_j\sum_{i=1}^n W_{jkli}(Y_i - \mu_{jkl})$ which is normally distributed as $N(0, \sigma_{jkl}^2)$, where σ_{jkl}^2 is defined below. It can also be shown that,

$$\sqrt{n}(\hat{\mu}_{jkl}^{IPWN} - \mu_{jkl}) = n^{-1/2}\sum_{i=1}^n \psi_{jkli} + o_p(1), \tag{2.6}$$

where $\psi_{jkli} = \kappa_j W_{jkli}(Y_i - \mu_{jkl})$ is the influence function of the estimator $\hat{\mu}_{jkl}^{IPWN}$ and $o_p(1)$ is a term that converges to zero in probability. Therefore, the asymptotic variance of $\sqrt{n}(\hat{\mu}_{jkl}^{IPWN} - \mu_{jkl})$ is given by,

$$\sigma_{jkl}^2 = \kappa_j \left[\frac{\pi_j}{P_k} \{\sigma_{A_j B_k}^2 + (1-\pi_j)^2 (\mu_{A_j B_k} - \mu_{A_j C_l})^2\} + \frac{1-\pi_j}{Q_l} \{\sigma_{A_j C_l}^2 + \pi_j^2 (\mu_{A_j B_k} - \mu_{A_j C_l})^2\} \right], \tag{2.7}$$

where π_j is the response rate for first stage treatment A_j; $\sigma_{A_j B_k}^2$ and $\sigma_{A_j C_l}^2$ are variances of the outcome in the population of patients who received the sequence of treatments $A_j B_k$ and $A_j C_l$, respectively; $\mu_{A_j B_k}$ and $\mu_{A_j C_l}$ are defined as before.

Following Ko and Wahed (2012), the variance formula in Equation (2.7) is derived as follows, $\sigma_{jkl}^2 = var(\psi_{jkli}) = var(\kappa_j W_{jkli}(Y_i - \mu_{jkl})) = \kappa_j^2 E\left[W_{jkli}(Y_i - \mu_{jkl})\right]^2$. This variance can be expressed in terms of subgroup-specific population parameters. For example, consider $\hat{\mu}_{111}^{IPWN}$. In this case, the weight is defined as $W_{111i} = X_{1i} \left\{ \frac{R_i Z_{1i}}{P_1} + \frac{(1-R_i) Z_{1i}^{(')}}{Q_1} \right\}$, and therefore, $W_{111i}^2 = X_{1i} \left\{ \frac{R_i Z_{1i}}{P_1^2} + \frac{(1-R_i) Z_{1i}^{'}}{Q_1^2} \right\}$. Then,

$$E\left[\kappa_1^2 W_{111i}^2 (Y_i - \mu_{111})^2\right] = \kappa_1^2 E\left[X_{1i} \left\{ \frac{R_i Z_{1i}}{P_1^2} + \frac{(1-R_i) Z_{1i}^{'}}{Q_1^2} \right\} (Y_i - \mu_{111})^2 \right] \tag{2.8}$$

.

Under assumptions (A1)-(A3), using a series of conditional expectations, we can show that

$$E\left[\kappa_1^2 W_{111i}^2 (Y_i - \mu_{111})^2\right] = \kappa_1 \left[\frac{\pi_1}{P_1} \left\{\sigma_{A_1 B_1}^2 + (\mu_{111} - \mu_{A_1 B_1})^2\right\} + \frac{(1-\pi_1)}{Q_1} \left\{\sigma_{A_1 C_1}^2 + (\mu_{111} - \mu_{A_1 C_1})^2\right\} \right].$$

$$
\begin{aligned}
&= \kappa_1^2 E\left[X_{1i}\left\{\frac{R_i Z_{1i}}{P_1^2}+\frac{(1-R_i)Z'_{1i}}{Q_1^2}\right\}(Y_i-\mu_{111})^2\right] \\
&= \kappa_1^2 EE\left[X_{1i}\left\{\frac{R_i Z_{1i}}{P_1^2}\right\}\{R_iY_i(A_1B_1)+(1-R_i)Y_i(A_1C_1)-\mu_{111}\}^2|R_i,X_{1i},Y_i(A_1B_1),Y_i(A_1C_1)\right]+ \\
&\quad \kappa_1^2 EE\left[X_{1i}\left\{\frac{(1-R_i)Z'_{1i}}{Q_1^2}\right\}\{R_iY_i(A_1B_1)+(1-R_i)Y_i(A_1C_1)-\mu_{111}\}^2|R_i,X_{1i},Y_i(A_1B_1),Y_i(A_1C_1)\right] \\
&= \kappa_1^2 E\left[Z_{1i}|R_i,X_{1i},Y_i(A_1B_1)\right]EE\left[X_{1i}\left\{\frac{R_i}{P_1^2}\right\}\{R_iY_i(A_1B_1)+(1-R_i)Y_i(A_1C_1)-\mu_{111}\}^2|R_i,X_{1i},Y_i(A_1B_1),Y_i(A_1C_1)\right]+ \\
&\quad \kappa_1^2 E\left[Z'_{1i}|R_i,X_{1i},Y_i(A_1C_1)\right]EE\left[X_{1i}\left\{\frac{(1-R_i)}{Q_1^2}\right\}\{R_iY_i(A_1B_1)+(1-R_i)Y_i(A_1C_1)-\mu_{111}\}^2|R_i,X_{1i},Y_i(A_1C_1)\right] \\
&= \kappa_1^2 E\left[Z_{1i}|R_i,X_{1i},Y_i(A_1B_1)\right]E\left[\frac{X_{1i}}{P_1^2}E\left[R_iY_i^2(A_1B_1)-2\mu_{111}\{R_iY_i(A_1B_1)+(1-R_i)Y_i(A_1C_1)\}+\mu_{111}^2|R_i,X_{1i}\right]\right]+ \\
&\quad \kappa_1^2 E\left[Z'_{1i}|R_i,X_{1i},Y_i(A_1C_1)\right]E\left[\frac{X_{1i}}{Q_1^2}E\left[(1-R_i)Y_i^2(A_1C_1)-2\mu_{111}\{R_iY_i(A_1B_1)+(1-R_i)Y_i(A_1C_1)\}+\mu_{111}^2|R_i,X_{1i}\right]\right] \\
&= \kappa_1^2 P_1 E\left[\frac{X_{1i}}{P_1^2}E\left[R_iY_i^2(A_1B_1)-2\mu_{111}\{R_iY_i(A_1B_1)\}+\mu_{111}^2|R_i,X_{1i}\right]\right]+ \\
&\quad \kappa_1^2 Q_1 E\left[\frac{X_{1i}}{Q_1^2}E\left[(1-R_i)Y_i^2(A_1C_1)-2\mu_{111}\{(1-R_i)Y_i(A_1C_1)\}+\mu_{111}^2|R_i,X_{1i}\right]\right] \\
&= \kappa_1^2 P_1 E\left[\frac{X_{1i}}{P_1^2}E\left[R_i(\sigma^2_{A_1B_1}+\mu^2_{A_1B_1})-2\mu_{111}R_i\mu_{A_1B_1}+\mu_{111}^2|R_i,X_{1i}\right]\right]+ \\
&\quad \kappa_1^2 Q_1 E\left[\frac{X_{1i}}{Q_1^2}E\left[(1-R_i)(\sigma^2_{A_1C_1}+\mu^2_{A_1C_1})-2\mu_{111}(1-R_i)\mu_{A_1C_1}+\mu_{111}^2|R_i,X_{1i}\right]\right] \\
&= \kappa_1^2 E\left[X_{1i}E\left[\frac{R_i}{P_1}\left\{\sigma^2_{A_1B_1}+\mu^2_{A_1B_1}-2\mu_{A_1B_1}\mu_{111}+\mu_{111}^2\right\}+\frac{(1-R_i)}{Q_1}\left\{\sigma^2_{A_1C_1}+\mu^2_{A_1C_1}-2\mu_{A_1C_1}\mu_{111}+\mu_{111}^2\right\}|X_{1i}\right]\right] \\
&= \kappa_1^2 E\left[X_{1i}\left[\frac{\pi_1}{P_1}\left\{\sigma^2_{A_1B_1}+(\mu_{111}-\mu_{A_1B_1})^2\right\}+\frac{(1-\pi_1)}{Q_1}\left\{\sigma^2_{A_1C_1}+(\mu_{111}-\mu_{A_1C_1})^2\right\}\right]\right] \\
&= \kappa_1\left[\frac{\pi_1}{P_1}\left\{\sigma^2_{A_1B_1}+(\mu_{111}-\mu_{A_1B_1})^2\right\}+\frac{(1-\pi_1)}{Q_1}\left\{\sigma^2_{A_1C_1}+(\mu_{111}-\mu_{A_1C_1})^2\right\}\right].
\end{aligned}
$$

Consequently, the asymptotic variance of $\hat{\mu}_{111}$ is given by,

$$
\begin{aligned}
var(\hat{\mu}_{111}^{IPWN}) &= \frac{\kappa_1}{n}\left[\frac{\pi_1}{P_1}\left\{\sigma^2_{A_1B_1}+(\mu_{111}-\mu_{A_1B_1})^2\right\}+\frac{(1-\pi_1)}{Q_1}\left\{\sigma^2_{A_1C_1}+(\mu_{111}-\mu_{A_1C_1})^2\right\}\right] \\
&= \frac{\sigma^2_{111}}{n}.
\end{aligned}
$$

Estimators that share the same first-stage treatment are correlated as they use a common group of observations. Consider $\hat{\mu}_{111}^{IPWN}$ and $\hat{\mu}_{112}^{IPWN}$. To derive the covariance between strategy means $\hat{\mu}_{111}^{IPWN}$ and $\hat{\mu}_{112}^{IPWN}$, we note that similar to $\sqrt{n}(\hat{\mu}_{111}^{IPWN}-\mu_{111})$, $\sqrt{n}(\hat{\mu}_{112}^{IPWN}-\mu_{112})$ is distributionally equivalent to $n^{-1/2}\kappa_1\sum_{i=1}^n W_{112i}(Y_i-\mu_{112})$. Therefore, the asymptotic covariance of $\sqrt{n}(\hat{\mu}_{111}^{IPWN}-\mu_{111})$ and $\sqrt{n}(\hat{\mu}_{112}^{IPWN}-\mu_{112})$ is given by,

$$
\begin{aligned}
\sigma_{111,112} = cov(\psi_{111i},\psi_{112i}) &= cov(\kappa_1 W_{111i}(Y_i-\mu_{111}),\kappa_1 W_{112i}(Y_i-\mu_{112})) \\
&= E\left[\kappa_1^2 W_{111i}W_{112i}(Y_i-\mu_{111})(Y_i-\mu_{112})\right].
\end{aligned}
$$

Since $W_{111i}W_{112i} = \frac{R_i Z_{1i} X_{1i}}{P_1^2}$, we can further simplify the above as,

$$\begin{aligned}
\sigma_{111,112} &= E\left[\kappa_1^2 \frac{R_i Z_{1i}}{P_1^2} X_{1i}(Y_i - \mu_{111})(Y_i - \mu_{112})\right] \\
\text{By the Consistency Assumption,} & \\
&= E\left\{E\left[\kappa_1^2 \frac{R_i Z_{1i}}{P_1^2} X_{1i}(Y_i(A_1B_1) - \mu_{111})(Y_i(A_1B_1) - \mu_{112})|R_i, X_{1i}, Y_i(A_1B_1)\right]\right\} \\
&= \kappa_1^2 E\left[X_{1i} E\left[\frac{R_i}{P_1^2}(Y_i(A_1B_1) - \mu_{111})(Y_i(A_1B_1) - \mu_{112})\right] E\{Z_{1i}|R_i, X_{1i}, Y_i(A_1B_1)\}\right] \\
&= \kappa_1^2 E E\left[X_{1i}\frac{R_i}{P_1}(Y_i(A_1B_1) - \mu_{111})(Y_i(A_1B_1) - \mu_{112})|X_{1i}, Y_i(A_1B_1)\right] \\
&= \kappa_1^2 \frac{\pi_1}{P_1} E\left[X_{1i}(Y_i(A_1B_1) - \mu_{111})(Y_i(A_1B_1) - \mu_{112})\right] \\
&= \kappa_1 \frac{\pi_1}{P_1}[\sigma^2_{A_1B_1} + \mu^2_{A_1B_1} - \mu_{111}\mu_{A_1B_1} - \mu_{112}\mu_{A_1B_1} + \mu_{111}\mu_{112}] \\
&= \kappa_1 \frac{\pi_1}{P_1}[\sigma^2_{A_1B_1} + (\mu_{A_1B_1} - \mu_{111})(\mu_{A_1B_1} - \mu_{112})].
\end{aligned}$$

Since, from Equation (2.2), $(\mu_{A_1B_1} - \mu_{111}) = \mu_{A_1B_1} - \pi_1\mu_{A_1B_1} - (1-\pi_1)\mu_{A_1C_1} = (1-\pi_1)(\mu_{A_1B_1} - \mu_{A_1C_1})$ and $(\mu_{A_1B_1} - \mu_{112}) = (1-\pi_1)(\mu_{A_1B_1} - \mu_{A_1C_2})$, it follows that the asymptotic covariance of $\hat{\mu}_{111}$ and $\hat{\mu}_{112}$ is given by

$$cov(\hat{\mu}^{IPWN}_{111}, \hat{\mu}^{IPWN}_{112}) = \frac{\kappa_1}{n}\frac{\pi_1}{P_1}[\sigma^2_{A_1B_1} + (1-\pi_1)^2(\mu_{A_1B_1} - \mu_{A_1C_1})(\mu_{A_1B_1} - \mu_{A_1C_2})]. \qquad (2.9)$$

A similar derivation could be employed to compute other covariances. Let $\Sigma = var(\psi_i)$, where ψ_i is the vector of eight influence functions $\psi_{jkli}, j, k, l = 1, 2$, denote the variance-covariance matrix where Equation (2.7) is used to form the diagonal elements and Equation (2.9) is used to form the off-diagonal entries, respectively.

2.4 OVERALL SAMPLE SIZE

The hypothesis of interest is whether there is a strategy-specific mean difference. The null hypothesis is $H_0 : \mu_{111}=\mu_{112}=\mu_{121}=\mu_{122}=\mu_{211}=\mu_{212}=\mu_{221}=\mu_{222}$, which is written as a linear equation $H_0 : C\mu=0$, where

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix},$$

and $\mu = [\mu_{111}, \mu_{112}, \mu_{121}, \mu_{122}, \mu_{211}, \mu_{212}, \mu_{221}, \mu_{222}]^T$. Under the null hypothesis, the statistic $n\hat{\mu}^T C^T [C\hat{\Sigma}C^T]^{-1} C\hat{\mu}$ follows a central chi-square distribution with degree of freedom equal to 7, the number of rows of the contrast matrix C. Here $\hat{\mu}$ and $\hat{\Sigma}$ denote estimated mean vector and covariance matrix given by,

$$\hat{\mu}^T = [\hat{\mu}_{111}, \hat{\mu}_{112}, \hat{\mu}_{121}, \hat{\mu}_{122}, \hat{\mu}_{211}, \hat{\mu}_{212}, \hat{\mu}_{221}, \hat{\mu}_{222}],$$

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_1 & \tilde{0} \\ \tilde{0} & \hat{\Sigma}_2 \end{bmatrix},$$

where

$$\hat{\Sigma}_1 = \begin{bmatrix} \hat{\sigma}^2_{111} & \hat{\sigma}_{111,112} & \hat{\sigma}_{111,121} & \hat{\sigma}_{111,122} \\ \hat{\sigma}_{112,111} & \hat{\sigma}^2_{112} & \hat{\sigma}_{112,121} & \hat{\sigma}_{112,122} \\ \hat{\sigma}_{111,112} & \hat{\sigma}_{121,112} & \hat{\sigma}^2_{121} & \hat{\sigma}_{121,122} \\ \hat{\sigma}_{122,111} & \hat{\sigma}_{122,112} & \hat{\sigma}_{122,121} & \hat{\sigma}^2_{122} \end{bmatrix},$$

$$\hat{\Sigma}_2 = \begin{bmatrix} \hat{\sigma}^2_{211} & \hat{\sigma}_{211,212} & \hat{\sigma}_{211,221} & \hat{\sigma}_{211,222} \\ \hat{\sigma}_{212,211} & \hat{\sigma}^2_{212} & \hat{\sigma}_{212,221} & \hat{\sigma}_{212,222} \\ \hat{\sigma}_{211,212} & \hat{\sigma}_{221,212} & \hat{\sigma}^2_{221} & \hat{\sigma}_{221,222} \\ \hat{\sigma}_{222,211} & \hat{\sigma}_{222,212} & \hat{\sigma}_{222,221} & \hat{\sigma}^2_{222} \end{bmatrix},$$

$$\tilde{0} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where $\hat{\mu}_{jkl}$ is defined in Equation (2.7), $\hat{\sigma}^2_{jkl}$ is obtained by substituting estimates of parameters on the RHS in Equation (2.7). For example,

$$\hat{\sigma}^2_{111} = \frac{\hat{\kappa}_1}{n}\left[\frac{\hat{\pi}_1}{\hat{P}_1}\left\{\hat{\sigma}^2_{A_1B_1} + (\hat{\mu}_{111} - \hat{\mu}_{A_1B_1})^2\right\} + \frac{(1-\hat{\pi}_1)}{\hat{Q}_1}\left\{\hat{\sigma}^2_{A_1C_1} + (\hat{\mu}_{111} - \hat{\mu}_{A_1C_1})^2\right\}\right],$$

where

$$\hat{\kappa}_1 = \frac{n}{\sum_{i=1}^n X_{1i}},$$

$$\hat{\pi}_1 = \frac{\sum_{i=1}^n X_{1i}R_i}{\sum_{i=1}^n X_{1i}},$$

$$\hat{\mu}_{A_1B_1} = \frac{\sum_{i=1}^n X_{1i}R_iZ_{1i}Y_i}{\sum_{i=1}^n X_{1i}R_iZ_{1i}},$$

$$\hat{\sigma}^2_{A_1B_1} = \frac{\sum_{i=1}^n (X_{1i}R_iZ_{1i}Y_i - \hat{\mu}_{A_1B_1})^2}{(\sum_{i=1}^n X_{1i}R_iZ_{1i})(\sum_{i=1}^n X_{1i}R_iZ_{1i} - 1)},$$

$$\hat{Q}_1 = \frac{\sum_{i=1}^n X_{1i}(1-R_i)Z'_{1i}}{\sum_{i=1}^n X_{1i}(1-R_i)},$$

and

$$\hat{P}_1 = \frac{\sum_{i=1}^n X_{1i}R_iZ_{1i}}{\sum_{i=1}^n X_{1i}R_i}.$$

Under the alternative hypothesis, the test statistic follows a non-central chi-squared distribution with the same degrees of freedom and a non-centrality parameter λ, where

$$\lambda = n\mu^T C^T [C\Sigma C^T]^{-1} C\mu.$$

Consequently, a straightforward manipulation leads to a sample size formula,

$$n = \frac{\lambda}{\mu^T C^T [C\Sigma C^T]^{-1} C\mu}. \tag{2.10}$$

To use the sample size formula in Equation (2.10), for a given power, we note that the power of the Wald test is the probability that we reject the null hypothesis, i.e., the probability that the test statistic is greater than the critical value. Thus,

$$power = P(\chi^2_{df=7,\lambda} \geq \chi^2_{df=7,1-\alpha}) = 1 - P(\chi^2_{df=7,\lambda} \leq \chi^2_{df=7,\alpha}), \tag{2.11}$$

where α is the level of significance of the test. For a given power and α, we can solve Equation (2.11) for λ. Having obtained λ, the sample size needed for achieving a given power is obtained by plugging in appropriate strategy means under the alternative hypothesis and their assumed variance-covariance matrix into the sample size expression above.

The knowledge of subgroup means and variances in the population will allow the computation of covariance terms.

2.5 POWERING PAIRWISE COMPARISONS

Above we developed a sample size formula for a global test that provides evidence that there are differences among at least one pair of strategy means. Next, it is natural to focus on pairwise comparisons and ask which strategy means are different. A popular two-sample pairwise test is the t-test. A sample size based on the usual t-test would not apply directly since the assumption of independence among strategy means does not hold. When strategies share first stage treatment, a pairwise treatment comparison should consider the between-strategy covariances in the traditional t-test based sample size formula. Suppose we are interested in the sample size of a test that truly rejects the null hypotheses at a pre-specified level of significance (α) and a given power. For instance, there are 8 regimes and 28 pairwise comparisons for Design 1. One possible pairwise comparison would be,

$$H_0 : \mu_{111} - \mu_{112} = \delta_1.$$

For each test different sample sizes are required to detect a difference between each pairwise comparison. To control type I error, Bonferroni correction can be used. That is, for a two-sided test the level of significance for each hypothesis will be α/g, where g is the total number

of pairwise comparisons. The aim is to compute the sample sizes for each pairwise comparison and then select maximum of the set of sample sizes that powers a test to identify difference between strategy means. The sample size formula that accounts dependency among strategy means is,

$$n = \frac{[\sigma^2_{jkl} + \sigma^2_{j'k'l'} - 2\sigma_{jkl,j'k'l'}][Z_{1-\alpha/2g} + Z_{1-\beta}]^2}{[\mu_{jkl} - \mu_{j'k'l'}]^2}, j, k, l = 1, 2 \tag{2.12}$$

where σ^2_{jkl}, $\sigma^2_{j'k'l'}$, and $\sigma_{jkl,j'k'l'}$ are obtained using Equations (2.7) and (2.9); μ_{jkl} and $\mu_{j'k'l'}$ are the strategy means under the alternative hypothesis. If the strategy means do not share the same initial treatments, the between-strategy means covariance is zero and the sample size formula (2.12) would mimic the one required for independent two-sample t-test.

Equation (2.12) has a more general use than it apparently implies. For example, suppose prior to designing the trial, researchers focus on $g_1 \leq g$ specific pairwise comparisons. Then the sample size for pairwise comparisons can be calculated using a level of significance $\frac{\alpha}{g_1}$ to ensure that the experiment-wise error is maintained at α for a pairwise comparison of g_1 pairs. Since the variance-covariance formula depends on the randomization probabilities, the researcher could potentially use randomization probabilities that allocate more observations to the strategies of interest. The other $(g-g_1)$ pairwise comparisons could remain unpowered but essentially provide valuable information for future studies.

Methods developed in Sections 3 and 4 can easily be applied to Designs 2 and 3. For example, for Design 2, to test the overall hypothesis $H_0 : \mu_{111} = \mu_{112} = \mu_{221} = \mu_{222}$, it is written as a linear equation $\mathrm{H}_0 : C\mu{=}0$, where

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix},$$

and $\mu = [\mu_{111}, \mu_{112}, \mu_{221}, \mu_{222}]^T$. The variance-covariance matrix Σ can be computed in a similar manner as Design 1 by replacing with 4 diagonal estimated mean strategy variances and 4 off-diagonal estimated covariances among strategy means. That is,

$$\Sigma = \begin{bmatrix} \sigma^2_{111} & \sigma_{111,112} & 0 & 0 \\ \sigma_{112,111} & \sigma^2_{112} & 0 & 0 \\ 0 & 0 & \sigma^2_{211} & \sigma_{211,212} \\ 0 & 0 & \sigma_{212,211} & \sigma^2_{212} \end{bmatrix}.$$

Note that in Design 2, since there is only one treatment option for the responders, $P_1{=}1$.

2.6 SIMULATION STUDY AND RESULTS

To evaluate the performance of the overall sample size formula, we conducted a number of simulations to see if the empirical power for detecting the alternative hypothesis is close to the nominal power. We presented four scenarios for each of the three designs in Tables 1, 2 and 3 by varying the nominal power, response rates and probabilities of second treatment assignment for responders (P_k) and non-responders (Q_l). For each subject in the population, $Y_i(A_jB_k)$ and $Y_i(A_jC_l)$ follow normal distribution with means $\mu_{A_jB_k}$ and $\mu_{A_jC_l}$, and variances $\sigma^2_{A_jB_k}$ and $\sigma^2_{A_jC_l}$, respectively for $j,k,l = 1,2$. The response status R_i was generated from a Bernoulli distribution with probability (response rate) π_1 to treatment A_1 and π_2 to treatment A_2. For each scenario we generated 10000 samples using the three designs.

The methods described so far are explained via Design 1, however, the formulas can be applied to Designs 2 and 3. There are six pairwise comparisons for Design 2 and fifteen for Design 3. Tables 4 and 5 show the pairwise sample size computation for Designs 2 and 3. Tables 1, 2 and 3 demonstrate sample size computation for different scenarios by assuming certain values for population parameters. Design 1 assumes subgroup means: $\mu_{A_jB_1} = \mu_{A_jC_2} = 15$, $\mu_{A_jC_1} = 20$, $\mu_{A_jB_2} = 22$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$, for $j,k,l = 1,2$. Subgroup variances are assumed to be the same for all designs considered. Depending on a specific design and scenario considered, the following range of response proportions π_j's are assumed: 0.2, 0.3, 0.5, 0.6 and 0.7. Similarly, depending on a specific design the following P_1 and Q_1 are assumed. Probability of treatment assignment for responders, P_1, is assumed to be 0.5, 0.7, 0.9 and 1. For non-responders, $Q_1(= 1 - Q_2)$, is assumed to be 0.5, 0.7, 0.9. Design 2 assumes the following subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_jC_2} = 15$, $\mu_{A_1C_1} = 20$, $\mu_{A_2C_1} = 22$, for $j,k,l = 1,2$. Design 3 assumes the following subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_3B_1} = 19$, $\mu_{A_jC_2} = 15$, $\mu_{A_1C_1} = 20$, $\mu_{A_2C_1} = 22$, $\mu_{A_3C_2} = 24$, for $j,k,l = 1,2,3$. The parameter values were chosen following those from Ko and Wahed [11]. The sample size changes as the values for π_j, P_l, Q_l, and power vary. The strategy means differ for each scenario in each table. In each scenario, having obtained the appropriate sample size using our formula, we evaluate the power of the Wald tests in rejecting the null hypothesis of no difference in treatment means when the strategies have different means. The

effect size is computed using the Mahalanobis distance (MD). One useful property of the MD is that it takes into account the correlation in the data. Effect sizes are common measures in psychology and other disciplines where they are useful in calculating and interpreting power especially if more than two groups are involved. Effect size estimates add validity to interpretation of results. The magnitude of effect sizes would capture experimental effects by protecting guaranteed significance due to large sample size [5].

The first row of Scenario 1 in Table 1 assumes strategy means $\mu_{111} = 17.5$ $\mu_{112} = 15$, $\mu_{121} = 21$, $\mu_{122} = 18.5$, $\mu_{211} = 17.5$, $\mu_{212} = 15$, $\mu_{221} = 21$, $\mu_{222} = 18.5$ when response rates π_1, π_2 were taken to be both 0.5; P_1 and Q_1 are assumed to be 0.5. 70 subjects would be required to detect the resulting effect size of 0.21 with power 80% at $\alpha = 0.05$. The empirical power is 85% which is slightly inflated compared to the nominal power of 80% used to compute the sample size. Row 3 of the same scenario shows that the empirical power of 92% is close to the nominal value of 90%. Similar patterns follow for all the rows in Scenarios 2, 3 and 4. If we observe across all scenarios (from 4 to 1), we note a small degree of increase in empirical power when P_1 increases.

The first row of Scenario 1 in Design 2 (Table 2) assumes strategy means $\mu_{111} = 17.5$, $\mu_{112} = 15$, $\mu_{211} = 19.5$, $\mu_{212} = 16$ when response rates π_1, π_2 were taken to be both 0.5; P_1=1 and Q_1=0.5. 142 subjects would be required to detect the resulting effect size of 0.08 with power 80% at $\alpha = 0.05$. The empirical power is 81% which is very close to the nominal power of 80% used to compute the sample size. Row 4 of scenario 3 shows that the empirical power of 93% is slightly inflated compared to the nominal value of 90%. For various response rates, the empirical power for each case in scenarios 1 to 3 nearly attain the nominal power. This indicates that the sample sizes calculated for Design 2 ensure enough power to detect differences among the four strategy means.

The first row of Scenario 1 in Design 3 (Table 3) assumes strategy means $\mu_{111} = 17.5$, $\mu_{112} = 15$, $\mu_{211} = 19.5$, $\mu_{212} = 16$, $\mu_{311} = 21.5$, $\mu_{312} = 17$ when response rates π_1, π_2, π_3 were taken to be all 0.5. 108 subjects would be required to detect the resulting effect size of 0.12 with power 80% at $\alpha = 0.05$. The empirical power is 83% which is slightly larger than the nominal power of 80% used to compute the sample size. We note that for small changes in response rates, sometimes the sample sizes do not change or change only slightly.

Table 1: **Overall sample size for Design 1.**

Scenario	π_1	π_2	P_1	Power	Overall Sample Size	Empirical Power	Effect Size (Mahalanobis Distance)
1	0.5	0.5	0.5	0.8	70	0.84	0.21
	0.5	0.5	0.7	0.8	79	0.85	0.18
	0.5	0.5	0.5	0.9	89	0.92	0.21
	0.5	0.5	0.8	0.9	120	0.92	0.15
2	0.2	0.5	0.5	0.8	83	0.82	0.17
	0.2	0.5	0.7	0.8	92	0.83	0.16
	0.2	0.5	0.5	0.9	106	0.9	0.17
	0.2	0.5	0.8	0.9	134	0.92	0.14
3	0.7	0.5	0.5	0.8	62	0.85	0.23
	0.7	0.5	0.7	0.8	71	0.85	0.20
	0.7	0.5	0.5	0.9	79	0.92	0.23
	0.7	0.5	0.7	0.9	91	0.92	0.20
4	0.2	0.7	0.5	0.8	72	0.84	0.20
	0.2	0.7	0.7	0.8	82	0.84	0.18
	0.2	0.7	0.5	0.9	92	0.91	0.20
	0.2	0.7	0.7	0.9	104	0.92	0.18

Sample size computation and simulation of empirical power (# replications=10000) for Design 1 where $Q_1 = 0.5$, subgroup means: $\mu_{A_jB_1} = \mu_{A_jC_2} = 15$, $\mu_{A_jC_1} = 20$, $\mu_{A_jB_2} = 22$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$, for $j, k, l = 1, 2$. Hypothesis of interest is H_0 : μ_{111}=μ_{112}=μ_{121}=μ_{122}=μ_{211}=μ_{212}=μ_{221}=μ_{222}; $\alpha = 0.05$.

Alternative is true with means: Scenario 1: $\mu_{111} = 17.5$, $\mu_{112} = 15.0$, $\mu_{121} = 21.0$, $\mu_{122} = 18.5$, $\mu_{211} = 17.5$, $\mu_{212} = 15.0$, $\mu_{221} = 21.0$, $\mu_{222} = 18.5$

Scenario 2: $\mu_{111} = 19.0$, $\mu_{112} = 15.0$, $\mu_{121} = 20.4$, $\mu_{122} = 16.4$, $\mu_{211} = 17.5$, $\mu_{212} = 15.0$, $\mu_{221} = 21.0$, $\mu_{222} = 18.5$

Scenario 3: $\mu_{111} = 16.5$, $\mu_{112} = 15.0$, $\mu_{121} = 21.4$, $\mu_{122} = 19.9$, $\mu_{211} = 17.5$, $\mu_{212} = 15.0$, $\mu_{221} = 21.0$, $\mu_{222} = 18.5$

Scenario 4: $\mu_{111} = 19.0$, $\mu_{112} = 15.0$, $\mu_{121} = 20.4$, $\mu_{122} = 16.4$, $\mu_{211} = 16.5$, $\mu_{212} = 15.0$, $\mu_{221} = 21.4$, $\mu_{222} = 19.9$

Table 2: **Overall sample size for Design 2.**

Scenario	π_1	π_2	Q_1	Power	Overall Sample Size	Empirical Power	Effect Size (Mahalanobis Distance)
1	0.5	0.5	0.5	0.8	142	0.81	0.08
	0.5	0.5	0.7	0.8	156	0.82	0.07
	0.5	0.5	0.5	0.9	185	0.91	0.08
	0.5	0.5	0.9	0.9	344	0.92	0.04
2	0.2	0.5	0.5	0.8	130	0.81	0.08
	0.2	0.5	0.7	0.8	144	0.82	0.07
	0.2	0.5	0.5	0.9	169	0.91	0.08
	0.2	0.5	0.9	0.9	448	0.92	0.03
3	0.7	0.5	0.5	0.8	143	0.82	0.08
	0.7	0.5	0.7	0.8	143	0.83	0.08
	0.7	0.5	0.5	0.9	186	0.91	0.08
	0.7	0.5	0.9	0.9	241	0.93	0.06
4	0.7	0.2	0.5	0.8	94	0.82	0.12
	0.7	0.2	0.7	0.8	88	0.84	0.12
	0.7	0.2	0.5	0.9	122	0.9	0.12
	0.7	0.2	0.9	0.9	131	0.94	0.11

Sample size computation and simulation of empirical power (# replications=10000) for Design 2 where $P_1 = 1$, subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_jC_2} = 15$, $\mu_{A_1C_1} = 20$, $\mu_{A_2C_1} = 22$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$ for $j, k, l = 1, 2$. Hypothesis of interest is $\mathrm{H}_0 : \mu_{111}{=}\mu_{112}{=}\mu_{211}{=}\mu_{212}$.

Alternative is true with means: Scenario 1: $\mu_{111} = 17.5$, $\mu_{112} = 15.0$, $\mu_{221} = 19.5$, $\mu_{222} = 16.0$

Scenario 2: $\mu_{111} = 19.0$, $\mu_{112} = 15.0$, $\mu_{221} = 19.5$, $\mu_{222} = 16.0$

Scenario 3: $\mu_{111} = 16.5$, $\mu_{112} = 15.0$, $\mu_{221} = 19.5$, $\mu_{222} = 16.0$

Scenario 4: $\mu_{111} = 16.5$, $\mu_{112} = 15.0$, $\mu_{221} = 21.0$, $\mu_{222} = 15.4$

Table 3: **Overall sample size for Design 3.**

Scenario	π_1	π_2	Power	Overall Sample Size	Empirical Power	Effect Size (Mahalanobis Distance)
1	0.5	0.5	0.8	108	0.83	0.12
	0.2	0.5	0.8	111	0.83	0.12
	0.5	0.5	0.9	139	0.91	0.12
	0.2	0.5	0.9	142	0.91	0.12
2	0.2	0.2	0.8	95	0.84	0.14
	0.2	0.6	0.8	116	0.84	0.11
	0.2	0.2	0.9	122	0.92	0.14
	0.2	0.6	0.9	149	0.91	0.11
3	0.3	0.5	0.8	111	0.83	0.12
	0.3	0.6	0.8	116	0.82	0.11
	0.3	0.5	0.9	142	0.91	0.12
	0.3	0.6	0.9	149	0.92	0.11
4	0.4	0.5	0.8	110	0.83	0.12
	0.4	0.6	0.8	115	0.82	0.11
	0.4	0.5	0.9	141	0.92	0.12
	0.4	0.6	0.9	148	0.91	0.11

Sample size computation and simulation of empirical power (# replications=10000) for Design 3 where $P_1 = 1$, subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_3B_1} = 19$, $\mu_{A_jC_2} = 15$, $\mu_{A_1C_1} = 20$, $\mu_{A_2C_1} = 22$, $\mu_{A_3C_1} = 24$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$ for $j, k, l = 1, 2, 3$. Response rate for induction treatment A_3 is assumed to be 50%. Hypothesis of interest is $\mathrm{H}_0 : \mu_{111} = \mu_{113} = \mu_{221} = \mu_{223} = \mu_{331} = \mu_{332}$.

Alternative is true with means: Scenario 1: $\mu_{111} = 17.5$, $\mu_{112} = 15.0$, $\mu_{221} = 19.5$, $\mu_{222} = 16.0$, $\mu_{331} = 21.5$, $\mu_{332} = 17.0$

Scenario 2: $\mu_{111} = 19.0$, $\mu_{112} = 15.0$, $\mu_{221} = 21.0$, $\mu_{222} = 15.4$, $\mu_{331} = 21.5$, $\mu_{332} = 17.0$

Scenario 3: $\mu_{111} = 18.5$, $\mu_{112} = 15.0$, $\mu_{221} = 19.5$, $\mu_{222} = 16.0$, $\mu_{331} = 21.5$, $\mu_{332} = 17.0$

Scenario 4: $\mu_{111} = 18.0$, $\mu_{112} = 15.0$, $\mu_{221} = 19.5$, $\mu_{222} = 16.0$, $\mu_{331} = 21.5$, $\mu_{332} = 17.0$

Table 4: **Pairwise sample size computation for Design 2.**

Hypothesis	Overall MC Adjusted Sample Size	Not Corrected for Multiple Comparison	δ_i
$H_1 : \mu_{111}-\mu_{112}=\delta_1$	532	345	2.5
$H_2 : \mu_{111}-\mu_{221}=\delta_2$	1107	717	-2.0
$H_3 : \mu_{111}-\mu_{222}=\delta_3$	1882	1220	1.5
$H_4 : \mu_{112}-\mu_{221}=\delta_4$	207	134	-4.5
$H_5 : \mu_{112}-\mu_{222}=\delta_5$	4008	2598	-1.0
$H_6 : \mu_{221}-\mu_{222}=\delta_6$	280	181	3.5

Subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_1C_1} = 20$, $\mu_{A_1C_2} = 15$, $\mu_{A_2C_1} = 22$, $\mu_{A_2C_2} = 15$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$ for $j, k, l = 1, 2$. Here π_1=0.5, π_2=0.5, P_1=1, Q_1=0.5, and power=0.8.

For example, row 4 of scenarios 2 and 3 have the same sample size (149). The sample size did not change as π_1 changed slightly from 0.2 to 0.3.

In many clinical trials, testing of an overall hypothesis may not be of primary interest, rather some or all of the pairwise comparisons are. To show how the sample size for a SMAR trial is determined in such cases, we present the sample sizes required for Design 2 when all six pairwise comparisons are powered simultaneously in the second column of Table 4. The third column provides the sample sizes when individual tests are powered. For example, under the setting described in Table 4, Design 2 requires 4008 patients to power all pairwise comparisons. However, if the interest, for example, is in powering the single hypothesis $H_0 : \mu_{111} = \mu_{112}$ leaving other pairs as exploratory, the trial could be conducted using a sample as small as 345. Similarly, Table 5 provides sample sizes for Design 3 when fifteen pairwise comparisons are powered simultaneously (Column 2) and when only three pairwise comparisons are considered (Column 3). From Column 2, Design 3 requires 30,704 patients (maximum of the sample sizes) to power all pairwise comparisons. However, if the interest is in powering only three pairwise hypotheses such as $H_0 : \mu_{112} = \mu_{113}$, $H_0 : \mu_{112} = \mu_{221}$, and $H_0 : \mu_{112} = \mu_{223}$, the trial would require a sample size of 2,441. On the other hand, if the

Table 5: **Pairwise sample size computation for Design 3.**

Hypothesis	Overall MC Adjusted Sample Size	Partially MC Adjusted Sample Size	δ_i
$H_1 : \mu_{112}-\mu_{113}=\delta_1$	941	690	2.5
$H_2 : \mu_{112}-\mu_{221}=\delta_2$	1955	1435	-2.0
$H_3 : \mu_{112}-\mu_{223}=\delta_3$	3326	2441	1.5
$H_4 : \mu_{112}-\mu_{331}=\delta_4$	489	359	-4.0
$H_5 : \mu_{112}-\mu_{332}=\delta_5$	30704	22535	0.5
$H_6 : \mu_{113}-\mu_{221}=\delta_6$	366	269	-4.5
$H_7 : \mu_{113}-\mu_{223}=\delta_7$	7082	5198	-1.0
$H_8 : \mu_{113}-\mu_{331}=\delta_8$	176	129	-6.5
$H_9 : \mu_{113}-\mu_{332}=\delta_9$	1819	1335	-2.0
$H_{10} : \mu_{221}-\mu_{223}=\delta_{10}$	494	362	3.5
$H_{11} : \mu_{221}-\mu_{331}=\delta_{11}$	1955	1435	-2.0
$H_{12} : \mu_{221}-\mu_{332}=\delta_{12}$	1228	901	2.5
$H_{13} : \mu_{223}-\mu_{331}=\delta_{13}$	247	182	-5.5
$H_{14} : \mu_{223}-\mu_{332}=\delta_{14}$	7339	5386	-1.0
$H_{15} : \mu_{331}-\mu_{332}=\delta_{15}$	314	230	4.5

Subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_3B_1} = 19$, $\mu_{A_1C_1} = 20$, $\mu_{A_1C_2} = 15$, $\mu_{A_2C_1} = 22$, $\mu_{A_2C_2} = 15$, $\mu_{A_3C_1} = 24$, $\mu_{A_3C_2} = 15$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$ for $j, k, l = 1, 2, 3$. Here π_1=0.5, π_2=0.5, P_1=1, Q_1=0.5, and power=0.8. Second column provides sample size which powers all pairwise comparisons whereas the third column assumes that only three of the fifteen hypotheses are of interest.

interest is only in comparing the three pairs, H_4, H_6, and H_8 then the sample size required will be $n = 359$.

Outcomes in the above simulation scenarios were generated from a normal distribution. We wanted to conduct the sensitivity of our formula to non-normal responses. To do this, we further generated data from logistic (symmetric) and gamma (skewed) distributions and calculated the empirical power based on the sample size calculated using Equation(2.10). Basically, we selected one row from each scenario of Tables 1 to 3 to perform sensitivity analysis of our formula using data from the logistic and gamma distributions ensuring the same means and variances for the subpopulations and keeping all other parameters the same. From each table, we selected the first row for Scenarios 1 and 3 while we chose the fourth row for Scenarios 2 and 4. Therefore, the results presented in Table 6 have twelve rows in total. In general, the nominal power is maintained and is consistent across the three distributions. This shows that our sample size formula is robust to the misspecification of outcome distribution.

2.7 DISCUSSION

Complex multi-stage diseases require decision-based multi-stage treatments depending on the response to prior-stage treatments. SMAR designs provide efficient and unbiased inference to compare staged strategies for complex conditions. We presented a sample size formula that is applicable for various SMAR designs to ensure adequately powered comparisons of these treatment strategies. The usual design is to randomize responders (or non-responders) to available treatments. A slight variation to that is a design where responders (or non-responders) would not be randomized any further in the second stage. Designs 2 and 3 are such examples. In Design 2, only the non-responders are randomized to C_1 or C_2 and C_1' or C_2' respectively depending on whether they received A_1 or A_2 in the first stage. Responders would stay on the same first stage treatment. Equivalently, responders will be randomized with probability 1 to whatever treatment they received in the first stage. There are four strategies resulting from this design and the sample size required to detect differences among

Table 6: **Robustness of the Sample Size Formula against misspecification of outcome distributions.**

Design	Scenario	π_1	π_2	P_1	Q_1	Power	Overall Sample Size	Empirical Power: Normal	Empirical Power: Gamma	Empirical Power: Logistic
Design 1	1	0.5	0.5	0.5	0.5	0.8	70	0.84	0.86	0.86
	2	0.2	0.5	0.8	0.5	0.9	134	0.92	0.93	0.93
	3	0.7	0.5	0.5	0.5	0.8	62	0.85	0.86	0.87
	4	0.2	0.7	0.7	0.5	0.9	104	0.92	0.92	0.92
Design 2	1	0.5	0.5	1	0.5	0.8	142	0.81	0.83	0.82
	2	0.2	0.5	1	0.9	0.9	448	0.92	0.91	0.91
	3	0.7	0.5	1	0.5	0.8	143	0.82	0.85	0.83
	4	0.7	0.2	1	0.9	0.9	131	0.94	0.96	0.94
Design 3	1	0.5	0.5	1	0.5	0.8	108	0.83	0.86	0.85
	2	0.2	0.6	1	0.5	0.9	149	0.91	0.93	0.92
	3	0.3	0.5	1	0.5	0.8	111	0.83	0.86	0.85
	4	0.4	0.6	1	0.5	0.9	148	0.91	0.93	0.92

For Design 1 the following parameter values were considered: $Q_1 = 0.5$, subgroup means: $\mu_{A_jB_1} = \mu_{A_jC_2} = 15$, $\mu_{A_jC_1} = 20$, $\mu_{A_jB_2} = 22$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$, for $j, k, l = 1, 2$. The hypothesis tested is $\mathrm{H}_0 : \mu_{111}{=}\mu_{112}{=}\mu_{121}{=}\mu_{122}{=}\mu_{211}{=}\mu_{212}{=}\mu_{221}{=}\mu_{222}$; $\alpha = 0.05$. For Design 2 the following parameter values were considered: $P_1 = 1$, subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_jC_2} = 15$, $\mu_{A_1C_1} = 20$, $\mu_{A_2C_1} = 22$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$ for $j, k, l = 1, 2$. The hypothesis tested is $\mathrm{H}_0 : \mu_{111}{=}\mu_{112}{=}\mu_{211}{=}\mu_{212}$. For Design 3 the following parameter values were considered: $P_1 = 1$, subgroup means: $\mu_{A_1B_1} = 15$, $\mu_{A_2B_1} = 17$, $\mu_{A_3B_1} = 19$, $\mu_{A_jC_2} = 15$, $\mu_{A_1C_1} = 20$, $\mu_{A_2C_1} = 22$, $\mu_{A_3C_2} = 24$; subgroup variances: $\sigma^2_{A_jB_k} = 6^2$, $\sigma^2_{A_jC_l} = 8^2$ for $j, k, l = 1, 2, 3$. Response rates to induction treatment A_3 is assumed to be 50%. The hypothesis tested is $\mathrm{H}_0 : \mu_{111}{=}\mu_{113}{=}\mu_{221}{=}\mu_{223}{=}\mu_{331}{=}\mu_{332}$.

the four strategies is computed. In Design 3 each patient is randomized to a set of treatments (A_1, A_2, A_3) in the first stage and these treatments are continued until they fail due to disease worsening. The patient is then re-randomized among a set of the same first stage treatments with the exception of the treatment s/he received initially. There are six strategies of interest in this design. We showed in the simulation how to compute sample size formula for this design and showed that the formula ensures nominal power under various scenarios involving many outcome distributions.

In contrast to our formula, Murphy's [22] formula is not applicable to designs powering multi-strategy comparison or to designs comparing strategies that share the same initial treatments commonly referred to as shared-path strategies [10] or overlapping strategies [3]. Moreover, their formula requires specifying the variance of the response under the strategies being compared, although the effect sizes can be specified per standard deviations of mean difference assuming equal variance across strategies.

Dawson and Lavori [3, 4] provide a sample size formula for comparing pairs of overlapping or non-overlapping/treatment strategies based on semiparametric efficient variances. The formula requires one to specify the variance of the response under each strategy and the variance inflation factor, the latter depending on the coefficient of determinations based on the regression of counterfactual strategy response on stage-specific states. Correct specification of such quantities is difficult, if not impossible, in the absence of a similar SMAR trial. However, when correctly specified, Dawson and Lavori's formula provide smaller sample sizes than those proposed in Murphy [22] or the ones provided here. One advantage of both Murphy [22] and Dawson and Lavori's [3, 4] formula over our method is that they can be applied to compare strategies from SMAR trial with more than two stages. However, like Murphy's formula, Dawson and Lavori's formula also focuses on comparing pairs of treatment strategies.

The simplicity of our procedure compared to Dawson and Lavori [3] (even in the two-stage SMAR trial settings) relies on the specification of the parameters. Our formula requires one to specify sub-group-specific means and variances. Our sample size formula requires specification of subgroup means and variances for patients following different treatment paths. These parameters are usually available from observational studies or stage-specific individual

non-SMAR trials. For example, there are many cancer clinical trials that compare frontline treatments (e.g. Estey et al. [15]). Even though such trials are terminated once the recruitment is over and the primary endpoint is observed or the trial period ends, patients are often followed and medication information (salvage treatments used) is collected for patients who become resistant to frontline therapy or for patients with disease progression. The collection of salvage treatment information is often done only for the purpose of safety, however, such information allows the researchers to obtain meaningful information on subgroup means and variances based on the salvage therapies received within each frontline treatments. Mental health and other clinical research by their very nature, are concerned with sequences of treatments and hence the means and variances of responses under a particular treatment sequence are most likely to be available from observational studies or from electronic medical records. Fortunately, there are already existing SMAR trials in mental health (STAR*D [33], CATIE [35]) that can provide useful information on subgroups to be used in future trial design.

The Murphy [22] and Dawson and Lavori [3] methods require fewer unknown quantities to be specified compared to what is required by our formula; our parameters are basically means and variances of response among subpopulations. Generally, these parameters can be obtained from pilot studies, non-SMAR trials or observational studies. Therefore, these parameters are less likely to be mis-specified as compared to the parameters in Murphy's [22] and Dawson and Lavori's [3] methods. Moreover, our focus is to compare multiple treatment strategies for which specification of effect size does not necessarily reduce the number of unknown parameters.

The Oetting et al. [25] sample size formula for comparing two strategies is derived under the assumption that response rates are the same across the two first stage treatments. While a sensitivity analysis was carried out in the simulation, this assumption may not be reasonable in practice. Finally, our formula does not address the issue of finding an optimal treatment strategy, which is a separate issue that is dealt with in Oetting et al. [25].

Future research could investigate sample size formulas for various k-stage designs with emphasis on specific and meaningful number of strategies. Issues of missing data is another design concern in SMAR trials that needs to be addressed.

3.0 ANALYSIS OF SEQUENTIALLY RANDOMIZED TRIALS THROUGH ARTIFICIAL RANDOMIZATION

3.1 INTRODUCTION

Treatment of patients whose disease severity change over time is best managed by employing strategies that treat patients dynamically. Naturally, administering treatments in steps according to changing disease phases would lead to better management and likely better outcome. Psychiatrists commonly treat mental illnesses by first stabilizing the patient in the first stage and then preventing relapse in the second stage [33]. Treatment strategies that are adapted to patients' response status are therefore attractive to clinicians and researchers alike. Public health interest in adaptive treatment strategies (ATS) is growing basically for two reasons:

i. Increasingly convincing evidence are becoming available that using adaptive treatment strategies, overall health outcome could be improved and better compared to traditional once-and-for-all treatments, and

ii. ATSs are implemented in a manner a patient's disease is naturally managed.

Research in the area of ATS has so far focused on comparison of adaptive treatment strategies and estimation of optimal strategy. The crux of the problem in this chapter is how to compare strategy means that account for baseline covariates. Regression methods such as multiple linear, logistic, or survival (Cox or accelerated failure) regression that allow for comparison of treatment strategies flexibly adjusting for baseline covariates are not as straight-forward to apply in SMART designs. This is because a patients in a SMAR trial can belong to multiple strategies making it challenging, if not impossible to apply regression

Table 7: **A hypothetical data set from a two-stage SMAR trial.**

Patient	Initial Treatment	Response Status	Second Treatment	A_1B_1	A_1B_2	A_2B_1	A_2B_2
1	A_1	No	B_1	1	0	0	0
2	A_1	Yes	NA	1	1	0	0
3	A_1	No	B_2	0	1	0	0
4	A_2	Yes	B_1	0	0	1	0
5	A_2	Yes	NA	0	0	1	1
6	A_2	No	B_2	0	0	0	1

techniques. To clarify this, let us consider a two-stage SMAR design similar to the one described in Figure 3 on page 12. Four strategies in this design are A_jB_k which denotes "treat with A_j followed by B_k if he/she is a non-responder $(j, k = 1, 2)$". One may be tempted to compare the four strategies: A_1B_1, A_1B_2, A_2B_1, A_2B_2 using a regression model. Note that a patient responding to A_1 is, by definition, consistent with both strategies A_1B_1 and A_1B_2 which poses a challenge for data analysts as it violates basic assumptions of regression modeling of unique group membership. As shown in [37], Table 7 summarizes a hypothetical situation whereby a patient can belong to multiple strategies. In this hypothetical data set we present 6 patients participating in a trial represented in Figure 3 (page 12). It gives the initial treatment assignment, response status, second treatment assignment, and an indicator (0 or 1) to indicate if that particular patient belongs to one of the four specific strategies. For instance, patient 2 is a responder to initial treatment A_1. Therefore, his or her data could be used to draw inference for two strategies namely, A_1B_1 and A_1B_2. Similarly, patient 5 is a responder to A_2 whose treatment path is consistent with following strategies A_2B_1 and A_2B_2. Thus, the same patient can have a treatment trajectory that makes him eligible to be counted in multiple strategies poses a challenge for the analyst to use regression modeling techniques. Hence, we propose an "artificial randomization" technique to make the data appear that each subject belongs to a unique strategy. This will enable us to use regular regression methods by inserting treatment strategy indicators as covariates in a regression model along with other covariates.

Table 8: **A hypothetical data set from a two-stage SMAR trial: Illustrating Artificial Randomization.**

Patient	Initial Treatment	Response Status	Second Treatment	Z^*	A_1B_1	A_1B_2	A_2B_1	A_2B_2
1	A_1	No	B_1	B_1	1	0	0	0
2	A_1	Yes	NA	B_1	1	0	0	0
3	A_1	No	B_2	B_2	0	1	0	0
4	A_2	No	B_1	B_1	0	0	1	0
5	A_2	Yes	NA	B_2	0	0	0	1
6	A_2	No	B_2	B_2	0	0	0	1

G-computation [27] and inverse probability weighting (IPW) [18] are commonly used to estimate treatment regime effects and thus can be used for Wald-type hypothesis testing to compare strategy means. Unlike G-computation, IPW has become popular in applications for its ease of implementation and generality. For example, IPW does not estimate or make assumptions about distributions of intermediate outcomes. Various papers have suggested methods to compare different treatment strategies using IPW. Except for the limited work of Hernan et al., and Tang and Wahed [8, 38], most of these methods do not allow regression to adjust for baseline covariates. Using observational data, Hernan et al. [8] applied a Cox regression with artificial censoring to compare different two treatment strategies. Due to wide range of possible treatment regime options, observations were routinely artificially censored if they depart from the pre-specified regimes of interest. Hernan's approach is a motivation to build a unified regression model with the option of adjusting for covariates and interaction terms (between covariates and treatment strategies) in a model. Using IPW in a two-stage randomization setting, Lokhnygina and Helterbrand [17] derived an estimator for the log-hazard and a score test to compare treatment strategies under Cox model. However, this is limited to the comparison of a pair of treatment regimes without adjusting for time-varying treatment effects or other auxiliary variables. Tang and Wahed [38] suggested modeling each treatment strategy using Cox regression model that adjusts for both time dependent and time independent covariates. They fit a stratified Cox model that allows the underlying hazard to vary across strategies. Model comparisons among treatment strategies is performed using log

ratio of estimated cumulative hazards. Their approach to model each strategy and make a statistical comparison across strategies is the only work known thus far to compare strategies by adjusting for covariates.

We propose a consistent simple multiple artificial randomization tool (SMART) estimator that combines 'artificially randomized' estimators according to Rubin's (1977, 1987) [31, 32] multiple imputation method. Subsequently, the 'artificial randomization' approach is utilized in regression models to compare strategies adjusting for baseline covariates, and hence inform further useful strategies.

This chapter is organized as follows. In Section 3.2, we describe counterfactual and observed outcomes followed by an introduction to artificial randomization and the proposed SMART estimator in Section 3.2.1. This estimator is compared with IPW-based estimators in Section 3.2.2. Within this subsection, we derive model-based and robust variance estimates for each estimator. In Section 3.3, we conduct simulation studies to examine properties of our proposed estimator with respect to existing estimators. In Section 3.4, we demonstrate how regression analysis through artificial randomization allows adjusted comparison of ATSs and helps develop new strategies involving covariates. In Section 3.5.1, we apply the adopted procedures to the CALGB data and report strategy means, standard errors and confidence intervals. Appropriately, depending on significant interaction between covariate and strategies, we define new informative covariate-specific strategies. We wrap up the chapter with results in Section 3.5.2. In the CALGB data the outcome is a survival time but we treat it as continuous in our methodology and drop censored times. With only few censored observations (7%) the assumption seems to be reasonable.

3.2 ARTIFICIAL RANDOMIZATION, SMART, AND OTHER ESTIMATORS

The design described in Figure 3 (page 12) allows estimation of the effect of four treatment regimes, namely, $A_jB_k; j,k = 1,2$. Let X_{ji} be the assignment indicator for first-stage treatment A_j, where $X_{ji} = 1$, if the i^{th} patient is randomized to $A_j, j = 1,2; i = 1,2,...,n$; 0

otherwise. The indicator for individual response status, R_{ji}, is 1 for responders and 0 for non-responders to $A_j, j = 1, 2; i = 1, 2, ..., n$. For non-responders, $R_{ji} = 0$, define variable Z_{ki} to indicate assignment of second-stage treatments, $B_k, k = 1, 2$; conditional on $R_{ji} = 0$, $Z_{ki} = 1$, if the i^{th} patient receives B_k assignment; 0 otherwise. Assume V_i and Y_i to denote a vector of baseline covariates and the final outcome, respectively.

Inferences from sequentially randomized trials are often facilitated through defining counterfactual outcomes and corresponding expectations [40, 18, 22]. The goal is to estimate, for example, the mean of the outcome under a given strategy. Let $Y_i(A_j), j = 1, 2$ indicate outcomes under A_j if the patient had received A_j and responded to A_j (and hence there were no second randomization). Similarly, $Y_i(A_jB_k), j, k = 1, 2$ denotes outcomes under $A_jB_k, j, k = 1, 2$ if the patient had received $A_j, j = 1, 2$, did not respond and followed by $B_k, k = 1, 2$. The potential outcome if patient i follows regime A_jB_k would then be $Y_{jki} = R_{ji}Y_i(A_j) + (1 - R_{ji})Y_i(A_jB_k), j, k = 1, 2; i = 1, ..., n$, where Y_{jki} is the potential outcome for individual i under strategy A_jB_k. The observed outcome, for patient i following treatment regime A_jB_k, is connected to the potential outcome via the following formula,

$$Y_i = \sum_{i=1}^{n} \sum_{j=1}^{2} X_{ji} \left\{ R_{ji}Y_i(A_j) + (1 - R_{ji}) \sum_{k=1}^{2} Z_{ki}Y_i(A_jB_k) \right\}$$

which is known as the consistency assumption (Assumption 2.3, page 16). As defined in Chapter 2 (Section 2.3, page 15), the strategy mean is, $\mu_{jki} = \pi_j\mu_{A_j} + (1 - \pi_j)\mu_{A_jB_k}$, where $\mu_{A_j} = \mathrm{E}\{Y(A_j)\}$ is the sub-group mean of the population receiving A_j followed by A_j, $\mu_{A_jB_k} = \mathrm{E}\{Y(A_jB_k)\}$ is sub-group mean of the population receiving A_j followed by B_k.

3.2.1 ARTIFICIAL RANDOMIZATION AND THE SMART ESTIMATOR

Analysis of ATSs obtained from SMAR designs has thus far been limited to comparing means of strategies using Wald-type hypothesis testing. Comparing means of strategies using regression models that adjust for baseline covariates may be of further interest. However, as discussed before, such adjustments are not straightforward. The reason is that a regression model requires subjects belong uniquely to one strategy, which is not the case for SMAR trials. For example, a responder to A_1 is consistent with following strategies A_1B_1 and

A_1B_2. Therefore, to analyze the data from this specific design using regression or ANOVA framework, we need to devise a way to attach each subject to a unique strategy. This is what is the topic of this research.

Since responders are not randomized to further treatments, for the purpose of identifying them to unique strategies, at the analysis stage, we create an artificial randomization indicator for responders that assigns each responder to one of the second stage treatment, B_1 or B_2. It is important to emphasize that responders are artificially randomized for the sake of statistical analysis after all the data have been accumulated and locked. They are not re-randomized in the actual trial. The purpose is to uniquely associate each responder to one of the four strategies $A_jB_k, j, k = 1, 2$ so that we can readily employ ANOVA or regression techniques.

At the analysis stage, we create a randomization indicator, using the same mechanism that was used to randomize the non-responders. For example, if B_1 and B_2 were assigned randomly with probabilities Q_1 and $1 - Q_1$ respectively, then this artificial randomization indicator Z_i^* is defined as,

$$Z_i^* = \begin{cases} Z_i & \text{if } R_i{=}0 \\ \text{Bernoulli}(Q_1) & \text{if } R_i{=}1 \end{cases}$$

where Bernoulli(Q_1) indicates a randomly generated value of a Bernoulli random variable with probability Q_1. With this artificial augmentation, the observed data looks like $(V_i, X_i, R_i, Z_i^*, Y_i), i = 1, 2, ..., n$. Define $Z_{1i}^* = Z_i^*$ and $Z_{2i}^* = 1 - Z_i^*$ be the indicators for B_1 and B_2 (actual or artificial assignment). In this observed data, X_i and Z_i^* uniquely assigns a participant to one strategy. For example, patient 2 in Table 8, $X_i = 0$ and $Z_i^* = 0$, we know that the participant 'followed' (treated to have followed) strategy A_1B_1. With this 'augmented observed data' one can easily construct estimators for strategy means or use regression-based methods to compare strategies adjusting for baseline covariates. Here we illustrate by defining estimates of strategy means. In Section 3.4, we explain how this method can be adapted for regression or ANOVA.

Since X_i and Z_i^* uniquely identifies a participant with a specific strategy, an estimator of the $(j,k)^{th}$ strategy mean would be

$$\hat{\mu}_{jk}^{AR} = \frac{\sum_{i=1}^{n} X_{ji} Z_{ki}^* Y_i}{\sum_{i=1}^{n} X_{ji} Z_{ki}^*}, \tag{3.1}$$

the simple average of outcome over those who 'followed' strategy $A_j B_k$. Note that we have used the superscript 'AR' to indicate that this estimator uses artificial randomization. To show consistency of this estimator, note that,

$$\begin{aligned}
E(X_{ji} Z_{ki}^*) &= E\left[X_{ji} E\left[Z_{ki}^* | X_{ji}\right]\right] \\
&= E\left[X_{ji} E\{E(Z_{ki}^* | X_{ji}, R_i)\}\right] \\
&= E\left[X_{ji} E\{R_i Q_k + (1 - R_i) Q_k\}\right] \\
&= E\left[X_{ji} Q_k\right] \\
&= Q_k \kappa_j.
\end{aligned}$$

where $E\left[X_{ji}\right] = \kappa_j$. Also,

$$\begin{aligned}
E(X_{ji} Z_{ki}^* Y_i) &= E\left[X_{ji} Z_{ki}^* Y_{jki}\right] \qquad \text{by consistency assumption (Assumption 2.3, page 16),} \\
&= E\left[Y_{jki} E\{X_{ji} Z_{ki}^* | Y_{jki}\}\right] \\
&= E\left[Y_{jki} E(X_{ji} Z_{ki}^*)\right] \\
&= Q_k E\left[Y_{jki}\right] \kappa_j \\
&= Q_k \mu_{jk} \kappa_j.
\end{aligned}$$

Therefore, using the law of large numbers,

$$\begin{aligned}
\hat{\mu}_{jk}^{AR} &= \frac{\frac{1}{n}\sum_{i=1}^{n} X_{ji} Z_{ki}^* Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_{ji} Z_{ki}^*} \\
&\xrightarrow{p} \frac{E(X_{ji} Z_{ki}^* Y_i)}{E(X_{ji} Z_{ki}^*)} \\
&= \frac{Q_k \mu_{jk} \kappa_j}{Q_k \kappa_j} \\
&= \mu_{jk},
\end{aligned}$$

where "$\xrightarrow{p}$" denotes "convergence in probability".

Hence, $\hat{\mu}_{jk}^{AR}$ is a consistent estimator of μ_{jk}. The artificial randomization procedure facilitates independent estimation of strategy means to compare ATSs. However, additional uncertainty introduced into the artificial estimator due to the augmented randomization indicator needs to be accounted for before it can be compared with existing estimators.

In the spirit of Rubin's average estimator for multiple imputations, we propose to conduct multiple artificial randomization and average them to form a single estimator. We refer to this as a "SMART" (simple multiple artificial randomized tool) estimator. This is similar to the multiple imputation method that captures uncertainty due to stochastic randomness introduced via the imputation process. Denote by $\hat{\mu}_{jk}^{AR(m)}$ the estimate of μ_{jk}, from the mth "artificially randomized" data set ($m = 1, 2, ..., M$). The SMART estimate of μ_{jk} is the simple average of the artificial estimates,

$$\hat{\mu}_{jk}^{SMART} = \frac{1}{M} \sum_{m=1}^{M} \hat{\mu}_{jk}^{AR(m)}. \tag{3.2}$$

Rubin (1987) [32] introduced variance formula that has within- and between-imputation variability. To mimic his approach, define W and B to be the average "within-randomization" and "between-randomization" covariances of $\hat{\mu}_{jk}^{AR(m)}$, that is,

$$W = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_{jk}^{AR(m)} \tag{3.3}$$

and

$$B = \frac{1}{(M-1)} \sum_{m=1}^{M} (\hat{\mu}_{jk}^{AR(m)} - \hat{\mu}_{jk}^{SMART})^2, \tag{3.4}$$

where $\hat{V}_{jk}^{AR(m)}$ is the estimated variance for $\hat{\mu}_{jk}^{AR(m)}$ for the m^{th} artificial randomization. The estimated variance of $\hat{\mu}_{jk}^{SMART}$ is then given by,

$$T = W + \left(\frac{M+1}{M}\right) B. \tag{3.5}$$

To summarize the steps, we create M data sets from M artificial randomizations. The M data sets are analyzed using standard procedures to compute the means and then results from M analyses are combined to obtain a single estimator given in Equation 3.2.

Equation (3.1) gives an estimate of μ_{jk} based on single artificial randomization, which is equivalent to a "single imputation" technique that does not take into account variability between "imputations" (or randomizations in our case) (Schafer, 1999) [34]. To make up for the disadvantages associated with single imputation (such as underestimated standard errors), multiple imputation technique has been suggested (Rubin, 1977, 1987) [31, 32]. The proposed SMART estimator is mechanistically equivalent to multiple imputation technique. SMART would account for both within and between-randomization variance.

This estimator enables us to apply tests such as the F-test and the t-test that are inconvenient or even impossible to be applied when observations are not independent. Large sample properties of SMART are compared with the four existing estimators using simulation. Below is a formula for the variance of artificial estimator which is derived in Section 4.1.

$$\begin{aligned} \text{var}\left(\hat{\mu}_{jk}^{AR}\right) = \frac{1}{nQ_k\kappa_j}\left[\pi_j\{\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2 - \sigma^2_{A_jB_k} - (\mu_{jk} - \mu_{A_jB_k})^2\}\right] \\ + \frac{1}{nQ_k\kappa_j}\left[\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right]. \end{aligned} \tag{3.6}$$

The asymptotic variance of $\hat{\mu}_{jk}^{AR}$ can be estimated either by the empirical estimator (Section 4.2), or the model-based estimator which is obtained by substituting estimated quantities $\hat{\kappa}_j, \hat{Q}_k, \hat{\pi}_j, \hat{\sigma}^2_{A_j}, \hat{\mu}_{A_j}, \hat{\sigma}^2_{A_jB_k}, \hat{\mu}_{A_jB_k}, \hat{\mu}_{jk}$ into Equation(3.6). It is similarly computed for all other estimators. Section 4.1 refers to the variance formulas of estimators and Section 4.2 refers to their robust variance estimates.

3.2.2 IPW-BASED ESTIMATORS

We compare our SMART estimator to other competing estimators such as those based on inverse probability weighting. To motivate the inverse probability weighting, suppose we are interested in estimating μ_{11}, mean outcome for strategy A_1B_1. Imagine that everyone in the sample follows the regime A_1B_1 in Design 2 (Figure 3, page 12). In such a case, we would estimate the mean, μ_{11}, using the sample average. However, as in all SMAR designs, in our sample patients receive treatment sequence other than A_1B_1 (those who received B_2 or the non-responders to A_1 who are randomized to B_2) which makes sample average estimation

biased. To correct this bias, each subject following A_1B_1 is weighted by the reciprocal of the probability of receiving treatment B_1 (i.e., Q_1) to account for the "missing" or unaccounted individuals who received B_2. To elaborate, a non-responder who received B_1 could have equally likely received B_2. Thus, those who received B_1 would be weighed by $1/Q_1$ because they represent those who were randomized to B_2. To estimate the A_1B_1 adaptive treatment strategy, we consider outcomes of subjects consistent with following that treatment path.

The inverse probability weighted estimator (IPW) of μ_{jk} is given by,

$$\hat{\mu}_{jk}^{IPW} = \frac{1}{n}\sum_{i=1}^{n} W_{jki} Y_i, \tag{3.7}$$

where $W_{jki} = \frac{X_{ji}}{\kappa_j}\left\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right\}$ and κ_j is the probability assignment to first-stage treatments $A_j, j = 1, 2$. The corresponding asymptotic variance for this estimator is,

$$\text{var}\left(\hat{\mu}_{jk}\right) = \frac{1}{n}\left[\frac{1}{\kappa_j}\left\{\frac{\pi_j}{Q_k}(\sigma^2_{A_j} + \mu^2_{A_j}) + (1-\pi_j)(\sigma^2_{A_jB_k} + \mu^2_{A_jB_k})\right\} - \mu^2_{jk}\right]. \tag{3.8}$$

The model-based variance estimate is obtained by substituting $\hat{\kappa}_j, \hat{Q}_k, \hat{\pi}_j, \hat{\sigma}^2_{A_j}, \hat{\mu}_{A_j}, \hat{\sigma}^2_{A_jB_k}, \hat{\mu}_{A_jB_k}, \hat{\mu}_{jk}$ into Equation 3.8. Another variance estimate can be the so-called sandwich estimator (see Section 4.2).

A variation of the IPW is its normalized version. As shown in [11], the normalized inverse probability weighted (NIPW) is shown to have generally smaller variance than its counterpart IPW. The NIPW estimator is given by,

$$\hat{\mu}_{jk}^{NIPW} = \frac{\sum_{i=1}^{n} W_{jki} Y_i}{\sum_{i=1}^{n} W_{jki}}. \tag{3.9}$$

The variance of NIPW is given by,

$$\text{var}(\hat{\mu}_{jk}) = \frac{1}{n}\left[\frac{\pi_j}{\kappa_j}\left\{\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2\right\} + \frac{(1-\pi_j)}{\kappa_j Q_k}\left\{\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right\}\right]. \tag{3.10}$$

The model-based variance estimate is obtained by substituting estimated quantities $\hat{\kappa}_j, \hat{Q}_k, \hat{\pi}_j, \hat{\sigma}^2_{A_j}$, $\hat{\mu}_{A_j}, \hat{\sigma}^2_{A_jB_k}, \hat{\mu}_{A_jB_k}, \hat{\mu}_{jk}$ into Equation (3.10). The robust variance estimator is given in Section 4.2.

In the above IPW and NIPW estimators, the randomization probabilities are assumed to be known. But, one may wish to estimate them from the data. Therefore, below we present

another version of IPW and NIPW estimators (denoted as IPW1 and NIPW1, respectively) that estimates κ_j and Q_k's from the data:

$$\hat{\mu}_{jk}^{IPW1} = \frac{1}{n}\sum_{i=1}^{n} \hat{W}_{jki} Y_i \tag{3.11}$$

and

$$\hat{\mu}_{jk}^{NIPW1} = \frac{\sum_{i=1}^{n} \hat{W}_{jki} Y_i}{\sum_{i=1}^{n} \hat{W}_{jki}}, \tag{3.12}$$

where $\hat{W}_{jki} = \frac{X_{ji}}{\hat{\kappa}_j}\left\{R_i + \frac{(1-R_i)Z_{ki}}{\hat{Q}_k}\right\}$, $\hat{\kappa}_j = \frac{n}{\sum_{i=1}^{n} X_{ji}}$, $\hat{Q}_k = \frac{\sum_{i=1}^{n} X_{ji}(1-R_i)Z_{ki}}{\sum_{i=1}^{n} X_{ji}(1-R_i)}$. Details of variance formulas and their robust estimates are shown in Sections 4.1 and 4.2. In the next subsection we present the results from a simulation study to see how SMART estimator compare to these existing estimators.

3.3 SIMULATION STUDY

We conducted a simulation study to evaluate the large sample properties of SMART estimator with moderate sample sizes (150 and 300) and compare its performance with IPW-based estimators. For each individual in the population, a hypothetical design similar to Design 3 on page 12 assumes that subgroup populations $Y_i(A_j)$ and $Y_i(A_jB_k)$ come from a normal distribution with subgroup means: $\mu_{A_j} = \mu_{A_jB_2} = 15$, $\mu_{A_jB_1} = 20$, $\mu_{A_jB_2} = 22$; subgroup variances: $\sigma^2_{A_j} = 2^2$, $\sigma^2_{A_jB_k} = 8^2$, for $j, k = 1, 2$. The initial treatment indicator X_i is generated from a binomial distribution with probability 0.5 of randomization to either A_1 or A_2. The response indicator R_i is drawn from a Bernoulli distribution with probability of success π_1 for initial treatment A_1 and π_2 for A_2. Randomization indicator Z_i for non-responders is generated from Bernoulli distribution with probability Q_1. By definition, the outcome Y_i under strategy A_jB_k should follow a normal distribution since $Y_i(A_j)$ and $Y_i(A_jB_k)$ are generated from a normal distribution. We considered many simulation scenarios for various sample sizes and population parameters such as μ_{A_j}, $\mu_{A_jB_k}$, $\sigma^2_{A_j}$, $\sigma^2_{A_jB_k}$, π_1, π_2, and Q_k. The true strategy means are computed using the formula $\mu_{jk} = \pi_j\mu_{A_j} + (1-\pi_j)\mu_{A_jB_k}$.

We generated 10000 samples from the above population to compare the competing estimators. To evaluate the performance of the estimators we calculate Monte Carlo (MCSE), relative bias (RB, %), model-based (MBSE), robust (RSE) standard errors and coverage probability (CP) of 95% Wald confidence intervals. These quantities are arranged as columns in Tables 9 and 10. The rows correspond to strategy estimators. Specifically, there are four strategies: A_1B_1, A_1B_2, A_2B_1 and A_2B_2.

Theoretically, we expect SMART estimators to be less efficient than NIPW as it uses less number of non-responders. We provide simulation evidence summarizing this fact. SMART are consistent estimators which makes them desirable in their application in regression based adaptive strategy comparison (Section 3.4). The main results of the simulation are presented in Tables 9 and 10, where the ratio of subgroup variance is held $\frac{\sigma^2_{A_j}}{\sigma^2_{A_jB_k}} = \frac{2^2}{8^2}$ for a given sample size (n=150, n=300). We present mean and standard error estimates for SMART when M=1 (SMART1) and M=5 (SMART5) based on Equation (3.5). The table rows are estimators SMART1, SMART5, IPW, IPW1, NIPW, NIPW1. The first thing to note in both tables is that the new estimator is approximately unbiased. Absolute relative biases (RB) were less than 0.087% for both SMART1 and SMART5. Both MBSE and RSE are close to the MCSE for SMART1 and SMART5. SMART5 has slightly higher MBSE than SMART1. As anticipated, the variance for SMART1 is slightly smaller, as the between imputation variance (B) is zero. They both approximately guarantee 95% CP. In Table 9 RSE and MCSE comparison indicates that SMART1 attains very close estimates to SMART5.

In addition to SMART1 and SMART5, Table 9 provides four estimators produce unbiased estimators of strategy means: μ_{11}, μ_{12}, μ_{21}, μ_{22}. Since it is shown that the IPW has greater estimated strategy mean variance than its normalized version [11], we will focus on the comparison of SMART estimators with the normalized versions of IPW across the MBSE, RSE and MCSE. The IPW-based estimators are approximately unbiased with relative biases less than 0.2% for IPW, less than 0.071% for IPW1, less than 0.3% for NIPW and less than 0.071% for NIPW1. Comparing columns MBSE, RSE and MCSE for SMART1 and SMART5 against the other estimators show a slightly higher standard error and hence are less efficient. However, quite surprisingly, the IPW1 is almost equally efficient as the normalized versions (NIPW and NIPW1).

The agreement between the asymptotic standard errors and Monte Carlo standard error suggests accuracy of the variance formulas of the estimators.

For the most part, all the estimators achieved coverage probabilities close to the nominal 95%. One can note, the NIPW and NIPW1 provided lower coverage probabilities of 90% and 90%, respectively, for strategy A_2B_1. This calls for more examination of this strategy.

Table 10 shows simulation results when the sample size is increased to 300. Expectedly, the standard errors of the estimators decreased. For the SMART estimator, the conclusions were analogous to results in Table 9. In the same vein, comparisons across estimators are consistent with results in Table 9. Due to equal response rates, π_j's, Tables 9 and 10 show that estimated means and standard errors are nearly the same for strategies $A_jB_1, j = 1, 2$. Common in both tables is that IPW estimator strategies A_1B_1 and A_2B_1 display higher standard errors across columns MBSE, RSE and MCSE.

3.4 ARTIFICIAL RANDOMIZATION AND REGRESSION

It might seem that the proposed SMART approach to estimating the strategy means from SMART designs is somewhat against the principle of sequential randomization. While this provides approximately unbiased estimates, they are not more efficient than NIPW estimators. This is expected, as the SMART approach uses less observations (e.g., half of the responders) than NIPW. However, the utility of SMART approach is in its simplicity to allow users to estimate/test strategy effects adjusting for other covariates. To show this, consider that in a population where patients follow particular treatment strategies, the mean response is modeled by

$$Y_i = \beta_0 + \sum_{j,k=1,2} \beta_{jk} S_{jki} + \gamma^T V_i + \sum_{j,k=1,2} \alpha_{jk}^T V_i S_{jki} + \epsilon_i, \tag{3.13}$$

with $\mathrm{E}(\epsilon_i) = 0$, where parameters β_{jk}, γ and α_{jk} represent vector of coefficients for strategies S_{jk}, covariates V and their interaction $S * V$. S_{jk} is defined as $S_{jk} = 1$ if patient i follows strategy A_jB_k, 0, otherwise.

Table 9: **Simulation results of estimators based on 10000 Monte Carlo (MC) samples of size 150.**

Estimator	EST	RB	MBSE	RSE	MCSE	CP
SMART1	15.01	-0.087	1.27	1.25	1.30	0.93
	20.01	-0.035	0.96	0.95	0.97	0.94
	14.99	0.027	1.27	1.25	1.34	0.92
	20.02	-0.075	0.96	0.95	0.97	0.94
SMART5	15.01	-0.087	1.39	1.36	1.23	0.96
	20.01	-0.045	0.98	0.97	0.94	0.96
	15.02	-0.15	1.39	1.36	1.30	0.94
	19.99	0.025	0.98	0.97	0.96	0.95
IPW	15.03	-0.20	1.85	1.83	1.86	0.93
	19.99	0.067	2.52	2.47	2.61	0.93
	14.97	0.18	1.86	1.83	1.92	0.93
	19.99	0.032	2.53	2.47	2.66	0.93
IPW1	15.01	-0.051	1.16	1.10	1.11	0.94
	19.99	0.071	0.94	0.94	0.95	0.93
	14.99	0.03	1.15	1.10	1.19	0.93
	19.99	0.04	0.95	0.94	0.97	0.92
NIPW	15.01	-0.29	1.16	1.16	1.19	0.92
	19.99	0.074	0.92	0.92	0.93	0.93
	14.99	-0.30	1.17	1.16	1.26	0.90
	19.99	0.038	0.93	0.93	0.94	0.93
NIPW1	15.00	-0.051	1.18	1.10	1.12	0.92
	19.99	0.071	0.94	0.94	0.95	0.93
	15.00	0.026	1.18	1.10	1.19	0.90
	19.99	0.041	0.94	0.94	0.97	0.93

Monte Carlo mean estimates (EST), relative bias (RB, %), Monte Carlo standard errors (MCSE), model-based standard errors (MBSE), robust standard errors (RSE) and 95% coverage probabilities (CP). CP is computed based on robust variance estimates. True strategy means: $\mu_{11} = \mu_{21} = 15$, $\mu_{12} = \mu_{22} = 20$.

Table 10: **Simulation results of estimators based on 10000 Monte Carlo (MC) samples of size 300.**

Estimator	EST	RB	MBSE	RSE	MCSE	CP
SMART1	14.99	0.067	0.89	0.88	0.91	0.94
	20.01	-0.05	0.68	0.67	0.67	0.95
	14.99	0.067	0.89	0.88	0.96	0.92
	20.02	-0.10	0.68	0.67	0.67	0.95
SMART5	15.02	-0.13	0.97	0.88	0.92	0.93
	20.02	-0.085	0.69	0.67	0.68	0.94
	14.99	0.01	0.97	0.88	0.97	0.92
	20.01	-0.045	0.69	0.67	0.67	0.95
IPW	14.99	0.015	1.31	1.30	1.33	0.94
	20.01	-0.097	1.78	1.76	1.87	0.93
	14.99	0.01	1.31	1.30	1.34	0.94
	20.01	-0.031	1.77	1.76	1.85	0.94
IPW1	15.00	-0.061	0.86	0.78	0.78	0.94
	20.01	-0.055	0.67	0.66	0.67	0.93
	14.99	0.024	0.86	0.78	0.83	0.94
	20.00	-0.01	0.67	0.66	0.68	0.93
NIPW	15.03	-0.23	0.83	0.83	0.84	0.93
	20.01	-0.052	0.66	0.66	0.67	0.94
	15.02	-0.12	0.83	0.83	0.89	0.92
	20.00	-0.01	0.66	0.66	0.67	0.94
NIPW1	15.01	-0.061	0.83	0.78	0.78	0.93
	20.01	-0.055	0.67	0.66	0.67	0.94
	14.99	0.024	0.83	0.78	0.83	0.92
	20.00	-0.01	0.67	0.66	0.68	0.94

Monte Carlo mean estimates (EST), relative bias (RB, %), Monte Carlo standard errors (MCSE), model-based standard errors (MBSE), robust standard errors (RSE) and 95% coverage probabilities (CP). CP is computed based on robust variance estimates. True strategy means: $\mu_{11} = \mu_{21} = 15$, $\mu_{12} = \mu_{22} = 20$.

This is a multiple linear regression model, and if data were available from an upfront randomization design (Ko and Wahed, 2012) [11], the parameters could be estimated using regular regression analysis. As explained in the previous section, due to the non-separability of the strategy indicators (a given observation can be associated to more than one strategy) such an analysis is not feasible for SMAR designs. But note that artificial randomization allows us to fit this regression model. As defined in Section 3.2.1, in terms of our artificially randomized data, for a subject following regime A_jB_k would belong to $S_{jki} = X_{ji}Z_i^*$.

3.5 DATA ANALYSIS OF CALGB DATA

3.5.1 SAMPLE AND MEASURES

We apply the artificial randomization method and fit the model to the CALGB data described in Chapter 1. There were a total of 388 patients who participated in the CALGB 8923 study of which 193 were randomized to the GM-CSF arm, and 195 were randomized to the placebo arm. Of the the GM-CSF arm, 79 were responders and of the placebo arm, 90 were responders. Out of 79 responders to GM-CSF, 37 were randomized to intensification therapy I while 45 out of 90 placebo responders were randomized to intensification therapy I. The rest were randomized to intensification therapy II. The artificial randomization was applied to the subset of non-responders (114 from the GM-CSF arm and 105 from the placebo arm).

Non-responders were superficially assumed to have received either intensification therapy I or II with probability 0.5. Accordingly, we constructed strategy indicators to be used as independent variables in the regression model as described in Section 3.4. We used R (version 2.15) to fit Equation 3.13 and compute the estimators in Table 11. Table 11 depicts mean estimates of the five strategies using SMART1, SMART5 IPW, IPW1, NIPW, NIPW1 estimators. In the analyses, we considered age (mean=69.5), sex (220 males, 168 females), and strategy indicators as covariates in the artificial regression. The main outcome is survival time but it is treated as being continuous.

Table 11: **Estimated strategy means and robust standard errors from the analysis of CALGB 8923 data.**

Strategy	SMART1 (SE)	SMART5 (SE)	IPW (SE)	IPW1 (SE)	NIPW (SE)	NIPW1 (SE)
A_1B_1	487.16 (70.58)	468.56 (70.11)	454.0 (70.9)	478.5 (57.9)	468.5 (59.5)	478.5 (66.5)
A_1B_2	521.21 (75.97)	536.46 (80.74)	454.0 (70.9)	528.0 (69.0)	468.5 (59.5)	528.0 (69.0)
A_2B_1	660.16 (88.23)	623.52 (90.12)	623.6 (91.3)	620.4 (71.6)	620.4 (73.4)	620.4 (71.6)
A_2B_2	592.72 (83.68)	627.48 (99.09)	632.6 (91.3)	629.4 (79.5)	620.4 (73.4)	629.4 (79.5)

3.5.2 RESULTS

From Table 11 rows are strategies and columns are strategy mean estimators. For each strategy the estimates are similar across each estimator. The estimated means of strategies sharing A_1 are less than the strategies sharing A_2 for all estimators. Similar observations can be made about the corresponding standard errors. The NIPW, NIPW1 and IPW1 gave smaller estimated standard errors than SMART which was noted in the simulation studies. SMART5 and IPW have greater standard errors than other estimators. In particular, standard error estimates for SMART5 are considerably large.

As part of the modeling strategy we identified age, sex and strategy indicators from the CALGB data as potential predictors in our regression model. Further, we explore possible interaction between the strategies and age and sex. Practically, we expect significant interactions to generate new strategies that specify long term survival of AML patients. In more detail, we did the following. Three interaction models were fitted. The first contained two interactions, namely, strategy by age and strategy by sex. The second and third models have strategy by age and strategy by sex interactions. All the interaction models were found to be not significant. To apply our method, we decide to keep the model with strategy and sex interaction. Had the interaction between strategy and sex were significant, we would choose the model in Table 12 for purposes of constructing informative strategies. Except for age (P=0.0024), strategy, sex and the interaction terms are not significant. The p-value for the interaction between strategy and sex is 0.086. Assuming significant interaction exists

Table 12: **Parameter Estimates for Model of CALGB 8923 data.**

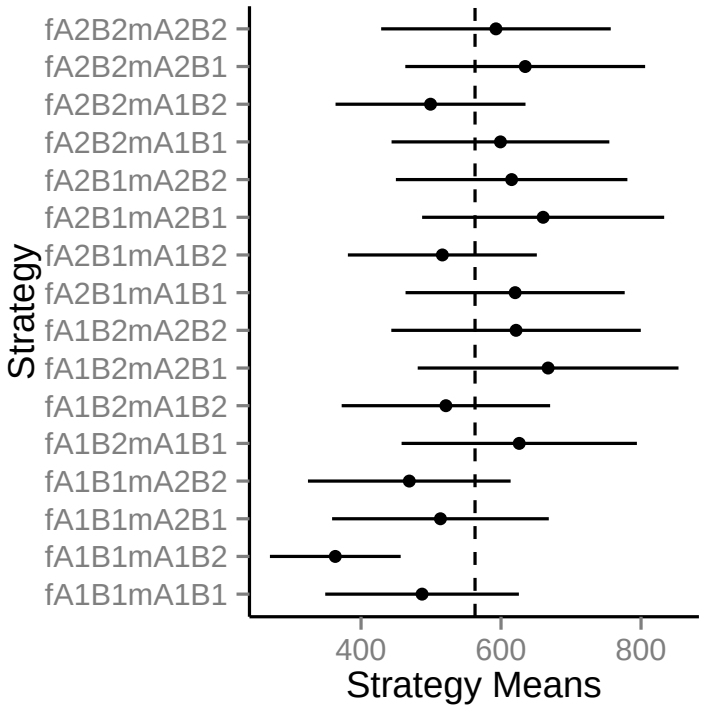
Parameter	Estimate	SE	Z	Pr $>\|Z\|$
Intercept	1996.75	487.42	4.10	<0.001
$S_{A_1B_1}$	25.90	146.64	0.18	0.86
$S_{A_1B_2}$	-156.38	149.07	-1.10	0.29
$S_{A_2B_1}$	71.50	154.77	0.46	0.64
Sex	95.74	154.75	0.62	0.54
Age	-20.98	6.87	-3.05	0.0024
$S_{A_1B_1}$*Sex	-309.39	230.22	-1.34	0.18
$S_{A_1B_2}$*Sex	281.88	221.34	1.27	0.20
$S_{A_2B_1}$*Sex	-23.34	224.15	-0.10	0.92

between sex and treatment strategies, we defined 16 new strategies to treat patients according to their sex. One strategy could be, "If female treat with A_1B_1, else treat with A_1B_1", which is denoted by "fA_1B_1mA_1B_1". Table 13 shows estimated means. Also, displayed are robust standard errors and confidence intervals for the sixteen strategies. Figure 5 displays the mean estimates and corresponding 95% confidence interval. Significant strategies are either to the right or left of the mean dotted vertical line. Similar plot called forest plot is used to display the effect of odds ratio for logistic regression results. As can be gathered from the table and figure, Strategies "fA_1B_1mA_1B_2" and "fA_1B_2mA_2B_1" have the smallest and highest mean and standard errors for the SMART estimators, respectively. As our simulation results would support the robust standard error estimates for SMART5 estimator is greater than SMART1 estimator.

This chapter introduced an unbiased SMART estimator that is based on multiple imputation concepts. As a consequence, it introduced the artificial randomization technique to compare strategies from SMAR design using regression methods.

Table 13: **Analysis of CALGB 8923 data for sixteen new strategies.**

Strategy	SMART1	SE	CI1	SMART5	SE	CI5
fA_1B_1mA_1B_1	487.16	70.58	(348.82, 625.50)	487.16	70.58	(348.82, 625.50)
fA_1B_1mA_1B_2	363.18	47.65	(269.79, 456.57)	363.18	47.65	(269.79, 456.57)
fA_1B_1mA_2B_1	513.41	78.98	(358.61, 668.21)	513.41	78.98	(358.61, 668.21)
fA_1B_1mA_2B_2	468.93	73.87	(324.15, 613.72)	468.93	73.87	(324.14, 613.72)
fA_1B_2mA_1B_1	626.00	85.76	(457.91, 794.10)	626.00	190.45	(252.72, 999.28)
fA_1B_2mA_1B_2	521.21	75.97	(372.31, 670.11)	521.21	207.92	(113.69, 928.73)
fA_1B_2mA_2B_1	667.22	95.05	(480.92, 853.52)	667.22	211.00	(253.66, 1080.78)
fA_1B_2mA_2B_2	621.43	90.93	(443.21, 799.65)	621.43	207.73	(214.28, 1028.58)
fA_2B_1mA_1B_1	620.14	79.88	(463.58, 776.70)	620.14	181.40	(264.60, 975.68)
fA_2B_1mA_1B_2	516.16	68.88	(381.16, 651.16)	516.16	199.62	(124.91, 907.41)
fA_2B_1mA_2B_1	660.16	88.23	(487.23, 833.09)	660.16	200.22	(267.73, 1052.59)
fA_2B_1mA_2B_2	615.13	84.38	(449.74, 780.51)	615.13	197.95	(227.15, 1003.11)
fA_2B_2mA_1B_1	599.14	79.38	(443.56, 754.72)	599.14	158.46	(288.56, 909.72)
fA_2B_2mA_1B_2	499.26	69.24	(363.55, 634.97)	499.26	180.48	(145.52, 853.00)
fA_2B_2mA_2B_1	634.54	87.41	(463.22, 805.86)	634.54	172.20	(297.03, 972.05)
fA_2B_2mA_2B_2	592.73	83.68	(428.72, 756.74)	592.73	173.18	(253.30, 932.16)



Means of sixteen strategies with their confidence intervals are displayed in a forest plot where the dots are strategy mean estimates, and lines indicate ± SE.

Figure 5: **SMART1.**

4.0 VARIANCE FORMULAS OF ESTIMATORS

4.1 VARIANCE FORMULAS

M-estimation or synonymously referred to as estimating equations are empirical or quasi-likelihood based functions. Quasi-likelihood functions depend on specified means and variance relations of each observation [1]. Thus, M-estimation derives estimators by solving estimating equations that are not necessarily constructed from the derivative of the log-likelihood. Huber (1964, 1967) introduced and developed the asymptotic properties of M-estimators. The estimating equation can be written as $\sum_{i=1}^{n} \psi(Y_i, \Theta) = 0$, where the M-estimator $\hat{\Theta}$ satisfies the equation,

$$\sum_{i=1}^{n} \psi(Y_i, \hat{\Theta}) = 0.$$

The vector $Y_i, i = 1, ..., n$ are independent and may not be identically distributed, Θ is p-dimensional parameter, and ψ is a known $(p \times 1)$ influence function independent of i or n. It is shown that $\hat{\Theta} \sim N\left(\Theta, \frac{V(\Theta)}{n}\right)$ as n $\to \infty$, where $V(\Theta) = A(\Theta)^{-1}B(\Theta)\{A(\Theta)^{-1}\}^T$. Formally, assuming $Y_1, Y_2, ..., Y_n$ are i.i.d and follow the distribution F, model-based estimators for A and B are given by

$$A(Y, \hat{\Theta}) = E_F\{-\psi^{'}(Y_i, \hat{\Theta})\}$$

and

$$B(Y, \hat{\Theta}) = E_F\{\psi(Y_i, \hat{\Theta})\psi(Y_i, \hat{\Theta})^T\}$$

respectively.

Using averages instead of expectations, the empirical estimators for A and B are given by

$$A_n(Y,\hat{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}\{-\psi^{'}(Y_i,\hat{\Theta})\}$$

and

$$B_n(Y,\hat{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i,\hat{\Theta})\psi(Y_i,\hat{\Theta})^T$$

respectively.

Below we derive the asymptotic variance for the proposed and existing estimators and in Section 4.2, we compute the empirical variance estimators.

1. **Variance of IPW Estimator**

   Let $\Theta_0 = (\mu_{jk})$ and the ψ- function $\psi(Y_i,\Theta_0) = \psi_1(Y_i,\Theta_0) = \{W_{jki}Y_i - \mu_{jk}\}$, $\psi^{'} = -1$, $\psi(Y_i,\Theta_0)\psi^T(Y_i,\Theta_0) = \{W_{jki}Y_i - \mu_{jk}\}^2$. Then $A(\Theta) = E\{-\psi^{'}(Y_i,\Theta_0)\} = 1$, $B(\Theta) = E\{\psi(Y_i,\Theta_0)\psi^T(Y_i,\Theta_0)\} = E\left[\left(\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\}Y_i - \mu_{jk}\right)^2\right]$

$$\begin{aligned}
&= E\left[\left(\frac{X_{ji}}{\kappa_j}\{R_iY_i(A_j) + \frac{(1-R_i)Z_{ki}Y_i(A_jB_k)}{Q_k}\} - \mu_{jk}\right)^2\right] \\
&= E\left[\frac{X_{ji}}{\kappa_j^2}\{R_iY_i^2(A_j) + \frac{(1-R_i)Z_{ki}Y_i^2(A_jB_k)}{Q_k^2}\}\right] \\
&\quad -2E\left[\mu_{jk}\frac{X_{ji}}{\kappa_j}\{R_iY_i(A_j) + \frac{(1-R_i)Z_{ki}Y_i(A_jB_k)}{Q_k}\} + \mu_{jk}^2\right] \\
&= E\left[\frac{X_{ji}}{\kappa_j^2}\{R_iY_i^2(A_j) + \frac{(1-R_i)Z_{ki}Y_i^2(A_jB_k)}{Q_k^2}\} - \mu_{jk}^2\right] \\
&= E\left[E\left[\frac{X_{ji}}{\kappa_j^2}\{R_iY_i^2(A_j) + \frac{(1-R_i)Z_{ki}Y_i^2(A_jB_k)}{Q_k^2}\} - \mu_{jk}^2\right]|X_{ji},R_i,Y_i(A_j),Y_i(A_jB_k)\right]
\end{aligned}$$

   After a few steps of iterative expectations, we obtain

$$B(\Theta) = \frac{1}{\kappa_j}\{\frac{\pi_j}{Q_k}(\sigma^2_{A_j} + \mu^2_{A_j}) + (1-\pi_j)(\sigma^2_{A_jB_k} + \mu^2_{A_jB_k})\} - \mu^2_{jk}.$$

   Therefore, the IPW variance estimator is

$$\begin{aligned}
\text{var}(\hat{\Theta}) &= \frac{1}{n}A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T \\
&= \frac{1}{n\kappa_j}\{\frac{\pi_j}{Q_k}(\sigma^2_{A_j} + \mu^2_{A_j}) + (1-\pi_j)(\sigma^2_{A_jB_k} + \mu^2_{A_jB_k})\} - \mu^2_{jk}.
\end{aligned}$$

2. **Variance of NIPW Estimator**

Let $\Theta_0 = (\mu_{jk})$ and the ψ- function $\psi(Y_i, \Theta_0) = W_{jki}(Y_i - \mu_{jk})$, where $\psi^{'}(Y_i, \Theta_0) = -W_{jki}$, $\psi(Y_i, \Theta_0)\psi^T(Y_i, \Theta_0) = \{W_{jki}(Y_i - \mu_{jk})\}^2$. Then $A(\Theta) = E\{-\psi^{'}\} = E\{W_{jki}\}$, $B(\Theta) = E\{\psi(Y_i, \Theta_0)\psi^T(Y_i, \Theta_0)\} = E\left[\frac{X_{ji}}{\kappa_j^2}\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\}^2\{Y_i - \mu_{jk}\}^2\right]$. Let's expand and evaluate $A(\Theta)$ and $B(\Theta)$.

$$
\begin{aligned}
A(\Theta) &= E\{-\psi^{'}\} = E\{W_{jki}\} \\
&= E\left[\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_k}{Q_k}\}\right] \\
&= E\left[E\left[\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_k}{Q_k}\}|X_{ji}, R_i\right]\right] \\
&= E\left[\frac{X_{ji}}{\kappa_j}E\left[\{R_i + \frac{(1-R_i)Z_k}{Q_k}\}|X_{ji}, R_i\right]\right] \\
&= E\left[\frac{X_{ji}}{\kappa_j}\left[R_i + (1-R_i)\right]\right] \\
&= E\left[\frac{X_{ji}}{\kappa_j}\right] = 1.
\end{aligned}
$$

$$
\begin{aligned}
B(\Theta) &= E\{\psi(Y_i, \Theta_0)\psi^T(Y_i, \Theta_0)\} \\
&= E\left[\frac{X_{ji}}{\kappa_j^2}\left\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right\}^2\{Y_i - \mu_{jk}\}^2\right] \\
&= E\left[E\left[\frac{X_{ji}}{\kappa_j^2}\left\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right\}^2\{R_iY_i(A_j) + (1-R_i)Y_i(A_jB_k) - \mu_{jk}\}^2\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&= E\left[E\left[\frac{X_{ji}R_i}{\kappa_j^2}\left[R_iY_i^2(A_j) - 2R_i\mu_{jk}Y_i(A_j) + \mu_{jk}^2\right]\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&\quad + E\left[E\left[\frac{X_{ji}(1-R_i)Z_{ki}}{\kappa_j^2Q_k^2}\left[(1-R_i)Y_i^2(A_jB_k)\right]\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&\quad - E\left[E\left[\frac{X_{ji}(1-R_i)Z_{ki}}{\kappa_j^2Q_k^2}\left[2(1-R_i)\mu_{jk}Y_i(A_jB_k) + \mu_{jk}^2\right]\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&= E\left[E\left[\frac{X_{ji}}{\kappa_j^2}\left\{R_i\left[Y_i^2(A_j) - 2\mu_{jk}Y_i(A_j) + \mu_{jk}^2\right]\right\}\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&\quad + E\left[Z_{ki}|X_{ji}, R_i, Y(A_j), Y(A_jB_k)\right] \\
&\quad E\left[E\left[\frac{X_{ji}}{\kappa_j^2}\left\{\frac{(1-R_i)}{Q_k^2}\left[Y_i^2(A_jB_k) - 2\mu_{jk}Y_i(A_jB_k) + \mu_{jk}^2\right]\right\}\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right]
\end{aligned}
$$

Taking the expected values and rearranging the terms leads to

$$= \frac{\pi_j}{\kappa_j}\left[\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2\right] + \frac{(1-\pi_j)}{\kappa_j Q_k}\left[\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right].$$

Therefore, the NIPW variance estimator is

$$\text{var}(\hat{\Theta}) = \frac{1}{n}A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T$$

$$= \frac{\pi_j}{n\kappa_j}\left[\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2\right] + \frac{(1-\pi_j)}{n\kappa_j Q_k}\left[\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right].$$

3. **Variance of AR Estimator**

Let $\Theta_0 = (\mu_{jk})$ and the ψ- function $\psi(Y_i, \Theta_0) = W_{jki}(Y_i - \mu_{jk})$, where $\psi'(Y_i, \Theta_0) = -W_{jki}$, $\psi(Y_i, \Theta_0)\psi^T(Y_i, \Theta_0) = \{W_{jki}(Y_i - \mu_{jk})\}^2$. Then $A(\Theta) = E\{-\psi'\} = E\{W_{jki}\}$, $B(\Theta) = E\{\psi(Y_i, \Theta_0)\psi^T(Y_i, \Theta_0)\} = E\left[X_{ji}\left\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\right\}^2\left\{Y_i - \mu_{jk}\right\}^2\right]$. Let's expand and evaluate $A(\Theta)$ and $B(\Theta)$.

$$A(\Theta) = E\{-\psi'\} = E\{W_{jki}\}$$

$$= E\left[X_{ji}\left\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\right\}\right]$$

$$= E\left[E\left(X_{ji}\left\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\right\}|X_{ji}, R_i\right)\right]$$

$$= E\left[X_{ji}\left\{R_iP_{ki} + (1-R_i)P_{ki}\right\}\right]$$

$$= E\left[X_{ji}P_{ki}\right] = \kappa_j Q_k.$$

$$B(\Theta) = E\{\psi(Y_i, \Theta_0)\psi^T(Y_i, \Theta_0)\}$$

$$\begin{aligned}
&= E\left[X_{ji}\left\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\right\}^2\left\{Y_i - \mu_{jk}\right\}^2\right] \\
&= E\left[E\left[X_{ji}\left\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\right\}^2\left\{R_iY_i(A_j) + (1-R_i)Y_i(A_jB_k) - \mu_{jk}\right\}^2\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&= E\left[E\left[X_{ji}R_iZ^*_{ki}\left[R_iY_i^2(A_j) - 2R_i\mu_{jk}Y_i(A_j) + \mu^2_{jk}\right]\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&\quad + E\left[E\left[X_{ji}(1-R_i)Z_{ki}\left[(1-R_i)Y_i^2(A_jB_k) - 2(1-R_i)\mu_{jk}Y_i(A_jB_k) + \mu^2_{jk}\right]\right]|R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&= E\left[E\left[X_{ji}R_i\left(Y_i^2(A_j) - 2\mu_{jk}Y_i(A_j) + \mu^2_{jk}\right)|X_{ji}, R_i, Y_i(A_j), Y_i(A_jB_k)\right]\right] \\
&\quad E\left[Z^*_{ki}|X_{ji}, R_i, Y_i(A_j), Y_i(A_jB_k)\right] \\
&\quad + E\left[E\left[X_{ji}(1-R_i)\left[(1-R_i)Y_i^2(A_jB_k) - 2(1-R_i)\mu_{jk}Y_i(A_jB_k) + \mu^2_{jk}\right]|X_{ji}, R_i, Y_i(A_j), Y_i(A_jB_k)\right]\right] \\
&\quad E\left[Z_{ki}|X_{ji}, R_i, Y_i(A_j), Y_i(A_jB_k)\right]
\end{aligned}$$

Taking the expected values and rearranging the terms leads to,

$$= Q_k\kappa_j\pi_j\left[\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2 - \sigma^2_{A_jB_k} - (\mu_{jk} - \mu_{A_jB_k})^2\right] + Q_k\kappa_j\left[\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right].$$

Therefore, variance of AR estimator is

$$var(\hat{\Theta}) = \frac{1}{n}A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T$$

$$= \frac{1}{nQ_k\kappa_j}\left[\pi_j\{\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2 - \sigma^2_{A_jB_k} - (\mu_{jk} - \mu_{A_jB_k})^2\}\right] + \frac{1}{nQ_k\kappa_j}\left[\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right].$$

4. **<u>Variance of NIPW1 Estimator</u>**

Let $\Theta_0 = (\kappa_j, Q_k, \mu_{jk})$ and ψ- function

$$\psi(Y_i, \Theta_0) = \begin{bmatrix} X_{ji} - \kappa_j \\ X_{ji}(1 - R_i)(Z_{ki} - Q_k) \\ \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]\{Y_i - \mu_{jk}\} \end{bmatrix},$$

where

$$-\psi'(Y_i, \Theta_0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1 - R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_k}{Q_k}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_k}{Q_k^2}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_k}{Q_k}\right] \end{bmatrix}.$$

Then

$A(\Theta) = E\{-\psi(Y_i, \Theta_0)\}$

$$= E\begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1 - R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_{ki}}{Q_k^2}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right] \end{bmatrix},$$

where

i. $E\{X_{ji}(1 - R_i)\} = E\left[E\{X_{ji}(1 - R_i)\}|X_{ji}\right] = E\left[X_{ji}E\{(1 - R_i)\}|X_{ji}\right] = E\{X_{ji}(1 - \pi_j)\} = \kappa_j(1 - \pi_j)$.

ii. $E\frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]\{Y_i - \mu_{jk}\} = 0$, since $\mathrm{E}(Y_i) = \mu_{jk}$.

iii. $\frac{X_{ji}}{\kappa_j}\left[\frac{R_i Z_{ki}}{P_k^2}\right]\{Y_i - \mu_{jk}\} = 0$, since $\mathrm{E}(Y_i) = \mu_{jk}$.

iv. $E\left[\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\}\right]$

$$= E\left[E\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\}|X_{ji}, R_i\right]$$
$$= E\left[E\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\}|X_{ji}, R_i\right]$$
$$= E\left[\frac{X_{ji}}{\kappa_j}E\{R_i + (1-R_i)\}|X_{ji}\right]$$
$$= E\left[\frac{X_{ji}}{\kappa_j}\right] = 1.$$

$B(\Theta) = \mathrm{E}\{\psi\psi^T\}$

$$= \mathrm{E}\begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix},$$

where

i. $B_{11} = \mathrm{E}(X_{ji} - \kappa_j)^2 = \mathrm{var}(X_{ji}) = \kappa_j(1-\kappa_j)$.

ii. $B_{12} = B_{21} = \mathrm{E}\{(X_{ji} - \kappa_j)X_{ji}(1-R_i)(Z_k - Q_k)\} = 0$.

iii. $B_{13} = B_{31} = \mathrm{E}\left[(X_{ji} - \kappa_j)\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_k}{Q_k}\}(Y_i - \mu_{jk})\right] = 0$, since $\mathrm{E}Y_i = \mu_{jk}$.

iv. $B_{22} = \mathrm{E}\{X_j(1-R_i)(Z_k - Q_k)^2\}$

$$= E\left[\{X_{ji}(1-R_i)Z_{ki} - X_{ji}(1-R_i)Q_k\}^2\right]$$

$$= E\left[X_j(1-R_i)Z_k + X_{ji}(1-R_i)Q_k^2 - 2X_{ji}(1-R_i)Q_k\right]$$
$$= E\left[E\left[X_{ji}(1-R_i)Z_k + X_{ji}(1-R_i)Q_k^2 - 2X_{ji}(1-R_i)Q_k\right]|X_{ji}, R_i\right]$$
$$= E\left[E\left[X_{ji}(1-R_i)Q_k + X_{ji}(1-R_i)Q_k^2 - 2X_{ji}(1-R_i)Q_k\right]|X_{ji}, R_i\right]$$
$$= E\left[E\left[X_{ji}(1-R_i)Q_k^2 - X_{ji}(1-R_i)Q_k^2\right]|X_{ji}, R_i\right]$$
$$= E\left[X_{ji}E\left[(Q_k^2 - Q_k)X_{ji}(1-\pi_j)\right]|X_{ji}\right]$$
$$= \kappa_j(1-\pi_j)(Q_k^2 - Q_k).$$

v. $B_{23} = B_{32} = \mathrm{E}\left[X_{ji}(1-R_i)(Z_{ki} - Q_k)\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\}(Y_i - \mu_{jk})\right] = 0$, since $\mathrm{E}Y_i = \mu_{jk}$.

vi. $B_{33} = \mathrm{E}\left[\frac{X_{ji}}{\kappa_j}\{R_i + \frac{(1-R_i)Z_k}{Q_k}\}(Y_i - \mu_{jk})\right]^2$.

$$= \frac{\pi_j}{\kappa_j}\left[\sigma^2_{A_j} + (\mu_{jk} - \mu_{A_j})^2\right] + \frac{(1-\pi_j)}{\kappa_j Q_k}\left[\sigma^2_{A_jB_k} + (\mu_{jk} - \mu_{A_jB_k})^2\right],$$

which is derived above for NIPW estimator.

The variance of NIPW1 estimator is the (3, 3) entry of the matrix $\frac{1}{n}\{A^{-1}(\Theta)B(\Theta)\{A^{-1}(\Theta)\}^T\}$.

5. **Variance of IPW1 Estimator**

Let $\Theta_0 = (\kappa_j, Q_k, \mu_{jk})$ and the ψ- function

$$\psi(Y_i, \Theta_0) = \begin{bmatrix} X_{ji} - \kappa_j \\ X_{ji}(1-R_i)(Z_k - Q_k) \\ \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_k}{Q_k}\right]Y_i - \mu_{jk} \end{bmatrix},$$

where

$$-\psi'(Y_i, \Theta_0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1-R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_k}{Q_k}\right]Y_i & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_k}{Q_k^2}\right]Y_i & 1 \end{bmatrix}.$$

Then $A(\Theta) = E\{-\psi(Y_i, \Theta_0)\}$

$$= E\begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1-R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_k}{Q_k^2}\right]Y_i & 1 \end{bmatrix},$$

where

i. $E\{X_{ji}(1-R_i)\} = E\left[E\{X_{ji}(1-R_i)\}|X_{ji}\right] = E\left[X_{ji}E\{(1-R_i)\}|X_{ji}\right] = E\{X_{ji}(1-\pi_j)\} = \kappa_j(1-\pi_j)$.

ii. $E\left\{\frac{X_{ji}}{\kappa_j^2}\left[R_i+\frac{(1-R_i)Z_k}{Q_k}\right]Y_i\right\}$

$$= E\left[E\frac{X_{ji}}{\kappa_j^2}\left[R_i+\frac{(1-R_i)Z_k}{Q_k}\right]Y_i|X_{ji},R_i,Y_i(A_j),Y_i(A_jB_k)\right]$$

$$= E\frac{X_{ji}}{\kappa_j^2}\left[E\left[R_iY_i(A_j)+\frac{(1-R_i)Z_kY_i(A_jB_k)}{Q_k}\right]Y_i|X_{ji},R_i,Y_i(A_j),Y_i(A_jB_k)\right]$$

$$= E\frac{X_{ji}}{\kappa_j^2}\left[E\left[R_iY_i(A_j)+(1-R_i)Y_i(A_jB_k)\right]|X_{ji},Y_i(A_j),Y_i(A_jB_k)\right]$$

$$= E\frac{X_{ji}}{\kappa_j^2}\left[E\left[\pi_jY_i(A_j)+(1-\pi_j)Y_i(A_jB_k)\right]|X_{ji},Y_i(A_j),Y_i(A_jB_k)\right]$$

$$= E\frac{X_{ji}}{\kappa_j^2}\left[E\left[\pi_j\mu_{A_j}+(1-\pi_j)\mu_{A_jB_k}\right]|X_{ji}\right]$$

$$= \frac{1}{\kappa_j}\left[\pi_j\mu_{A_j}+(1-\pi_j)\mu_{A_jB_k}\right]=\frac{1}{\kappa_j}\mu_{jk}.$$

iii. $E\left\{\frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_k}{Q_k^2}\right]Y_i\right\}$

$$= E\left[E\left\{\frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_k}{Q_k^2}\right]Y_i|X_{ji},R_i\right\}\right]$$

$$= E\left\{\frac{X_{ji}}{\kappa_j}E\left[\frac{(1-R_i)Z_k}{Q_k^2}\right]|X_{ji},R_i\right\}E\left\{E\left[Y_i|R_i\right]\right\}$$

$$= E\left\{\frac{X_{ji}}{\kappa_j}E\left[\frac{(1-R_i)}{Q_k}\right]|X_{ji}\right\}\mu_{jk}$$

$$= E\left\{\frac{X_{ji}}{\kappa_j}\frac{(1-\pi_j)}{Q_k}\right\}\mu_{jk}$$

$$= \frac{(1-\pi_j)}{Q_k}\mu_{jk}.$$

$B(\Theta)=\mathrm{E}\{\psi\psi^T\}$

$$= \mathrm{E}\begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix},$$

where

i. $B_{11}=\mathrm{E}(X_{ji}-\kappa_j)^2=\mathrm{var}(X_{ji})=\kappa_j(1-\kappa_j)$.

ii. $B_{12}=B_{21}=\mathrm{E}\{(X_{ji}-\kappa_j)X_{ji}(1-R_i)(Z_k-Q_k)\}=0$.

iii. $B_{13} = B_{31} = \mathrm{E}(X_{ji} - \kappa_j)\left[\frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i - \mu_{jk}\right]$

$$= E\left[(X_{ji} - \kappa_j)\frac{X_{ji}}{\kappa_j}E\left[\{R_iY_i(A_j) + \frac{(1-R_i)Z_kY_i(A_jB_k)}{Q_k}\} - \mu_{jk}\right]|X_{ji}, R_i, Y_i(A_j), Y_i(A_jB_k)\right]$$

$$= E\left[(X_{ji} - \kappa_j)\frac{X_{ji}}{\kappa_j}E\left[\{R_iY_i(A_j) + (1-R_i)Y_i(A_jB_k)\} - \mu_{jk}\right]|X_{ji}, Y_i(A_j), Y_i(A_jB_k)\right]$$

$$= E\left[(X_{ji} - \kappa_j)\frac{X_{ji}}{\kappa_j}\mu_{jk} - \mu_{jk}\right]$$

$$= \frac{\mu_{jk}}{\kappa_j}E\left[(X_{ji} - \kappa_j)^2\right]$$

$$= \frac{\mu_{jk}}{\kappa_j}\kappa_j(1-\kappa_j) = (1-\kappa_j)\mu_{jk}.$$

iv. $B_{22} = \mathrm{E}\{X_i(1-R_i)(Z_{ki} - Q_k)^2\} = \kappa_j(1-\pi_j)(Q_k^2 - Q_k)$, which is the same as entry B_{22} for NIPW.

v. $B_{23} = B_{32} = \mathrm{E}\left\{X_{ji}(1-R_i)(Z_{ki} - Q_k)\{\frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i - \mu_{jk}\}\right\} = 0.$

vi. $B_{33} = \mathrm{E}\left[\frac{X_{ji}}{\kappa_j}\left\{R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right\}Y_i - \mu_{jk}\right]^2$

$$= \frac{1}{\kappa_j}\{\frac{\pi_j}{Q_k}(\sigma^2_{A_j} + \mu^2_{A_j}) + (1-\pi_j)(\sigma^2_{A_jB_k} + \mu^2_{A_jB_k})\} - \mu^2_{jk},$$

which is computed for IPW above.

The variance of NIPW1 estimator is the (3, 3) entry of the matrix, $\frac{1}{n}\{A^{-1}(\Theta)B(\Theta)\{A^{-1}(\Theta)\}^T\}$,

4.2 VARIANCE ESTIMATORS

1. **<u>IPW Robust Variance Estimator</u>**

$$\hat{\mu}^{IPW}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}Y_i$$

$$= \frac{1}{n}\sum_{i=1}^{n}W_{jki}Y_i$$

where $W_{jki} = \frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}$.

Let the ψ function for $\hat{\Theta} = \hat{\mu}_{A_jB_k}$ be $\psi = \{W_{jki}Y_i - \mu_{jk}\}$; $\psi' = -1$, $\psi\psi^T = \{W_{jki}Y_i - \mu_{jk}\}^2$,

$A(\Theta) = \frac{1}{n}\sum_{i=1}^{n} -\psi' = 1$, $B(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\psi\psi^T = \frac{1}{n}\sum_{i=1}^{n}\{W_{jki}Y_i - \mu_{jk}\}^2$,

$\widehat{\text{var}(\hat{\Theta})}$ is computed as follows

$$\begin{aligned} &= \frac{1}{n}\{A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T\} \\ &= \frac{1}{n}\{\frac{1}{n}\sum_{i=1}^{n}\{W_{jki}Y_i - \hat{\mu}_{jk}\}^2\} \\ &= \frac{1}{n^2}\{\sum_{i=1}^{n}\{W_{jki}Y_i - \hat{\mu}_{jk}\}^2\}. \end{aligned}$$

2. **NIPW Robust Variance Estimator**

$$\hat{\mu}^{NIPW}_{A_jB_k} = \frac{\sum_{i=1}^{n}\frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}Y_i}{\sum_{i=1}^{n}\frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}}$$

$$= \frac{\sum_{i=1}^{n}W_{jki}Y_i}{\sum_{i=1}^{n}W_{jki}}$$

where $W_{jki} = \frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}$.

Let the ψ function for $\hat{\Theta} = \hat{\mu}_{A_jB_k}$ be $\psi = W_{jki}(Y_i - \mu_{jk})$; $\psi' = -W_{jki}$, $\psi\psi^T = \{W_{jki}(Y_i - \mu_{jk})\}^2$, $A(\Theta) = \frac{1}{n}\sum_{i=1}^{n} -\psi' = \frac{1}{n}\sum_{i=1}^{n}W_{jki}$, $B(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\psi\psi^T = \frac{1}{n}\sum_{i=1}^{n}\{W_{jki}(Y_i - \mu_{jk})\}^2$,

$\widehat{\text{var}(\hat{\Theta})}$ is computed as follows,

$$\begin{aligned} &= \frac{1}{n}\{A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T\} \\ &= \frac{n}{\sum_{i=1}^{n}W_{jki}}\{\frac{1}{n^2}\sum_{i=1}^{n}\{W_{jki}(Y_i - \hat{\mu}_{jk})\}^2\}\frac{n}{\sum_{i=1}^{n}W_{jki}} \\ &= \frac{\sum_{i=1}^{n}\{W_{jki}(Y_i - \hat{\mu}_{jk})\}^2\}}{(\sum_{i=1}^{n}W_{jki})^2} \end{aligned}$$

3. **AR Robust Variance Estimator**

The artificial randomized (AR) estimators is given by,

$$\hat{\mu}^{SMART}_{jk} = \frac{\sum_{i=1}^{n}X_{ji}\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\}Y_i}{\sum_{i=1}^{n}X_{ji}\{R_iZ^*_{ki} + (1-R_i)Z_{ki}\}}$$

$$= \frac{\sum_{i=1}^{n}W_{jki}Y_i}{\sum_{i=1}^{n}W_{jki}}$$

where $W_{jki} = X_{ji}\{R_i Z^*_{ki} + (1-R_i)Z_{ki}\}$.

Let the ψ function for $\hat{\Theta} = \hat{\mu}_{A_jB_k}$ be $\psi = W_{jki}(Y_i - \mu_{jk})$, $\psi' = -W_{jki}$, $\psi\psi^T = \{W_{jki}(Y_i - \mu_{jk})\}^2$, $A(\Theta) = \frac{1}{n}\sum_{i=1}^n -\psi' = \frac{1}{n}\sum_{i=1}^n W_{jki}$, $B(\Theta) = \frac{1}{n}\sum_{i=1}^n \psi\psi^T = \frac{1}{n}\sum_{i=1}^n \{W_{jki}(Y_i - \mu_{jk})\}^2$,

$\widehat{\text{var}(\hat{\Theta})}$ is computed as follows

$$\begin{aligned}
&= \frac{1}{n} var(\hat{\Theta}) \\
&= \frac{1}{n}\{A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T\} \\
&= \frac{n}{\sum_{i=1}^n W_{jki}}\{\frac{1}{n^2}\sum_{i=1}^n \{W_{jki}(Y_i - \mu_{jk})\}^2\}\frac{n}{\sum_{i=1}^n W_{jki}} \\
&= \frac{\sum_{i=1}^n \{W_{jki}(Y_i - \mu_{jk})\}^2\}}{(\sum_{i=1}^n W_{jki})^2}
\end{aligned}$$

4. **NIPW1 Robust Variance Estimator**

$$\begin{aligned}
\hat{\mu}^{NIPW}_{A_jB_k} &= \frac{\sum_{i=1}^n \frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}Y_i}{\sum_{i=1}^n \frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}} \\
&= \frac{\sum_{i=1}^n W_{jki}Y_i}{\sum_{i=1}^n W_{jki}}
\end{aligned}$$

where $W_{jki} = \frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}$.

Let the ψ-function for $\Theta = (\kappa_j, Q_k, \mu_{jk})$ be

$$\psi(\Theta, Y_i) = \begin{bmatrix} X_{ji} - \kappa_j \\ X_{ji}(1-R_i)(Z_{ki} - Q_k) \\ \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]\{Y_i - \mu_{jk}\} \end{bmatrix},$$

$$-\psi'(Y_i, \Theta_0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1-R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_{ki}}{Q_k^2}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right] \end{bmatrix},$$

$A(\Theta) = \frac{1}{n}\sum_{i=1}^{n} -\psi(\Theta, Y_i)$

$$= \sum_{i=1}^{n} \begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1-R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_{ki}}{Q_k^2}\right]\{Y_i - \mu_{jk}\} & \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right] \end{bmatrix}.$$

$B(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \psi\psi^T$

$$= \sum_{i=1}^{n} \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix},$$

where

i. $B_{11} = \sum_{i=1}^{n}(X_{ji} - \kappa_j)^2$.

ii. $B_{12} = B_{21} = \sum_{i=1}^{n}(X_{ji} - \kappa_j)X_{ji}(1-R_i)(Z_{ki} - Q_k)$.

iii. $B_{13} = B_{31} = \sum_{i=1}^{n}(X_{ji} - \kappa_j)\frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right](Y_i - \mu_{jk})$.

iv. $B_{22} = \sum_{i=1}^{n} X_i(1-R_i)(Z_{ki} - Q_k)^2$.

v. $B_{23} = B_{32} = \sum_{i=1}^{n} X_{ji}(1-R_i)(Z_{ki} - Q_k)\frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right](Y_i - \mu_{jk})$.

vi. $B_{33} = \sum_{i=1}^{n}\left[\frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right](Y_i - \mu_{jk})\right]^2$.

The robust variance estimator of the variance of NIPW1 estimator is the (3, 3) entry of matrix $\frac{1}{n}\{A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T\}$.

5. **IPW1 Robust Variance Estimator**

$$\hat{\mu}_{jk}^{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}Y_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} W_{jki}Y_i$$

where $W_{jki} = \frac{X_{ji}}{\kappa_j}\{R_i + \frac{1-R_i}{Q_k}Z_{ki}\}$.

Let the ψ-function for $\Theta = (\kappa_j, Q_k, \mu_{jk})$ be

$$\psi(\Theta, Y_i) = \begin{bmatrix} X_{ji} - \kappa_j \\ X_{ji}(1-R_i)(Z_{ki} - Q_k) \\ \frac{X_{ji}}{\kappa_j}\left[R_i + \frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i - \mu_{jk} \end{bmatrix},$$

$$-\psi^{'}(Y_i,\Theta_0)=\begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1-R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i+\frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_{ki}}{Q_k^2}\right]Y_i & 1 \end{bmatrix},$$

$A(\Theta)=\frac{1}{n}\sum_{i=1}^{n}-\psi(\Theta,Y_i)$

$$=\sum_{i=1}^{n}\begin{bmatrix} 1 & 0 & 0 \\ 0 & X_{ji}(1-R_i) & 0 \\ \frac{X_{ji}}{\kappa_j^2}\left[R_i+\frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i & \frac{X_{ji}}{\kappa_j}\left[\frac{(1-R_i)Z_{ki}}{Q_k^2}\right]Y_i & 1 \end{bmatrix}.$$

$B(\Theta)=\frac{1}{n}\sum_{i=1}^{n}\psi\psi^T$

$$=\sum_{i=1}^{n}\begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix},$$

where

i. $B_{11}=\sum_{i=1}^{n}(X_{ji}-\kappa_j)^2$.

ii. $B_{12}=B_{21}=\sum_{i=1}^{n}(X_{ji}-\kappa_j)X_{ji}(1-R_i)(Z_{ki}-Q_k)$.

iii. $B_{13}=B_{31}=\sum_{i=1}^{n}(X_{ji}-\kappa_j)\{\frac{X_{ji}}{\kappa_j}\left[R_i+\frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i-\mu_{jk}\}$.

iv. $B_{22}=\sum_{i=1}^{n}X_i(1-R_i)(Z_{ki}-Q_k)^2$.

v. $B_{23}=B_{32}=\sum_{i=1}^{n}X_{ji}(1-R_i)(Z_{ki}-Q_k)\{\frac{X_{ji}}{\kappa_j}\left[R_i+\frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i-\mu_{jk}\}$.

vi. $B_{33}=\sum_{i=1}^{n}\left[\frac{X_{ji}}{\kappa_j}\left[R_i+\frac{(1-R_i)Z_{ki}}{Q_k}\right]Y_i-\mu_{jk}\right]^2$.

The robust variance estimate of IPW1 is the (3, 3) entry of matrix $\frac{1}{n}\{A^{-1}(\hat{\Theta})B(\hat{\Theta})\{A^{-1}(\hat{\Theta})\}^T\}$.

5.0 DISCUSSION AND FUTURE WORK

Adapting or personalizing treatments due to heterogeneous response profiles is becoming important for treating patients with chronic diseases. Routine sequential dose adjustment, treatment choice based on history of covariates and past treatments are essential features of adaptive treatment strategies. SMAR designs enable construction of pre-specified strategies for future patients.

This chapter presents a synopsis of the dissertation and future work. This dissertation makes two fundamental contributions to the literature of adaptive treatment strategies.

In Chapter 2, we introduced sample size formulas that will be useful when designing two-stage SMAR trials. The formula is based on Wald-type test statistic. An overall sample size formula to detect any difference among all strategy means as well as a sample size formula to detect pairwise differences are provided. Often, the interest is to compare ATSs embedded within SMAR designs. One approach is to conduct hypothesis testing to compare the ATS means. This can be achieved using IPW or g-computation techniques to estimate strategy means and compare. However, investigators are interested in using regression models to estimate ATS means, compare among them and make pertinent inferences by allowing covariate adjustments. To specifically deal with this, in Chapter 3, we introduced the concept of 'artificial randomization'. Artificial randomization puts subjects into unique classes which then makes regression possible. Due to its sequential nature, data from SMAR designs are manipulated before it becomes available for regression methods. To elaborate implementation of the method we used the CALGB data to perform a data analysis. One of the key benefits of artificial randomization is that using the initial regression results, we are able to create more strategies that could inform disease management.

We have proposed an unbiased estimator called simple multiple artificial randomized tool (SMART) which is easier to implement. SMART specifies artificial randomization to create M=1 to M=5 data sets to create a single estimate and standard error. It is worth investigating properties of the estimator for higher values of M by changing assumed population parameters. Softwares such as R or SAS can be used to implement the regression as the data that is artificially randomized becomes readily available. There is a challenge to extend the methodology to SMAR designs with more than two stages. It could easily be adapted to binary, survival outcomes. Our method is limited to two-stage, specific designs where either first stage responders or non-responders do not get re-randomized.

When analyzing data from SMAR designs, it is of interest to test equality of adaptive treatment strategies (ATSs). In the usual setting, hypothesis testing is carried out parametrically using test statistics developed for IPW or g-estimation. Non-parametric tests are often attractive for small sample sizes. For future work, permutation tests could be considered.

BIBLIOGRAPHY

1. Boos, D. D., Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods.* New York: Springer.

2. Bembom, O. and van der Laan, M. J. (2007). Statistical methods for analyzing sequentially randomized trials. *Journal of the National Cancer Institute* **99(21),** 1577–1582.

3. Dawson, R. and Lavori, P. W. (2010). Sample Size calculations for Evaluating Treatment Policies in Multi-Stage Design. *Clin. Trials* **7,** 643–652. DOI: 10.1177/1740774510376418

4. Dawson, R. and Lavori, P. W. (2012). Efficient design and inference for multistage randomized trials of individualized treatment policies. *Biostatistics* **13(1),** 142–152.

5. Dwyer, J. H. (1974). Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin* **81(10),** 731–737.

6. Feng, W. and Wahed, A. S. (2008). A supremum log rank test for comparing adaptive treatment strategies and corresponding sample size formula. *Biometrika* **95(3),** 695–707.

7. Feng, W. and Wahed, A. S. (2009). Sample Size for Two-Stage Studies with Maintenance Therapy. *Statistics in Medicine* **28,** 2028–2041.

8. Hernan M. A., Lanoy E., Costagliola D., Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin. Pharmacol. Toxicol.* **98,** 237–242.

9. Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47(260),** 663–685.

10. Kidwell, K. W. and Wahed, A. S. (2013). Weighted log-rank statistic to compare shared-path adaptive treatment strategies. *Biostatistics* **14(2),** 299–312.

11. Ko, J. H. and Wahed, A. S. (2012). Up-front vs. Sequential Randomizations for Inference on Adaptive Treatment Strategies. *Statistics in Medicine* **31(9),** 812–830.

12. Lavori P. W., Dawson R, Rush A. J. (2000). Flexible treatment strategies in chronic disease: clinical and research implications. *Biol. Psychiatry* **48,** 605–14.

13. Lavori, P. W. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clin. Trials* **1,** 9–20.

14. Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. A. (2012). A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology* **8,** 14.1–14.28. DOI: 10.1146/annurev-clinpsy-032511-143152

15. Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. A. (1999). Randomized phase II study of fludarabine+ cytosine arabinoside+ idarubicinall-trans retinoic acidgranulocyte colony-stimulating factor in poor prognosis newly diagnosed acute myeloid leukemia and myelodysplastic syndrome. *Blood* **93(8),** 2478–2484.

16. Li, Z. and Murphy, S. A. (2011). Sample size formulae for two-stage randomized trials with survival outcomes. *Biometrika* **98(3),** 503–518.

17. Lokhnygina Y. and Helterbrand, J. D. (2007). Cox regression methods for two-stage randomization designs. *Biometrics* **63,** 422–428.

18. Lunceford, J. K., Davidian, M., and Tsiatis, A. A. (2002). Estimation of Survival Distributions of Treatment Policies in Two-Stage Randomization Designs in Clinical Trials. *Biometrics* **58(1),** 48–57.

19. Miyahara, S. and Wahed, A. S. (2010). Weighted Kaplan-Meier estimators for two-stage treatment regimes. *Statistics in Medicine* **29,** 2581–2591.

20. Moodie, E. E. M., and Richardson, T. S., Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63(2),** 447–455.

21. Murphy, S.A., Van Der Laan, M. J., Robins, J. M. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* **96(456),** 1410–1423.

22. Murphy, S. A. (2005). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine* **24,** 1455–1481. DOI: 10.1002/sim.2022

23. Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and alcohol dependence* **88,** S24–S30.

24. Nahum-Shani, I., Qian, M., Oslin, D., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., Murphy, S. A. (2007). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological methods* **88,** 457.

25. Oetting, A. I., Levy, J. A., Weiss, R. D., Murphy, S. A. (2011). Statistical methodology for a SMART design in the development of adaptive treatment strategies. In Shrout, P. E., editor. Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (pp.179-205). American Psychiatric Publishing, Arlington, VA.

26. Pelham, W. E. and Fabiano, G. A. (2008). Evidence-based psychosocial treatments for attention deficit/hyperactivity disorder. *J. Clin. Child Adolesc. Psychol.* **37,** 184–214.

27. Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period applications to control of the healthy workers survivor effect. *Mathematical Modeling* **7,** 1393–1512. DOI: 10.1016/0270-0255(86)90088-6

28. Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Nicholas P. Jewell, Klaus Dietz, and Vernon T. Farewell, editors. AIDS Epidemiology-Methodological Issues, (pages 297-331). Birkhauser, Boston, 1992.

29. Robins, J. M. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor. Latent Variable Modeling and Applications to Causality (pp.69-117). Springer, New York, NY.

30. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66(5),** 688-701.

31. Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72,** 538-543.

32. Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley: New York, NY.

33. Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., et al (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin. Trials* **25,** 119–142.

34. Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research* **8(1),** 3-15.

35. Schneider, L. S., Tariot, P. N., Lyketsos, C. G., Dagerman, K. S., Davis, S. M., Hsiao, J. K., Ismail, M. S., Lebowitz, B. D., Lyketsos, C. G., Ryan, J. M., Stroup, T. S., Sultzer, D. L., Weintraub, D, and Lieberman, J. A. (2001). National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE), Alzheimer disease trial methodology. *American Journal of Geriatric Psychiatry* **9(4),** 346–360.

36. Stone, R. M., Berg, D. T., George, S. L., Dodge, R. K., Paciucci, P. A., Schulman, P., Lee, E. J., Moore, J. O., Powell, B. L., and Schiffer, C. A. (1995). Granulocyte-macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. *New England Journal of Medicine* **332(25),** 1671–1677.

37. Tang, X. and Wahed, A. S. (2011). Comparison of treatment regimes with adjustment for auxiliary variables. *Journal of Applied Statistics* **38(12),** 2925–2938.

38. Tang, X. and Wahed, A. S. (2013). Cumulative Hazard Ratio Estimation for Treatment Regimes in Sequentially Randomized Clinical Trials. *Statistics in Biosciences* 1–18.

39. Thall, P. F., Wooten, L. K, Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine* **26,** 4687–4702.

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

41. Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **54(3),** 426–482.

42. Wallace, M. P. and Moodie, E. E. M. (2014). Personalizing medicine: a review of adaptive treatment strategies. *Pharmacoepidemiology and drug safety* **23(6),** 580–585.