SURVIVAL ANALYSIS OF SHARED-PATH ADAPTIVE TREATMENT STRATEGIES

by

Kelley M. Kidwell

B.S. in Mathematics, Bucknell University, 2007

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Kelley M. Kidwell

It was defended on

July 18, 2012

and approved by

Kaleab Abebe, PhD
Assistant Professor
School of Medicine
University of Pittsburgh

Joseph Costantino, DrPH
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Sally Morton, PhD
Professor and Chair
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Dissertation Advisor
Abdus S. Wahed, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

SURVIVAL ANALYSIS OF SHARED-PATH ADAPTIVE TREATMENT STRATEGIES

Kelley M. Kidwell, PhD

University of Pittsburgh, 2012

Adaptive treatment strategies closely mimic the reality of a physician's prescription process where the physician prescribes a medication to his/her patient and based on that patient's response to the medication, modifies the treatment. Two-stage randomization designs, more generally, sequential multiple assignment randomization trial (SMART) designs, are useful to assess adaptive treatment strategies where the interest is in comparing the entire sequence of treatments, including the patient's intermediate response. In this dissertation, we introduce the notion of shared-path and separate-path adaptive treatment strategies and propose weighted log-rank statistics to compare overall survival distributions of two shared-path or multiple two-stage adaptive treatment strategies. Large sample properties of the statistics are derived and the type I error rate and power of the tests are compared to standard statistics through simulation. We also propose a sample size equation to power a two-stage SMART comparing the overall survival of multiple adaptive treatment strategies.

Public health significance: The treatment of many diseases and illnesses, especially those which are chronic (cancer, AIDS, depression, substance abuse, ADHD), includes sequences of treatments based on the individual's characteristics, behaviors, and responses. Treatment is inherently dynamic, but often, clinical trials are not designed or analyzed to take this dynamic feature into account. We present methods to adequately power and analyze clinical trials with time-to-event data which aim to compare these individualized sequences of treatments or adaptive treatment strategies. Through these methods and by comparing adaptive treatment strategies, patient outcomes can be operationalized and improved over time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Adaptive treatment strategies are at the forefront of clinical trials and statistical methodology literature. As this area develops, new terminology and concepts abound. This chapter introduces the central concepts which provide the backdrop for this dissertation. We will introduce the topics listed below and culminate with the motivation and objectives of this work. If you are familiar with the following concepts, please feel free to jump to Section 1.5.

- Adaptive treatment strategies (ATS)
- Sequential multiple assignment randomized trial (SMART)
- Counterfactual framework
- Naive methods of analysis

1.1 ADAPTIVE TREATMENT STRATEGIES (ATS)

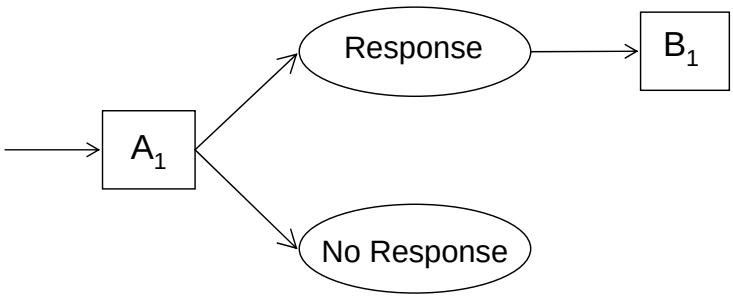
Adaptive treatment strategies (or dynamic treatment regimes), notably introduced and described in statistical literature by Lavori et al. (2000), Lavori and Dawson (2000), Lavori et al. (2004), Murphy and McKay (2004) and Collins et al. (2004), among others, consist of a sequence of treatments and decision rules that guide the choice of these treatments for a given individual. These strategies are dynamic/adaptive since the strategies allow personalizing/adapting treatments based on the patient's response to earlier treatments. They, therefore, closely mimic a physician's prescription process. Adaptive treatment strategies are currently a driving force in the biomedical field updating the way treatment and prevention trials are designed, implemented, and analyzed.

With increasing scientific knowledge and technology, as well as people's increasing lifespan, many illnesses and diseases require more than just one therapy during the course of treatment. Cancer usually requires a maintenance therapy in order to illicit the desired result, while AIDS, depression and substance abuse often require multiple treatments which are taken sequentially to enhance their effects, or require treatment routines which differ substantially between people and depend on personal intermediate outcomes. Thus, adaptive treatment strategies are especially relevant in clinical trials for chronic diseases. The objective in a clinical trial utilizing adaptive treatment strategies is to develop multi-stage decision-making strategies that improve patient outcomes over time. Note that the term 'adaptive' here refers to time-varying sequences of treatments for a given patient which are presumably chosen based on that patient's status (characteristics, response), and not therapy for the present patient which depends on past patients or where design parameters are altered mid-trial.

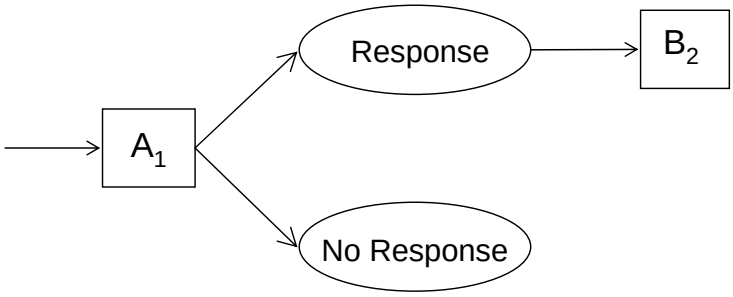
An adaptive treatment strategy generally begins with an initial treatment, followed by intermediate outcomes which are input into decision rules that then define the subsequent stages of treatment. The decision rules take patient characteristics and outcomes, such as patient history, health risk factors, patient response, and patient adherence, into account to provide a personalized therapy sequence. Explicitly, let us define a two-stage adaptive treatment strategy A_jB_k to be 'Treat initially with A_j, and then by second-line treatment B_k if the patient is eligible and consents to subsequent second-line therapy.' The eligibility of the patient would depend on the decision rule. An adaptive treatment strategy could consist of several stages with several treatment options at each stage.

We further define a more specific, two-stage adaptive treatment strategy with two treatment options at each stage. Here, patients are first treated with either A_1 or A_2, and patients who respond to the initial treatment and consent to further treatment, can receive either B_1 or B_2. Therefore, in this setting, we can have the set of strategies $\{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$, where, strategy A_1B_1, for example, dictates that the patient be treated with therapy A_1 and upon response and consent, be treated with B_1 (Figure 1.1). Notice from the definition of adaptive treatment strategies, as illustrated in Figure 1.1, that patients who do not respond to therapy A_1 are consistent with both strategies A_1B_1 and A_1B_2. Thus, a patient following

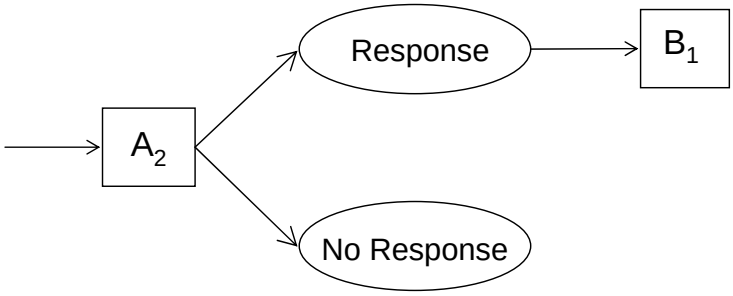
(a) Strategy A_1B_1



(b) Strategy A_1B_2



(c) Strategy A_2B_1



(d) Strategy A_2B_2

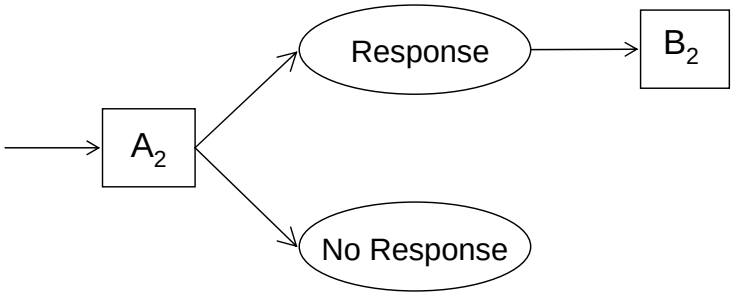


Figure 1.1: Example of two-stage adaptive treatment strategies

ATS A_1B_1 who did not respond to A_1 shares the same treatment path as a patient following A_1B_2, but did not respond to A_1. Similarly, patients who do not respond to therapy A_2 are consistent with both strategies A_2B_1 and A_2B_2. On the other hand, patients who follow A_1B_1 or A_1B_2 are not consistent with either strategies A_2B_1 or A_2B_2. This distinction motivates the classification of adaptive treatment strategies into strategies which do or do not share treatment paths. Sets of strategies such as $\{A_1B_1, A_1B_2\}$ and $\{A_2B_1, A_2B_2\}$ are shared-path adaptive treatment strategies. Two two-stage adaptive treatment strategies are shared-path if individuals treated with one strategy share a common path of treatment with individuals treated by the other strategy. Strategies which do not share a common path of treatment will be referred to as separate-path adaptive treatment strategies. For explicit definitions, please refer to Section 2.2.

1.2 SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIAL (SMART)

To develop decision rules and inform the construction of adaptive treatment strategies, Murphy (2005) introduced the concept of the sequential multiple assignment randomized trial (SMART), which had previously been studied under different terminology in Lavori and Dawson (2000), Lavori et al. (2004). In a SMART, individuals may be randomized several times, receiving multiple treatments sequentially over time. The design allows for the assessment and comparison of adaptive treatment strategies. Many trials over the past decade have assessed adaptive treatment strategies using the SMART design including Stone et al. (2001), Stroup et al. (2003), Rush et al. (2004), Winter et al. (2006), Marlowe et al. (2007) and Matthay et al. (2009).

SMART designs may include any number of randomization stages, but this dissertation will focus on two-stage randomization designs. Figure 1.2 depicts a two-stage SMART design where patients are initially randomized to treatments A_j, $j = 1, \ldots, J$, and then depending on response to these initial treatments, are randomized to treatments B_k, $k = 1, \ldots, K$, for responders, and B'_l, $l = 1, \ldots, L$ for non-responders. The investigator may be interested

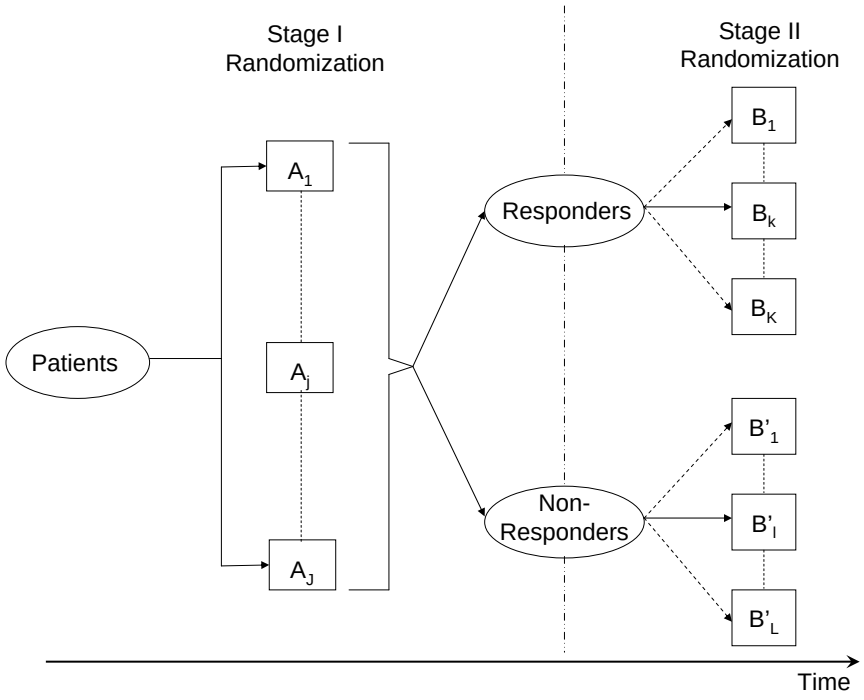


Figure 1.2: A general two-stage SMART design

in comparing adaptive treatment strategies $A_jB_kB'_l$ for any $j = 1, \ldots, J$, $k = 1, \ldots, K$, $l = 1, \ldots, L$, where $A_jB_kB'_l$ stands for 'Treat with A_j followed by B_k if the patient responds, or by B'_l if the patient does not'. Figure 1.2 illustrates a general two-stage design where every patient receives some therapy at each stage, but there may be cases where the therapy is stopped. For example, in older, terminally ill, leukemia patients, if the initial chemotherapy fails to achieve remission, usually no further therapy is given. In Chapters 2 and 3, the two-stage design is set up such that the non-responders are not randomized in the second stage, eliminating the branches involving B'_l, $l = 1, 2, \ldots, L$. Methods introduced in this dissertation, however, can adapt to other two-stage designs by re-defining response or altering the weighting equations presented in Section 2.6. When there are more than two randomization stages, intermediate outcomes dictate subsequent treatment randomization and the sequence of treatments continues.

The SMART design allows for the assessment and comparison of adaptive treatment strategies in an efficient manner. If strategies were predetermined and patients randomized upfront to all possible strategies, the trial would require a large number of patients, even

with few stages and few treatments at each stage. The SMART design allows for only the patients who meet predetermined criteria to be randomized into further treatment, specifying the adaptive treatment strategies. SMART designs still often require a larger number of subjects than most clinical trials comparing only two or more treatments at one stage, but these designs allow us to mimic the process of prescription and answer different questions about sequences of treatment to find the optimal sequence based on individual's response and/or characteristics.

1.3 COUNTERFACTUAL FRAMEWORK

To quantify the treatment effects of adaptive treatment strategies and construct estimands of interest from a population, the counterfactual (or potential) outcomes framework (Rubin, 1974, Holland, 1986) is often very useful. We introduce this concept by presenting a simple example. Suppose you had a headache. There are many different treatment options for a headache, but suppose this time, you take Excedrin. On this occasion, your outcome is that your headache disappears within 30 minutes. But, what would the outcome have been if you had taken Tylenol or Advil or perhaps did nothing? We only observed the outcome from Excedrin, but you would have had an outcome for each of those treatments, had you taken them and you would have had an outcome for any possible treatment. All of the possible outcomes, including those that we can and can not observe, establish the set of your counterfactual outcomes. If we knew every outcome from every treatment, observed and unobserved, we would compare each of them to each other and report back the optimal treatment to eliminate your headache as quickly as possible.

The concept has been simplified here, but it can be extended to the realm of adaptive treatment strategies. Patients following an ATS only receive one sequence of treatments based on their characteristics, behaviors, or response and randomization probability and we observe only one outcome. But, the patient could have received a different set of treatments and had a different outcome for each set. Throughout this dissertation, the counterfactual outcomes of interest are potential survival times and these outcomes will help us identify the

variables whose distributions are compared across treatment strategies. We will relate the set of counterfactual survival times to the observed survival time.

In order to operate within the counterfactual framework and make valid inference, we need to verify three assumptions (Rubin, 1974, Robins et al., 1994). The assumptions include the consistency assumption (or SUTVA, stable unit treatment value assumption) which connects the counterfactuals to the observed data, the sequential randomization assumption (a.k.a. no unmeasured confounding or conditional exchangeability) which requires the values of the counterfactuals to be missing at random given data on baseline covariates, and positivity (or experimental treatment assumption) which insists the probability of being assigned to each possible treatment is greater than zero. These assumptions are elaborated on below in a simplified setting.

To further explain how counterfactual outcomes help develop estimands while assessing treatment effects, suppose there are two competing treatments A and B whose effect on an outcome Y is of interest. Suppose Y_i^A and Y_i^B denote the counterfactual outcomes for treatments A and B for patient i. One would like to draw inference about the counterfactual mean difference $E(Y_i^A - Y_i^B)$. For a given patient i, both Y_i^A and Y_i^B are not observed, so we cannot estimate the average treatment difference by the marginal mean of the difference, $\frac{\sum_{i=1}^N \{Y_i^A - Y_i^B\}}{N}$. If we satisfy the assumptions above, however, we can estimate the average treatment difference from the observed data.

Under the consistency assumption and letting X_i denote treatment received, the observed outcome values can be defined as $Y_i = I(X_i = A)Y_i^A + I(X_i = B)Y_i^B$. Then for an individual who received treatment A, $Y_i = Y_i^A$ and likewise, for an individual who received treatment B, $Y_i = Y_i^B$. The second assumption, sequential randomization, states that the probability of receiving treatment A or B does not depend on the counterfactual outcomes, Y_i^A or Y_i^B, given baseline predictors. This is guaranteed in conditionally randomized studies. In observational studies, however, investigators need to collect as many predictors as feasibly possible to approximately satisfy this assumption.

Under these two assumptions, $E(Y_i^A - Y_i^B) = E(\bar{Y}^A - \bar{Y}^B)$, where we define $\bar{Y}^A = \sum_{i=1}^N I(X_i = A)Y_i / \sum_{i=1}^N I(X_i = A)$ and $\bar{Y}^B = \sum_{i=1}^N I(X_i = B)Y_i / \sum_{i=1}^N I(X_i = B)$. Thus, when these two assumptions are satisfied, $E(Y_i^A - Y_i^B)$ can be unbiasedly estimated by

$\bar{Y}^A - \bar{Y}^B$. Finally, we assume positivity holds such that investigators have assigned subjects into all the treatments of interest, here A and B. In the adaptive treatment strategies survival analysis setting, counterfactual outcomes will be developed in Section 2.3 and the assumptions will be verified in Section 2.4.

1.4 NAIVE METHODS OF ANALYSIS AND RELATED WORK

As the use of adaptive treatment strategies gains momentum in clinical and behavioral research, statisticians bear the responsibility to produce methodology that analyzes the findings accurately and efficiently. Prior to the invention of the term 'adaptive treatment strategies', survival data from SMART designs had been analyzed separately for each stage ignoring past or future treatment phases or analyzed conditionally on previous stages (for example, only analyzing the outcomes from the responders of the initial treatment). Both of these methods ignore the sequence of treatments and answer a different question than that of comparing adaptive treatment strategies. Conditional analysis also ignores information from another, potentially large, set of patients (for example, the non-responders to the initial treatment). In attempt to account for all of the patients, another method groups patients into the strategy which they followed and then compares the survival outcomes by group using the standard (unweighted) log-rank statistic. This method ignores the second-stage (or further stage) randomization(s) and assumes independence between the groups following each strategy. As we have seen in Section 1.1, this independence is not guaranteed since some groups of individuals share the same treatment paths. For example, referring to Figure 1.1 and assuming the two-stage strategy A_jB_k, $j, k = 1, 2$, which dictates 'Treat with A_j, $j = 1, 2$, followed by B_k, $k = 1, 2$, if the patient responds and is eligible,' we see that if we grouped all those who follow A_1B_1 and all those who follow A_1B_2, the set of non-responders to A_1 would be in both groups. This violates the independence assumption of the standard (unweighted) log-rank statistic.

In more recent years, methods have been developed to take into account both the sequence of treatments and the two-stage randomization in survival analysis of adaptive treat-

ment strategies. Guo (2005) proposed an inverse-probability weighted version of the log-rank test for comparing two adaptive treatment strategies in his unpublished PhD thesis. He developed weights to account for the second-stage randomization which we use and discuss in detail in Section 2.6. Li and Murphy (2011) formally presented the weighted log-rank statistic proposed in Guo (2005) with both time-dependent and time-independent weights. Lokhnygina and Helterbrand (2007) proposed a weighted version of the score equation and score test of the Cox proportional hazards model. Feng and Wahed (2008) utilized the weights developed by Guo and lessened the assumptions made by Lokhnygina and Helterbrand to present a supremum weighted log-rank statistic. Each of these methods, however, were created only to compare two separate-path adaptive treatment strategies. Li and Murphy (2011) presented a weighted Kaplan-Meier estimator to compare two shared-path adaptive treatment strategies, but only to compare point-wise survival estimates for two strategies. Thus, to compare overall survival between shared-path strategies, one could use a weighted log-rank statistic similar to that from Guo's thesis, assuming independence. This independent weighted log-rank statistic ignores the covariance between the groups which share treatment paths in its variance calculation. Thus, the naive methods of analysis of adaptive treatment strategies either ignore second-stage randomization or assume independence between groups. We consider both of these issues in the development of a weighted log-rank statistic to compare two or more shared-path adaptive treatment strategies.

1.5 MOTIVATION AND OBJECTIVES

Medicine has been practiced in a personalized manner throughout history, but only in the past decade has this concept been explored and enhanced in both medical and statistical literature. Developing innovative trial designs to compare adaptive treatment strategies will increase their efficiency and practicality, allowing for a more realistic and personalized treatment process. Generating the appropriate statistical techniques for analysis will allow us to benefit from this improved efficacy and effectiveness. This promising area offers many openings to develop and strengthen statistical methods. Specifically, the distinction between

shared-path and separate-path adaptive treatment strategies and the lack of methods to analyze two shared-path or multiple strategies motivated the methodology in this dissertation.

This dissertation aims to compare the overall survival distributions from two or more two-stage adaptive treatment strategies which may share the same treatment paths. The presented weighted log-rank tests comparing the overall survival distribution of two shared-path or multiple adaptive treatment strategies include covariance terms to account for patients who are consistent with more than one treatment strategy. These tests, therefore, are more efficient than naive methods for comparing shared-path adaptive treatment strategies (Section 1.4).

To facilitate the implementation of adaptive treatment strategies in clinical trials, we also present a sample size equation. This equation will allow statisticians and clinicians to adequately power a SMART when the interest is in comparing survival distributions of multiple adaptive treatment strategies, some of which may be shared-path. The sample size formula asks for inputs that the physician could easily approximate given prior knowledge or pilot studies. Using this information and given sample size, clinics can design and implement SMARTs, leading to the construction of effective adaptive treatment strategies.

2.0 LOG-RANK STATISTICS TO COMPARE TWO SHARED-PATH ADAPTIVE TREATMENT STRATEGIES

2.1 INTRODUCTION

Physicians rarely choose treatment for a patient randomly from competing treatments, but rather they prescribe treatments based on their clinical experience in treating patients with similar characteristics and those patients' individual history of response and adverse reactions to prior treatments. Thus, physicians inherently practice personalized medicine, yet many clinical trials continue to compare two or more treatments at specific time points using randomized, independent groups. These randomized controlled trial designs lack the dynamic aspect of assessing patients' intermediate outcomes and possibly modifying therapies in order to elicit a desired response. Sequential multiple assignment randomized trials, SMART, (Murphy, 2005) have been developed to investigate a sequence of time-varying treatments subject to modification based on the individual's response, more alike treatment strategies that are adopted by physicians in practice. The SMART design allows for the assessment and comparison of adaptive treatment strategies (also known as dynamic treatment regimes), which consist of a sequence of individually tailored therapies during the course of treatment. In a SMART design, a patient's intermediate outcome is measured at specific time points whereupon the treatment or its dosage is adjusted accordingly. Biomedical studies, especially clinical trials for chronic diseases such as cancer, AIDS, depression, and substance abuse, are utilizing the SMART design to reach conclusions about personalized adaptive treatment strategies.

To better illustrate the emerging paradigm of adaptive treatment strategies, consider the following examples for treating moderate depression. One adaptive strategy for moderate

depression treatment is, "First treat the patient with Sertraline for 8 weeks, if the patient does not respond (Beck Depression Inventory, BDI, score over 12), treat the patient with Sertraline as well as with cognitive behavioral therapy (CBT); if the patient responds (BDI score of 12 or under), continue Sertraline." Similarly, other adaptive strategies could be considered where alternative treatment options are prescribed at one or more stages. Another example of an adaptive treatment strategy is, "First treat the patient with Escitalopram for 8 weeks, if the patient does not respond, treat the patient additionally with Bupropion; if the patient responds, continue Escitalopram." At the end, one would be interested to compare not just Sertraline to Escitalopram, but rather, the entire sequence of Sertraline alone or Sertraline followed by CBT and Escitalopram alone or Escitalopram followed by the addition of Bupropion. Thus, strategies consisting of initial treatment, intermediate response and maintenance or second-line treatment are compared to find an optimal course of treatment for an individual.

Individualized medicine has been one of the major concentrations of the medical community in recent years and thus, the last decade has brought about a surge in the application of SMART designs for comparing adaptive strategies in clinical and behavioral research (Stone et al., 1995, 2001, Stroup et al., 2003, Rush et al., 2004, Winter et al., 2006, Marlowe et al., 2007, Matthay et al., 2009), although not all of these studies had comparisons of adaptive strategies as their main aim. As a consequence of the increased use of SMART designs, statistical literature experienced a similar surge in the development of statistical methods for analyzing data arising from such trials (Thall et al., 2000, Murphy, 2003, 2005, Dawson and Lavori, 2004, Wahed and Tsiatis, 2004, Wahed, 2010, Orellana et al., 2010). This dissertation focuses on time-to-event outcome data and hence the review of literature will emphasize statistical methods for survival analysis in SMART designs.

Prior to the invention of the terms 'adaptive treatment strategies' or 'dynamic treatment regimes' survival data from SMART designs had been analyzed separately for each stage ignoring past or future treatment phases. Lunceford et al. (2002) first showed how to estimate point-wise survival probabilities or overall mean survival for adaptive treatment strategies arising from two-stage SMART designs. Methods proposed therein basically used marginal models employing inverse-probability-of-treatment-weighting for estimation. Their analy-

sis, while improving upon stage-specific analysis, was not applicable for comparing overall survival curves under different treatment strategies.

The first valid attempt in developing a test comparing overall survival curves under two adaptive treatment strategies was taken by Guo in his 2005 dissertation. He provided an inverse-weighted version of the log-rank test for comparing two separate-path adaptive treatment strategies (strategies that do not share the same treatment paths, see Section 2.2). Lokhnygina and Helterbrand (2007) extended the idea of Lunceford et al. (2002) to the Cox proportional hazards model and proposed a weighted version of the score equation and score test to compare induction strategies for a fixed second-stage treatment. Generalizing the proportional hazards assumption and creating a more robust statistic, Feng and Wahed (2008) utilized the inverse-probability-of-treatment-weighting method developed in Guo (2005) to present a supremum weighted log-rank statistic, but again only to compare two separate-path adaptive strategies.

Comparison of shared-path adaptive treatment strategies is challenging since the correlation between survival curves needs to be accounted for in the estimation process. Accounting for this correlation, for example, allows us to compare treatment strategies that share the same initial treatment. The goal of this chapter is to present methods for comparing two shared-path adaptive treatment strategies (strategies that share some of the same treatment paths, see Section 2.2).

2.2 DEFINITIONS

Consider a two-stage SMART design in which patients are first randomized to receive treatment A, level A_1 or A_2, and those who respond to the initial treatment A and consent to another randomization, receive maintenance treatment B, randomly allocated to the levels B_1 or B_2 (see Figure 2.1). For simplicity, we will use response to indicate 'response to the previous treatment and consent to the following treatment'. We are interested in the outcomes of patients who follow the various treatment strategies A_jB_k, $j,k = 1,2$, where the strategy A_jB_k is defined as follows.

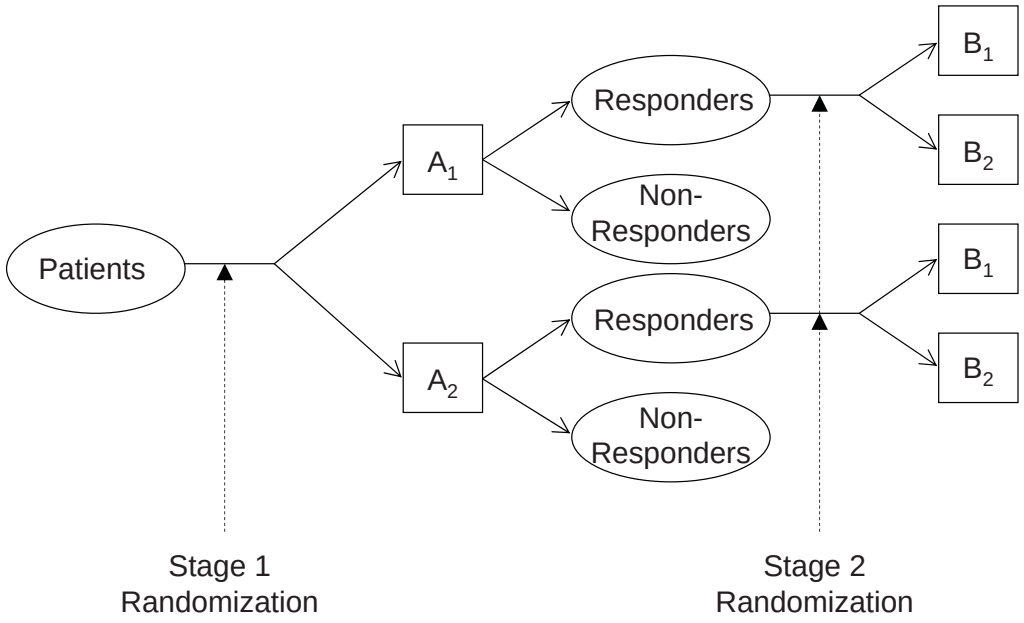


Figure 2.1: An example of a two-stage SMART design where only responders receive maintenance therapy

Definition 1. *Adaptive Treatment Strategy A_jB_k: 'Treat with A_j followed by B_k if the patient is eligible and consents to subsequent second-line therapy'.*

Furthermore, we classify strategies into shared-path and separate-path adaptive treatment strategies as follows:

Definition 2. *Shared-Path Adaptive Treatment Strategies: Two-stage adaptive treatment strategies are shared-path if individuals treated with one strategy share a common path of treatment with individuals treated with the other strategy.*

For example, consider strategies A_1B_1 and A_1B_2. Strategy A_1B_1 dictates that a patient be treated with A_1 and then by B_1 only if the patient responds to A_1. Similarly, strategy A_1B_2 dictates that a patient be treated with A_1 and then by B_2 only if the patient responds to A_1. Thus, a patient who is treated under strategy A_1B_1 but did not respond to A_1 will receive exactly the same sequence of treatment as a patient who is treated under strategy A_1B_2 but did not respond. Therefore, strategies A_1B_1 and A_1B_2 are shared-path adaptive treatment strategies. Similarly, the pair (A_2B_1, A_2B_2) are shared-path.

Strategies that do not share a common path of treatment will be referred to as separate-path treatment strategies. As an example, strategies A_1B_1 and A_2B_1 are separate-path adaptive treatment strategies since patients treated with A_1B_1 can not receive a treatment sequence received by patients treated with A_2B_1. Similarly, pairs (A_1B_1, A_2B_2), (A_1B_2, A_2B_1), and (A_1B_2, A_2B_2) are also separate-path.

2.3 COUNTERFACTUALS

Counterfactual (or potential) outcomes (Rubin, 1974, Holland, 1986) are often used to construct estimands of interest from a population. In reality, every individual follows one specific treatment strategy, therefore for each individual, we observe only one outcome for the specific treatment strategy he/she followed. In theory, however, individuals in the population could follow any treatment strategy A_jB_k, and for each individual, one can envision one outcome for each possible strategy, hence every individual has his/her own set of imaginary (potential) outcomes for every possible treatment strategy. The entire set of possible outcomes for an individual is referred to as his/her counterfactual outcomes. These outcomes will help us identify the variables whose distributions are compared across treatment strategies.

In order to define patients' counterfactual outcomes, which in this setting are potential survival times, we introduce the following notation. For patient i, let $R_{ji} = 1$ if the ith patient responded to the initial treatment A_j and $R_{ji} = 0$ if the ith patient did not respond to initial treatment A_j. Let T_{ji}^{NR} be the survival time for patient i if he/she received but did not respond to therapy A_j. Further, let T_{jki}^{R} denote the the survival time for patient i if he/she responded to treatment A_j and received treatment B_k. For treatment strategy A_jB_k, patient i receives one induction treatment, A_1 or A_2, either responds or does not respond to that particular induction treatment, and at the second stage, based on the response from the first stage, either receives B_1, B_2, or no further treatment. Thus, every patient only follows one path within a treatment strategy and we cannot observe $(R_{1i}, R_{2i}, T_{1i}^{NR}, T_{2i}^{NR}, T_{11i}^{R}, T_{12i}^{R}, T_{21i}^{R}, T_{22i}^{R})$ for each patient i. Consequently, these variables are the counterfactuals or potential random variables; those variables which could potentially

occur under any possible treatment strategy. For patient i following strategy A_jB_k, the potential survival time, T_{jki}, can be expressed in terms of his/her counterfactual outcomes as

$$T_{jki} = (1 - R_{ji})T_{ji}^{NR} + R_{ji}T_{jki}^{R}.$$

We will use these potential survival times to construct a weighted log-rank statistic to compare two or more separate-path or shared-path adaptive treatment strategies. First we will focus on comparing two shared-path adaptive treatment strategies, A_1B_1 and A_1B_2 or, equivalently, the distributions of T_{11} and T_{12}, and then generalize our statistic to compare more than two strategies with a specific extension to compare all four adaptive treatment strategies, A_1B_1, A_1B_2, A_2B_1 and A_2B_2 in Chapter 3.

2.4 OBSERVED DATA & ASSUMPTIONS

The observed data for a two-stage design described in Figure 2.1 can be represented as a set of random vectors $\{X_i, R_i, R_iT_i^R, R_iZ_i, U_i, \delta_i\}$, for $i = 1, \ldots, n$, where $X_i = 2 - j$ if the ith patient is randomized to induction treatment A_j $(j = 1, 2)$, R_i is the observed response indicator such that $R_i = 1$ if the ith patient is a responder to A_j and $R_i = 0$ otherwise, $Z_i = 2 - k$ if patient i is assigned to treatment B_k $(k = 1, 2)$, the event time is $U_i = \min(T_i, C_i)$, where C_i is the potential censoring time and T_i is the survival time for patient i, and $\delta_i = I(T_i \leq C_i)$. If T_i^R denotes the time to response for patient i who has responded to initial treatment, then the observed response R_i can be expressed as $R_i = X_iR_{1i}I(C_i > T_i^R) + (1 - X_i)R_{2i}I(C_i > T_i^R)$, where R_{ji} is the counterfactual response defined in Section 2.3.

First we make the stable unit treatment value assumption or consistency (Rubin, 1974) to relate the uncensored survival time T_i to the counterfactual outcomes. Explicitly, this assumption states that under the treatment assignment consistent with the counterfactual outcome, the observed outcome is equal to the counterfactual ($T_i = \sum_{j=1}^{2} X_{ji}[(1-R_{ji})T_{ji}^{NR} + R_{ji}\{Z_iT_{j1i}^R + (1 - Z_i)T_{j2i}^R\}]$, where $X_{1i} = X_i$ and $X_{2i} = 1 - X_i$). Other frequently made

assumptions such as 'no unmeasured confounders' and positivity (all treatment strategies have positive probability of being observed) follow from random assignment of treatments (Orellana et al., 2010). Since most clinical trials have limited follow-up, the survival time here is restricted to time L, where L is some value less than the maximum follow-up time for all patients in the sample.

2.5 STANDARD UNWEIGHTED LOG-RANK STATISTIC

The standard unweighted log-rank test statistic is well known, well documented and commonly used to compare survival curves for independent groups following a specified strategy. If there were no second randomization and each patient was set to follow either A_1B_1 or A_1B_2, data from patients receiving A_1B_1 would be considered independent of the data from patients receiving A_1B_2. To compare the two independent groups of patients following predetermined strategies A_1B_1 and A_1B_2 (to test the null hypothesis of no difference between the two survival distributions) based on the observed data $\{U_{1ki} = min(T_{1ki}, C_i), \delta_{1ki} = I(T_{1ki} \leq C_i), k = 1, 2;\ i = 1, \ldots, n\}$, we would use the standard unweighted log-rank test statistic

$$Z_n(t) = \int_0^t \frac{Y_{11}(s)Y_{12}(s)}{Y_{11}(s) + Y_{12}(s)} \left\{ \frac{dN_{11}(s)}{Y_{11}(s)} - \frac{dN_{12}(s)}{Y_{12}(s)} \right\}, \tag{2.1}$$

where $N_{1ki}(s) = I(U_{1ki} \leq s, \delta_{1ki} = 1)$, $Y_{1ki}(s) = I(U_{1ki} \geq s)$, $N_{1k}(s) = \sum_{i=1}^{n} N_{1ki}(s)$, and $Y_{1k}(s) = \sum_{i=1}^{n} Y_{1ki}(s)$ for $k = 1, 2$. Under the null hypothesis, $n^{-1/2}Z_n(t)$ is asymptotically normally distributed with mean zero and a variance that can be consistently estimated from the observed event times. For details of the properties of the standard unweighted log-rank statistic, we refer the readers to Fleming and Harrington (1991).

The standard unweighted log-rank statistic is inadequate, however, to test survival curves in a two-stage randomized design. First, the standard unweighted log-rank statistic does not account for the second randomization in a two-stage SMART design. In such design, U_{11i} is not observed for patient i who responded to A_1, but is randomized to maintenance treatment B_2 and likewise, U_{12i} is not observed for patient i who responded to A_1, but is randomized to maintenance treatment B_1. Second, since non-responders to A_1 are consistent with both

adaptive treatment strategies A_1B_1 and A_1B_2, the non-responders to A_1 are common to both groups. Hence, the two groups of patients following adaptive treatment strategies A_1B_1 and A_1B_2 are not statistically independent.

The first inadequacy of the standard unweighted log-rank statistic has been addressed by Guo in his unpublished 2005 PhD thesis from North Carolina State University (Guo, 2005), where a weighted version of the log-rank statistic was proposed to account for the second randomization. This statistic weights the at-risk and event processes according to the response status and randomization probability for each individual. This weighted log-rank statistic and the corresponding supremum version (Feng and Wahed, 2008), however, are only applicable to testing separate-path strategies (e.g. A_1B_1 vs. A_2B_1). Since the second inadequacy of the standard unweighted log-rank statistic remains even with the weighted log-rank statistic, we will address it in this chapter. Specifically, we propose a weighted log-rank statistic to test the hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$, where $\Lambda_{jk}(t)$ is the cumulative hazard at time t of those following strategy A_jB_k, accounting for the fact that patients following A_1B_1 includes a group of patients who also follow A_1B_2.

2.6 WEIGHTED LOG-RANK STATISTIC

We present the notation for time-dependent weights which is adapted from Guo and Tsiatis (2005). Explicitly, let

$$W_{11i}(s) = \frac{X_i}{\phi}\left\{1 - R_i(s) + \frac{R_i(s)Z_i}{\pi}\right\} \tag{2.2}$$

be the weight assigned to the ith patient at time s for the purpose of estimating quantities related to the strategy A_1B_1, where $R_i(s) = R_iI(T_i^R \leq s)$, such that $R_i(s) = 1$ if the ith patient responded to A_1 by time s, 0, otherwise, π is the known probability of a patient being assigned to maintenance therapy B_1, and ϕ is the known probability of a patient being assigned to first-line therapy A_1. Similarly,

$$W_{12i}(s) = \frac{X_i}{\phi}\left\{1 - R_i(s) + \frac{R_i(s)(1 - Z_i)}{1 - \pi}\right\} \tag{2.3}$$

for estimating quantities related to the strategy A_1B_2. Note that if a patient has not responded by time s, $W_{11i}(s) = W_{12i}(s) = \frac{1}{\phi}$, confirming that the non-responders are consistent with both strategies and the only weight is due to the randomization probability to A_1; if the patient has responded and is randomized to B_1 by time s, $W_{11i}(s) = \frac{1}{\phi\pi}$ and $W_{12i}(s) = 0$; if the patient has responded and is randomized to B_2 by time s, however, $W_{11i}(s) = 0$ and $W_{12i}(s) = \frac{1}{\phi(1-\pi)}$. This construction of weights is based on the fundamental principle of inverse-probability-of-treatment-weighting (Robins et al., 1994).

Table 2.1: At-risk process notation

Term	Definition	Description
$Y_i(s)$	$I(U_i \geq s)$	$Y_i(s)$=1 when individual i is at-risk at time s regardless of what treatment he/she receives, 0 otherwise
$Y_{jki}(s)$	$I(U_{jki} \geq s, X_i = 2-j, Z_i = 2-k)$	$Y_{jki}(s)$=1 when individual i following treatment strategy A_jB_k is at-risk at time s, 0 otherwise
$\bar{Y}_{jk}(s)$	$\sum_{i=1}^n W_{jki}(s)Y_i(s)$	The weighted number of individuals at-risk at time s following treatment strategy A_jB_k
$Y_j^{NR}(s)$	$\sum_{i=1}^n I(X_i = 2-j)\{1-R_i(s)\}Y_i(s)$	The number of individuals who have yet to respond to treatment A_j and are at-risk at time s
$Y_{j.}(s)$	$\sum_{i=1}^n I(X_i = 2-j)Y_i(s)$	The number of individuals with initial treatment A_j and are at-risk at time s
$Y(s)$	$\sum_{i=1}^n Y_i(s)$	The number of all individuals at risk at time s regardless of what treatment they receive

To facilitate the derivation of the desired test statistic to compare shared-path adaptive treatment strategies and its asymptotic properties, we introduce further notation. For quick reference, we included these in Tables 2.1 and 2.2. The general at-risk process for all patients regardless of the strategy that they follow is $Y_i(s) = I(U_i \geq s)$, the weighted at-risk process is $\bar{Y}_{jk}(s) = \sum_{i=1}^n W_{jki}(s)Y_i(s)$, the at-risk process for only those who are non-responders to A_j is $Y_j^{NR}(s) = \sum_{i=1}^n I(X_i = 2-j)\{1-R_i(s)\}Y_i(s)$, the overall at-risk process for patients treated with A_j is $Y_{j.}(s) = \sum_{i=1}^n I(X_i = 2-j)Y_i(s)$ and the overall at risk-process for all patients is $Y(s) = \sum_{i=1}^n Y_i(s)$. Likewise, the general event process

Table 2.2: Event process notation

Term	Definition	Description
$N_i(s)$	$I(U_i \leq s, \delta = 1)$	$N_i(s)$=1 when individual i has an event at or before time s regardless of what treatment he/she receives, 0 otherwise
$N_{jki}(s)$	$I(U_{jki} \leq s, \delta_i = 1, X_i = 2-j, Z_i = 2-k)$	$N_{jki}(s)$=1 when individual i following treatment strategy A_jB_k has an event at or before time s, 0 otherwise
$\bar{N}_{jk}(s)$	$\sum_{i=1}^{n} W_{jki}(s)N_i(s)$	The weighted number of events at or before time s for individuals following treatment strategy A_jB_k
$N_j^{NR}(s)$	$\sum_{i=1}^{n} I(X_i = 2-j)\{1-R_i(s)\}N_i(s)$	The number of individuals who are yet to respond to treatment A_j and have an event at or before time s
$N_{j\cdot}(s)$	$\sum_{i=1}^{n} I(X_i = 2-j)N_i(s)$	The number of individuals with initial treatment A_j and have an event at or before time s
$N(s)$	$\sum_{i=1}^{n} N_i(s)$	The number of all individuals with an event at or before time s regardless of what treatment they receive

for any patient i is $N_i(s) = I(U_i \leq s, \delta_i = 1)$, the weighted event process is $\bar{N}_{jk}(s) = \sum_{i=1}^{n} W_{jki}(s)N_i(s)$, the event process for those who are non-responders to A_j is $N_j^{NR}(s) = \sum_{i=1}^{n} I(X_i = 2-j)\{1-R_i(s)\}N_i(s)$, the overall event process for patients treated with A_j is $N_{j\cdot}(s) = \sum_{i=1}^{n} I(X_i = 2-j)N_i(s)$, and the overall event process for all patients is $N(s) = \sum_{i=1}^{n} N_i(s)$. Based on these weighted processes, the inverse-probability-of-randomization weighted-log-rank statistic for testing H_0: $\Lambda_{11}(t) = \Lambda_{12}(t)$, where $\Lambda_{jk}(t)$ is the cumulative hazard at time t for those following strategy A_jB_k, is defined as

$$Z_n^W(t) = \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \left\{ \frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}_{12}(s)}{\bar{Y}_{12}(s)} \right\}. \tag{2.4}$$

The rationale behind this formulation of the test statistic is given in Feng and Wahed (2008). In short, the quantity $d\bar{N}_{1k}(s)/\bar{Y}_{1k}(s)$ is an unbiased estimator of the instantaneous

event rate at time s, $d\Lambda_{1k}(s)$. Therefore, it serves the same purpose of $dN_{1k}(s)/Y_{1k}(s)$ in the standard unweighted log-rank test defined in equation (2.1). Under the null hypothesis $\Lambda_{11}(t) = \Lambda_{12}(t)$, since the term $\{\bar{Y}_{11}(s)\bar{Y}_{12}(s)\}/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}$ is predictable (with respect to the filtration $\mathcal{F}(t) = \sigma\{X_i, R_i(s), R_i(s)Z_i, I(C_i \leq s), N_i(s), i = 1, \ldots, n;\ j = 1, 2;\ 0 \leq s \leq t\}$), the weighted log-rank statistic in equation (2.4) has expectation zero (see Section 2.7).

While the weighted log-rank statistic looks almost identical to that of the standard unweighted log-rank statistic, note that the terms $d\bar{N}_{11}(s)/\bar{Y}_{11}(s)$ and $d\bar{N}_{12}(s)/\bar{Y}_{12}(s)$ are correlated unlike the unweighted versions from the predetermined strategies in the standard log-rank statistic. The variance calculation will change substantially in order to account for this correlation between these two terms. The variance calculation presented in the next section addresses the second and remaining inadequacy of the standard log-rank and supremum log-rank tests. We will use a standardized version of the statistic from equation (2.4) to test the null hypothesis H_0: $\Lambda_{11}(t) = \Lambda_{12}(t)$.

2.7 ASYMPTOTIC PROPERTIES

First we note that $n^{-1/2}Z_n^W(t)$ in equation (2.4) can be expressed as a sum of two terms using the definition of martingale increments. Explicitly,

$$n^{-1/2}Z_n^W(t) = G_n(t) + R_n(t) \tag{2.5}$$

where

$$G_n(t) \;=\; n^{-1/2}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)}\left\{\frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{12}(s)}{\bar{Y}_{12}(s)}\right\} \tag{2.6}$$

and

$$R_n(t) \;=\; n^{-1/2}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)}\{d\Lambda_{11}(s) - d\Lambda_{12}(s)\}, \tag{2.7}$$

since $\bar{M}_{jk}(t) = \bar{N}_{jk}(t) - \int_0^t \bar{Y}_{jk}(s)d\Lambda_{jk}(s)$.

We must show the following two results regarding the weighted martingale to derive the asymptotic properties of the weighted log-rank statistic, $n^{-1/2}Z_n^W(t)$: (i) $d\bar{M}_{jk}(s) =$

$\sum_{i=1}^{n} W_{jki}(s)dM_{jki}(s)$ and (ii) $E\{d\bar{M}_{jk}(s)|\mathcal{F}(s_-)\} = 0$, where $M_{jki}(s)$ is the ith patient specific martingale, corresponding to $M_{jk}(s) = N_{jk}(s) - \int_0^s Y_{jk}(u)d\Lambda_{jk}(u)$, the usual martingale process for strategy A_jB_k, had there been no second randomization and each patient followed a pre-specified (perhaps randomized) treatment strategy. The first result directly follows from the argument laid out in Feng and Wahed (2008, p. 699). For the second result, arbitrarily setting $j = k = 1$,

$$
\begin{aligned}
&E\{d\bar{M}_{11}(s)|\mathcal{F}(s_-)\} \\
&\quad = E\left\{\sum_{i=1}^{n} W_{11i}(s)dM_{11i}(s)|\mathcal{F}(s_-)\right\} \\
&\quad = E\left[\sum_{i=1}^{n} \frac{X_i}{\phi}\left\{1 - R_i(s) + R_i(s)\frac{Z_i}{\pi}\right\}\{dN_{11i}(s) - Y_{11i}(s)d\Lambda_{11}(s)\} \mid \mathcal{F}(s_-)\right] \\
&\quad = \sum_{i=1}^{n} \frac{X_i}{\phi}\left\{1 - R_i(s) + R_i(s)\frac{Z_i}{\pi}\right\}[E\{dN_{11i}(s)|\mathcal{F}(s_-)\} - Y_{11i}(s)d\Lambda_{11}(s)] \\
&\quad = \sum_{i=1}^{n} \frac{X_i}{\phi}\left\{1 - R_i(s) + R_i(s)\frac{Z_i}{\pi}\right\}\{Y_{11i}(s)d\Lambda_{11}(s) - Y_{11i}(s)d\Lambda_{11}(s)\} \\
&\quad = 0.
\end{aligned}
$$

We have used the fact that the expected value of the increment of the event process corresponding to the patients who followed the treatment strategy of interest, here A_1B_1, given the history, is zero. That is, $E\{dN_{11i}(s)|\mathcal{F}(s_-)\} = Y_{11i}(s)d\Lambda_{11}(s)$. Likewise, the results follow for any $j, k = 1, 2$.

Then, under the null hypothesis, $\Lambda_{11}(t) = \Lambda_{12}(t)$, so $n^{-1/2}Z_n^W(t) = G_n(t)$ in equation (2.6). Since we have just shown that martingale increments have mean zero, $E\{Z_n^W(t)\} = 0$. Thus, $Z_n^W(t)$ has mean zero under the null hypothesis of no difference in hazards between two strategies. To derive the variance of $n^{-1/2}Z_n^W(t)$, we can further expand $G_n(t)$ in equation (2.6). Using the first result of weighted martingales ($d\bar{M}_{jk}(s) = \sum_{i=1}^{n} W_{jki}(s)dM_{jki}(s)$), $G_n(t)$ can be expressed as a difference of two martingale processes, $G_n(t) = G_n^{11}(t) - G_n^{12}(t)$:

$$
n^{-1/2}\left\{\sum_{i=1}^{n}\int_0^t \frac{\bar{Y}_{12}(s)W_{11i}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)}dM_{11i}(s) - \sum_{i=1}^{n}\int_0^t \frac{\bar{Y}_{11}(s)W_{12i}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)}dM_{12i}(s)\right\}. \tag{2.8}
$$

By the martingale central limit theorem (Fleming and Harrington, 1991, Ch. 5), $G_n^{1k}(t)$ converges to a Gaussian process with mean zero. Therefore, $G_n(t)$ converges to a Gaussian process with mean zero and variance equal to $var\{G_n^{11}(t)\}+var\{G_n^{12}(t)\}-2cov\{G_n^{11}(t), G_n^{12}(t)\}$. The variances of $G_n^{11}(t)$ and $G_n^{12}(t)$ can be calculated the same way as the variance for the weighted log-rank statistic in Feng and Wahed (2008). More explicitly, $var\{G_n^{1k}(t)\}$ is the limit of $n^{-1}\sum_{i=1}^n \int_0^t \frac{\bar{Y}^2_{1(3-k)}(s)W^2_{1ki}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2} Y_i(s)d\Lambda_{1k}(s)$, $k = 1, 2$. To find the covariance between two martingale processes, $cov\{G_n^{11}(t), G_n^{12}(t)\}$, we use the formula from Fleming and Harrington (1991, p. 70). Explicitly, if H_1 and H_2 are locally-bounded, predictable processes and M_1 and M_2 are local martingales then the covariance between $\int H_1 dM_1$ and $\int H_2 dM_2$ is $\int H_1 H_2 cov(dM_1, dM_2)$. Then, the asymptotic variance of $G_n(t)$ can be expressed as the limiting value of

$$n^{-1}\sum_{k=1}^{2}\sum_{i=1}^{n}\int_0^t \frac{\bar{Y}^2_{1(3-k)}(s)W^2_{1ki}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2} Y_i(s)d\Lambda_{1k}(s) \tag{2.9}$$

$$-2n^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}\sum_{i=1}^{n} W_{11i}(s)W_{12i}(s)cov\{dM_{11i}(s), dM_{12i}(s)\}. \tag{2.10}$$

First, note that $W_{11i}(s)W_{12i}(s) = X_i\{1 - R_i(s)\}/\phi^2$. Subsequently, under the null hypothesis, $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_0(t)$, the term inside the summation in the second line of equation (2.9) can be shown to be equal to

$$\sum_{i=1}^{n} X_i\{1 - R_i(s)\}\{Y_i(s)d\Lambda_1^{NR}(s)\}/\phi^2 = Y_1^{NR}(s)d\Lambda_1^{NR}(s)/\phi^2.$$

Thus the variance of $G_n(t)$ can now be expressed as the limiting value of

$$\begin{aligned} &n^{-1}\int_0^t \frac{d\Lambda_0(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}\left\{\bar{Y}_{12}^2(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s) + \bar{Y}_{11}^2(s)\sum_{i=1}^{n} W_{12i}^2(s)Y_i(s)\right\} \\ &-2(n\phi^2)^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\left\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\right\}^2}\sum_{i=1}^{n} X_i\{1 - R_i(s)\}Y_i(s)d\Lambda_1^{NR}(s). \end{aligned} \tag{2.11}$$

A consistent variance estimator of $n^{-1/2}Z_n^W(t)$ is then given by

$$\begin{aligned} \hat{\sigma}^2(t) = n^{-1}&\int_0^t \frac{\bar{Y}_{12}^2(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s) + \bar{Y}_{11}^2(s)\sum_{i=1}^{n} W_{12i}^2(s)Y_i(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}\left\{\frac{dN_{1.}(s)}{Y_{1.}(s)}\right\} \\ &-2(n\phi^2)^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\left\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\right\}^2}\left\{Y_1^{NR}(s)\frac{dN_1^{NR}(s)}{Y_1^{NR}(s)}\right\}. \end{aligned} \tag{2.12}$$

The notation used in the above equation or elsewhere in this chapter can be reviewed in Tables 2.1 and 2.2. The corresponding standardized weighted log-rank test statistic is given by $T_n^W(L)$, where

$$T_n^W(L) = n^{-1/2} Z_n^W(L)/\hat{\sigma}(L), \tag{2.13}$$

and L, as noted before, is less than the maximum follow-up time. The level α weighted log-rank test rejects the equality of two shared-path adaptive treatment strategies' cumulative hazards when $|T_n^W(L)| \geq Z_{1-\alpha/2}$ where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of a standard normal distribution.

2.8 SIMULATION STUDIES

2.8.1 Data Generation

To evaluate the performance of the weighted log-rank statistics for comparing two shared-path adaptive treatment strategies, we conducted a series of Monte Carlo simulations. We were interested in assessing the type I error rate under the null hypothesis of no difference in overall survival and in assessing the power of the weighted log-rank statistics under various alternative scenarios. As stated in the introduction, often, shared-path adaptive treatment strategies are compared either by turning them into independent groups by using only those patients who responded to the first-line treatment or pretending as if all those who follow each strategy form independent groups. Since the first comparison addresses a different question by comparing the second-stage treatments conditional on response (instead of comparing entire adaptive treatment strategies), we have not included this statistic for comparison in our simulation studies. Instead, in our simulation studies, to test the equality of two shared-path adaptive treatment strategies, we have compared the proposed weighted log-rank statistic, $T_n^W(L)$ from equation (2.13) referred to as WLR, to a similar weighted log-rank statistic that treats the two groups independently such that the variance ignores the covariance term, hence referred to as the independent weighted log-rank test (IWLR), and

to the standard unweighted log-rank (SLR) statistic applied to two groups of patients who followed each strategy. The groups for the standard unweighted log-rank statistic were formed by combining those who did not respond to A_j to those who responded to A_j and received treatment B_k. For example, the group representing adaptive treatment strategy A_1B_1 consists of all the non-responders to A_1 and all those who responded to A_1 and were subsequently assigned to receive B_1 and the group representing adaptive treatment strategy A_1B_2 consists of all the non-responders to A_1 and all those who responded to A_1 and were subsequently assigned to receive B_2.

We outline the general data generation process to compare two or more ATS here and provide specific parameters for each simulation in Sections 2.8.2 and 2.8.3 for comparing two shared-path ATS and in Sections 3.4.1 and 3.4.2 for comparing multiple ATS. The initial treatment indicator, X_i, was generated from a Bernoulli distribution with $pr(X_i = 1) = 0.5$ so that there were about an equal number of patients initially treated with A_1 and A_2. We took R_i, the response indicator, to be Bernoulli with $pr(R_i = 1) = \pi_R$, $\pi_R \in (0.4, 0.6)$, so that there were 40% or 60% of patients who responded to the initial treatment. When $R_i = 0$, a survival time $T_{ji}^{NR}, j = 1, 2$, was generated from an exponential distribution with mean μ_j^{NR}. When $R_i = 1$, the treatment B_1 indicator, Z_i, was generated from a Bernoulli(0.5) distribution. Also when $R_i = 1$, time to response, T_{ji}^{R}, $j = 1, 2$, was generated from an exponential distribution with mean θ_j^R and time from response to an event, T_{jki}^{RE}, $j, k = 1, 2$, was generated from an exponential distribution with mean θ_{jk}^{RE}. The total survival time for those who responded to A_j and were randomized to B_k is thus, $T_{jki}^{*} = T_{ji}^{R} + T_{jki}^{RE}$, for $j, k = 1, 2$. The variables of interest here are the time-to-events, T_{jki}, where $T_{jki} = (1 - R_i)T_{ji}^{NR} + R_iT_{jki}^{*}$, $j, k = 1, 2$. These variables reflect the overall survival time under strategy A_jB_k, $(j, k = 1, 2)$. The observed survival time for the ith individual in the absence of censoring is defined as $T_i = X_i[R_i\{Z_iT_{11i}^{*} + (1 - Z)T_{12i}^{*}\} + (1 - R_i)T_{1i}^{NR}] + (1 - X_i)[R_i\{Z_iT_{21i}^{*} + (1 - Z_i)T_{22i}^{*}\} + (1 - R_i)T_{2i}^{NR}]$. Additionally, a right censored time, C_i, was generated from a uniform distribution from zero to v, such that 30% or 50% of the population were censored. Censoring was independent and uninformative of response and survival time. The final observed time was then defined as $U_i = min(T_i, C_i)$ with corresponding complete case indicator, $\delta_i = I(T_i \leq C_i)$.

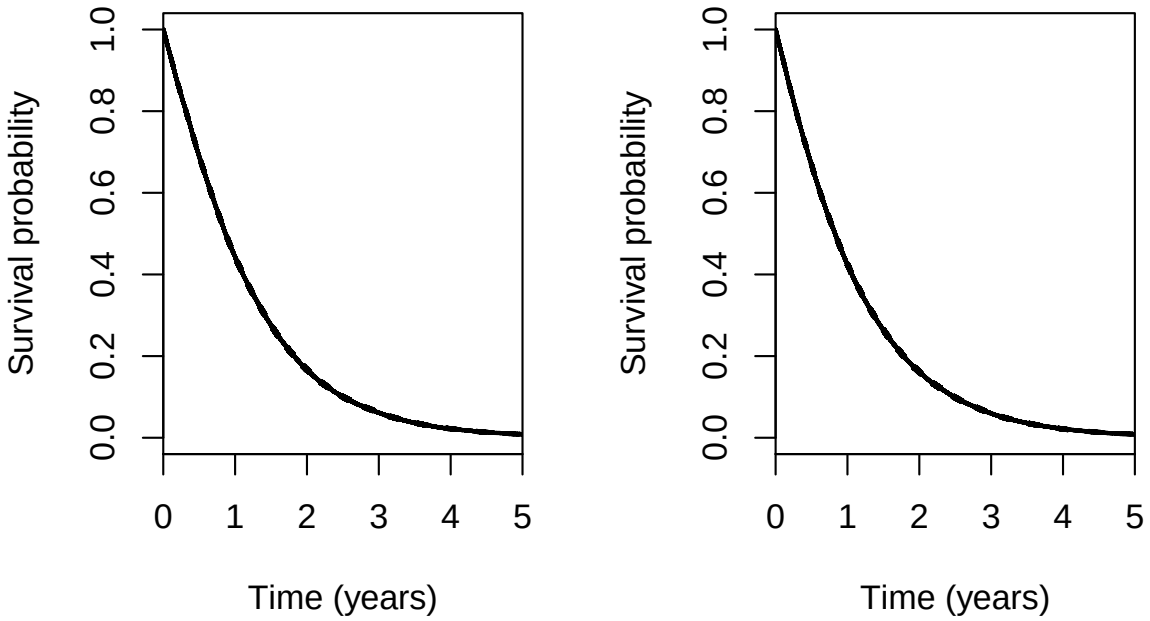


Figure 2.2: Survival curves under the null distribution that all adaptive treatment strategies' cumulative hazards are equal for 40% response rates (left panel) and 60% response rate (right panel)

For each generated dataset we conducted the weighted log-rank test described in Section 2.7 to test the hypotheses $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_0(t)$. We report the estimated type I error (proportion of samples for which the hypothesis was falsely rejected) for all tests in Tables 2.3 when $H_0$was true, and the estimated power (proportion of samples for which the hypothesis was correctly rejected) for all tests in Tables 2.4-2.6.

2.8.2 Simulation from the Null Distribution

To investigate the performance of the weighted log-rank statistic under the null hypotheses, we generated 5000 datasets with the following parameters: $\theta_1^R = 0.5$ and $\theta_{11}^{RE} = \theta_{12}^{RE} = 1$. With a 40% response rate, $\mu_1^{NR} = 0.91$ and the censoring paramter v was set to 3.80 and 2.00, and with a 60% response rate, $\mu_1^{NR} = 0.56$, v was set to 3.60 and 1.85 to produce about 30% and 50% censoring, respectively.

Table 2.3 presents the estimated type I error rates for testing the null hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_0(t)$ under several combinations of sample size, response rates and

Table 2.3: Type I error rate under null hypotheses $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$

Response		30% censoring			50% censoring		
Rate (%)	n	WLR	IWLR	SLR	WLR	IWLR	SLR
40	250	0.057	0.006	<0.001	0.060	0.004	<0.001
	500	0.052	0.005	<0.001	0.056	0.002	<0.001
	1000	0.049	0.006	<0.001	0.053	0.004	<0.001
50	250	0.055	0.010	0.002	0.058	0.009	<0.001
	500	0.061	0.014	0.003	0.055	0.008	<0.001
	1000	0.054	0.013	0.003	0.054	0.009	<0.001
60	250	0.052	0.017	0.008	0.053	0.010	0.002
	500	0.053	0.018	0.006	0.053	0.013	0.002
	1000	0.054	0.017	0.008	0.051	0.009	0.001

The target type I error rate is $\alpha = 0.05$. WLR is the weighted log-rank statistic in equation (2.13). IWLR is the independent weighted log-rank statistic. SLR denotes the standard unweighted log-rank statistic.

censoring for the WLR, IWLR, and SLR tests. For a sample size of 500, a response rate of 40% and censoring of 30%, the type I error for the WLR test was very close to the nominal level of 0.05. The IWLR statistic does not subtract the covariance term between the shared-path strategies and therefore rejects the null hypothesis less often leading to a more conservative test with an approximate error rate of 0.005 for a nominal level of 0.05. The SLR test, which combines and equally weights all patients who follow a strategy regardless of their response status, also yielded very conservative type I error rates with an estimate, in this case, of less than 0.001. Preserving the response rate at 40%, but increasing censoring led to a similar type I error rate, such that the WLR test had an estimated type I error rate of 0.056, the IWLR test and the SLR test had estimated error rates less than 0.001. In general, increasing censoring did not affect the estimated type I error rate for the WLR test as it maintained the nominal level of 0.05 in all scenarios with greater than 250 individuals. Preserving censoring at 30% or at 50% and increasing the response rate from 40% to 60% led to about the same estimated type I error rates for the WLR test and slightly higher rates for the IWLR test and the SLR test, but these two tests remained overly conservative.

2.8.3 Simulation from Alternative Distributions

Since the type I error rates were generally upheld, we explored a variety of scenarios performing 5000 iterations to test the power of the weighted log-rank test at a sample size of 250, 500, and 1000. Data were generated from populations under alternative hypotheses where the cumulative hazards of the adaptive treatment strategies were not equal. The true survival distributions under the alternative hypotheses, designated as scenarios (a)-(d), are plotted in Figure 2.3 when 60% of the population respond to A_1. The strategies A_1B_1 and A_1B_2 here, are the same as those used for comparison of alternative distributions of four ATS in Section 3.4.2. The parameters for scenario (a) were set as follows: $\mu_1^{NR} = \theta_1^R = 1$, $\theta_{11}^{RE} = 2$, and $\theta_{12}^{RE} = 3.33$. The parameters for scenario (b) the parameters were set as follows: $\mu_1^{NR} = \theta_1^R = \theta_{11}^{RE} = 1$, and $\theta_{12}^{RE} = 3.33$. The parameters for scenario (c) were set as follows: $\mu_1^{NR} = 1$, $\theta_1^R = 2$, $\theta_{11}^{RE} = 0.67$, and $\theta_{12}^{RE} = 0.5$. Finally, the parameters for scenario (d) were set as follows: $\mu_1^{NR} = 1.43$, $\theta_1^R = 0.2$, $\theta_{11}^{RE} = 1$, and $\theta_{12}^{RE} = 1.67$. The censoring parameter v was set to 5 for scenarios (a) and (b), 6 for scenario (c) and 5.5 for scenario (d).

Table 2.4 presents the results for testing the null hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_0(t)$ versus the alternative hypothesis that the cumulative hazards for the two shared-path adaptive treatment strategies differ for a sample size of 250. The WLR test had much greater power to correctly reject the null hypothesis than the IWLR test and especially when compared to the SLR test. In all cases, increasing the response rate from 40% to 60% increased the power of all tests, but the WLR test always maintained the greatest power. In particular, note the large difference in power in scenario (b). The WLR test maintained very high power in this situation, 0.848 for 40% responders, while the IWLR test had about half the power at 0.458, and the SLR test failed to pick up the difference in the survival curves in most of the iterations with power of 0.088.

Tables 2.5 and 2.6 present the results for testing the null hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_0(t)$ for a sample size of 500 and 1000, respectively. Since the sample size increased from 250 to 500 to 1000, the power for all statistics in Table 2.6 is higher than the power in Table 2.5 which is higher than the power in Table 2.4. For scenario (b), the power for both the WLR and IWLR is very high, with the power of the SLR reaching to above 80% with

60% responders for a sample size of 500 and 50% responders for a sample size of 1000. The pattern of increased power for all tests with an increasing percentage of responders remains the same. The survival curves for those following strategies A_1B_1 and A_1B_2 are very similar in scenario (c) and thus it would take a very large sample size with any amount of responders to find this difference.

In conclusion, the WLR test statistic generally maintained the type I error rates for sample sizes greater than 250, whereas the IWLR and SLR were overly conservative for all sample sizes. The power to detect differences in two shared-path ATS was highest for the WLR test statistic in all scenarios tested.

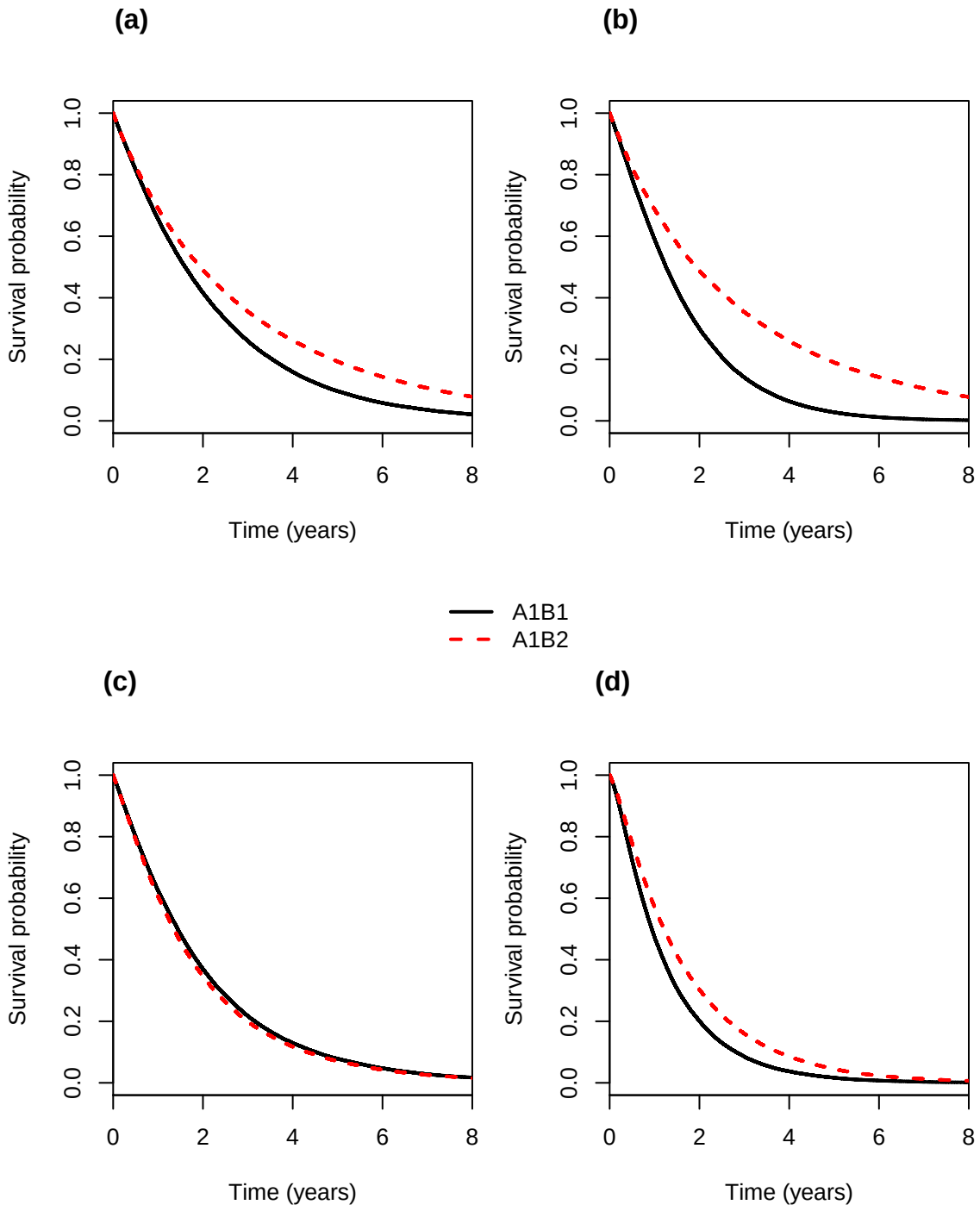


Figure 2.3: Survival curves for treatment strategies A_1B_1 (solid), A_1B_2 (dashes), under different alternative hypotheses scenarios for 60% responders.

Table 2.4: Power against alternative survival curves under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$ for sample size n=250

Scenario	Response Rate (%)	Censoring Rate (%)	Power WLR	IWLR	SLR
(a)	40	36	0.128	0.020	0.006
	50	40	0.138	0.036	0.020
	60	44	0.135	0.059	0.035
(b)	40	33	0.848	0.458	0.088
	50	36	0.899	0.676	0.228
	60	39	0.942	0.823	0.424
(c)	40	26	0.077	0.002	<0.001
	50	29	0.090	0.007	<0.001
	60	31	0.094	0.019	<0.001
(d)	40	26	0.210	0.077	0.006
	50	26	0.331	0.179	0.039
	60	26	0.419	0.300	0.098

See Figure 2.3 and Section 2.8.3 for a description of alternative survival scenarios (a)-(d). WLR is the weighted log-rank statistic in equation (2.13). IWLR is the independent weighted log-rank statistic. SLR denotes the standard unweighted log-rank statistic.

Table 2.5: Power against alternative survival curves under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$ for sample size n=500

Scenario	Response Rate (%)	Censoring Rate (%)	Power WLR	IWLR	SLR
(a)	40	36	0.230	0.055	0.021
	50	40	0.271	0.100	0.048
	60	44	0.304	0.155	0.088
(b)	40	33	0.987	0.865	0.331
	50	36	0.997	0.970	0.620
	60	39	0.999	0.993	0.841
(c)	40	26	0.099	0.004	<0.001
	50	29	0.113	0.012	<0.001
	60	31	0.119	0.036	<0.001
(d)	40	26	0.443	0.213	0.024
	50	26	0.605	0.419	0.121
	60	27	0.723	0.607	0.295

See Figure 2.3 and Section 2.8.3 for a description of alternative survival scenarios (a)-(d). WLR is the weighted log-rank statistic in equation (2.13). IWLR is the independent weighted log-rank statistic. SLR denotes the standard unweighted log-rank statistic.

Table 2.6: Power against alternative survival curves under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$ for sample size n=1000

Scenario	Response Rate (%)	Censoring Rate (%)	Power WLR	IWLR	SLR
(a)	40	36	0.468	0.169	0.056
	50	40	0.546	0.289	0.133
	60	44	0.628	0.435	0.258
(b)	40	33	1.000	0.997	0.786
	50	36	1.000	1.000	0.962
	60	39	1.000	1.000	0.996
(c)	40	26	0.155	0.012	<0.001
	50	29	0.168	0.036	<0.001
	60	31	0.201	0.068	<0.001
(d)	40	26	0.772	0.543	0.136
	50	26	0.902	0.801	0.423
	60	36	0.962	0.924	0.714

See Figure 2.3 and Section 2.8.3 for a description of alternative survival scenarios (a)-(d). WLR is the weighted log-rank statistic in equation (2.13). IWLR is the independent weighted log-rank statistic. SLR denotes the standard unweighted log-rank statistic.

3.0 LOG-RANK STATISTICS FOR MULTIPLE ADAPTIVE TREATMENT STRATEGIES

3.1 INTRODUCTION

In addition to comparing two shared-path adaptive treatment strategies, we would like to compare more than two ATS which may share the same treatment paths using test statistics similar to k-sample log-rank tests (Harrington and Fleming, 1982). Naive approaches to comparing the overall survival distributions of two shared-path or multiple adaptive treatment strategies include those mentioned in Section 1.4. Specifically, (i) ignoring the induction treatments, comparing second-line therapies conditioning on patients who were eligible to receive second-stage treatments, or (ii) using the statistics provided in Lokhnygina and Helterbrand (2007), Feng and Wahed (2008), or Li and Murphy (2011) but ignoring that these statistics were created for comparing separate-path adaptive treatment strategies, or (iii) forming groups where each group includes all of the patients who follow each adaptive treatment strategy and applying the standard unweighted log-rank test. The first option ignores the two-stage design and answers a different question than that is intended, the second option inflates the variance of the stated statistics, and the third option forms groups which contain some of the same patients violating the standard log-rank assumption that groups are statistically independent. The goal of this chapter is to address both the second-stage randomization as well as account for the covariance between shared-path ATS by extending the results from Chapter 2 and presenting the weighted log-rank statistic to compare multiple adaptive treatment strategies, some of which may be shared-path.

3.2 WEIGHTED LOG-RANK STATISTIC

In the setting described in Chapter 2, we would now like to extend the comparison to all four adaptive strategies, $A_jB_k,\ j,k=1,2$, and test the overall null hypothesis of no treatment effect. The null hypothesis that all cumulative hazards of those following ATS $A_jB_k,\ j,k=1,2$ are equal is stated as $H_0:\Lambda_{11}(t)=\Lambda_{12}(t)=\Lambda_{21}(t)=\Lambda_{22}(t)=\Lambda_0(t)$ against the alternative hypothesis, H_1: at least one cumulative hazard differs.

To derive the multivariate weighted log-rank statistic, we first notice that H_0 can be cast as a vectorized differences of cumulative hazards such that $H_0:\zeta(t)=0$ where $\zeta(t)=\{\Lambda_{11}(t)-\Lambda_{12}(t),\ \Lambda_{11}(t)-\Lambda_{21}(t),\ \Lambda_{11}(t)-\Lambda_{22}(t)\}^T$. Following Section 2.6, an unbiased estimator of $\zeta(\mathbf{t})$ is given by $\hat{\zeta}(t)=\int_0^t\{\frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)}-\frac{d\bar{N}_{12}(s)}{\bar{Y}_{12}(s)},\frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)}-\frac{d\bar{N}_{21}(s)}{\bar{Y}_{21}(s)},\frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)}-\frac{d\bar{N}_{22}(s)}{\bar{Y}_{22}(s)}\}^T$.

The corresponding weighted log-rank statistic for testing H_0 is defined as the vector of the weighted martingale differences, $Z_n^{MW}(t)=\{Z_n^{11.12}(t),Z_n^{11.21}(t),Z_n^{11.22}(t)\}^T$ where

$$Z_n^{jk.j'k'}(t)=\int_0^t\frac{\bar{Y}_{jk}(s)\bar{Y}_{j'k'}(s)}{\bar{Y}_{jk}(s)+\bar{Y}_{j'k'}(s)}\left\{\frac{d\bar{N}_{jk}(s)}{\bar{Y}_{jk}(s)}-\frac{d\bar{N}_{j'k'}(s)}{\bar{Y}_{j'k'}(s)}\right\}. \tag{3.1}$$

Under the null hypothesis, the statistic $Z_n^{MW}(t)$ has expectation zero. Since $Z_n^{MW}(t)$ is a linear combination of weighted Z^W-statistics defined in equation (2.4), by the multivariate central limit theorem for martingales (Fleming and Harrington, 1991), $n^{-1/2}Z_n^{MW}(t)$ follows a mean zero Gaussian process with asymptotic variance covariance matrix, $\Sigma(t)$, that can be estimated by

$$\hat{\Sigma}(t)=\begin{pmatrix}s_{11}(t) & s_{12}(t) & s_{13}(t)\\ s_{12}(t) & s_{22}(t) & s_{23}(t)\\ s_{13}(t) & s_{23}(t) & s_{33}(t)\end{pmatrix}, \tag{3.2}$$

where the elements of $\hat{\Sigma}(t)$ are defined as follows.

3.3 ASYMPTOTIC PROPERTIES

The estimated variance of the first component of $Z_n^{MW}(t)$, $s_{11}(t)$, is given in equation (2.12), except that the induction-treatment-specific processes, $N_{1\cdot}(s)$ and $Y_{1\cdot}(s)$, have been substituted with the overall processes $N(s)$ and $Y(s)$ to reflect that under the null hypothesis, all strategies have equal hazards. Explicitly,

$$\begin{aligned} s_{11}(t) =& n^{-1} \int_0^t \frac{\bar{Y}_{12}^2(s) \sum_{i=0}^n W_{11i}^2(s) Y_i(s) + \bar{Y}_{11}^2(s) \sum_{i=0}^n W_{12i}^2(s) Y_i(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}^2} \left\{ \frac{dN(s)}{Y(s)} \right\} \\ & - 2(n\phi^2)^{-1} \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}^2} \left\{ Y_1^{NR}(s) \frac{dN_1^{NR}(s)}{Y_1^{NR}(s)} \right\}. \end{aligned} \tag{3.3}$$

Similarly, the estimated variances of the second and third components of $Z_n^{MW}(t)$, $s_{22}(t)$ and $s_{33}(t)$, are given as,

$$s_{22}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{21}^2(s) \sum_{i=0}^n W_{11i}^2(s) Y_i(s) + \bar{Y}_{11}^2(s) \sum_{i=0}^n W_{21i}^2(s) Y_i(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{21}(s)\}^2} \left\{ \frac{dN(s)}{Y(s)} \right\} \tag{3.4}$$

$$s_{33}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{22}^2(s) \sum_{i=0}^n W_{11i}^2(s) Y_i(s) + \bar{Y}_{11}^2(s) \sum_{i=0}^n W_{22i}^2(s) Y_i(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{22}(s)\}^2} \left\{ \frac{dN(s)}{Y(s)} \right\}. \tag{3.5}$$

Note that equations (3.4) and (3.5) do not contain covariance terms since $d\bar{N}_{jk}(s)/\bar{Y}_{jk}(s)$ and $d\bar{N}_{j'k'}(s)/\bar{Y}_{j'k'}(s)$, $j \neq j'$, in $Z_n^{11.21}(t)$ and $Z_n^{11.22}(t)$, are conditionally independent given $\mathcal{F}(s_-)$.

To obtain an expression for the estimated covariance terms in equation (3.2), we first give the expressions for the asymptotic covariances and then present the corresponding estimates. We derive the covariance specifically for $\sigma_{12}(t) = n^{-1}cov\{Z_n^{11.12}(t), Z_n^{11.21}(t)\}$ corresponding to the estimated covariance $s_{12}(t)$; the derivations of $\sigma_{13}(t) = n^{-1}cov\{Z_n^{11.12}(t), Z_n^{11.22}(t)\}$ and $\sigma_{23}(t) = n^{-1}cov\{Z_n^{11.21}(t), Z_n^{11.22}(t)\}$ follow similarly. To begin, we define the covariance under the null hypothesis,

$$\begin{aligned} \sigma_{12}(t) &= n^{-1}cov\{Z_n^{11.12}(t), Z_n^{11.21}(t)\} \\ &= n^{-1}cov\left[\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \left\{ \frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}_{12}(s)}{\bar{Y}_{12}(s)} \right\}, \right. \\ &\qquad \left. \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{21}(s)} \left\{ \frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}_{21}(s)}{\bar{Y}_{21}(s)} \right\} \right] \\ &= n^{-1}cov\left[\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \left\{ \frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{12}(s)}{\bar{Y}_{12}(s)} \right\}, \right. \end{aligned}$$

$$\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} \left\{ \frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{21}(s)}{\bar{Y}_{21}(s)} \right\} \Bigg]. \tag{3.6}$$

Distributing the terms and further simplifying equation (3.6) using martingale properties,

$$\begin{aligned}
\sigma_{12}(t) &= n^{-1} cov \left\{ \int_0^t \frac{\bar{Y}_{12}(s)d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} - \int_0^t \frac{\bar{Y}_{11}(s)d\bar{M}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}, \right. \\
&\qquad \left. \int_0^t \frac{\bar{Y}_{21}(s)d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} - \int_0^t \frac{\bar{Y}_{11}(s)d\bar{M}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} \right\} \\
&= n^{-1} cov \left\{ \int_0^t \frac{\bar{Y}_{12}(s)d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}, \int_0^t \frac{\bar{Y}_{21}(s)d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} \right\} \\
&\quad - n^{-1} cov \left\{ \int_0^t \frac{\bar{Y}_{12}(s)d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}, \int_0^t \frac{\bar{Y}_{11}(s)d\bar{M}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} \right\} \\
&\quad - n^{-1} cov \left\{ \int_0^t \frac{\bar{Y}_{11}(s)d\bar{M}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}, \int_0^t \frac{\bar{Y}_{21}(s)d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} \right\} \\
&\quad + n^{-1} cov \left\{ \int_0^t \frac{\bar{Y}_{11}(s)d\bar{M}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}, \int_0^t \frac{\bar{Y}_{11}(s)d\bar{M}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} \right\} \\
&= n^{-1} E \int_0^t \frac{\bar{Y}_{12}(s)\bar{Y}_{21}(s) cov\{d\bar{M}_{11}(s), d\bar{M}_{11}(s) \mid \mathcal{F}(s_-)\}}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}} \\
&\quad - n^{-1} E \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s) cov\{d\bar{M}_{12}(s), d\bar{M}_{11}(s) \mid \mathcal{F}(s_-)\}}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}}.
\end{aligned} \tag{3.7}$$

In the intermediate steps to reach equation (3.7), we have used the fact that the pairs of strategies (A_1B_1, A_2B_1) and (A_1B_2, A_2B_1) are separate path, or that $cov\{d\bar{M}_{11}(s), d\bar{M}_{21}(s)\} = cov\{d\bar{M}_{12}(s), d\bar{M}_{21}(s)\} = 0$. By expanding the weighted martingales using $d\bar{M}_{jk}(t) = \sum_{i=1}^n W_{jki}(t)dM_{jki}(t)$, the covariances of interest can be expressed as expectations of integrals with respect to the filtration, or history up to time s defined in Section 2.6, such that $cov\{d\bar{M}_{11i}(s), d\bar{M}_{11i}(s)|\mathcal{F}(s_-)\} = \sum_{i=1}^n W_{11i}^2(s)Y_i(s)d\Lambda_0(s)$ and $cov\{d\bar{M}_{12i}(s), d\bar{M}_{11i}(s)|\mathcal{F}(s_-)\} = \sum_{i=1}^n X_i\{1-R_i(s)\}Y_i(s)d\Lambda_1^{NR}(s)/\phi^2$. Using derivations similar to the one used to derive the covariance between the increments of the martingales for strategies A_1B_1 and A_1B_2 (Section 2.7), we find

$$\begin{aligned}
\sigma_{12}(t) &= n^{-1} E \left[\int_0^t \frac{\bar{Y}_{12}(s)\bar{Y}_{21}(s)}{\omega_{12.21}(s)} \sum_{i=1}^n W_{11i}^2(s)Y_i(s)d\Lambda_0(s) \right. \\
&\qquad \left. - \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\omega_{12.21}(s)} \frac{1}{\phi^2} \sum_{i=1}^n X_i\{1-R_i(s)\}Y_i(s)d\Lambda_1^{NR}(s) \right] \\
&= n^{-1} E \left[\int_0^t \frac{\bar{Y}_{21}(s)}{\omega_{12.21}(s)} \left\{ \bar{Y}_{12}(s) \sum_{i=1}^n W_{11i}^2(s)Y_i(s)d\Lambda_0(s) - \frac{1}{\phi^2}\bar{Y}_{11}(s)Y_1^{NR}(s)d\Lambda_1^{NR}(s) \right\} \right],
\end{aligned}$$

where $\omega_{12.21}(s) = \{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s) + \bar{Y}_{21}(s)\}$. Similarly,

$$\sigma_{13}(t) = n^{-1}E\left[\int_0^t \frac{\bar{Y}_{22}(s)}{\omega_{12.22}(s)}\left\{\bar{Y}_{12}(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s)d\Lambda_0(s) - \frac{1}{\phi^2}\bar{Y}_{11}(s)Y_1^{NR}(s)d\Lambda_1^{NR}(s)\right\}\right] \tag{3.8}$$

$$\sigma_{23}(t) = n^{-1}E\left[\int_0^t \frac{1}{\omega_{21.22}(s)}\left\{\bar{Y}_{21}(s)\bar{Y}_{22}(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s)d\Lambda_0(s) + \frac{1}{(1-\phi)^2}\bar{Y}_{11}^2(s)Y_2^{NR}(s)d\Lambda_2^{NR}(s)\right\}\right], \tag{3.9}$$

where $\omega_{jk.j'k'}(s) = \{\bar{Y}_{11}(s) + \bar{Y}_{jk}(s)\}\{\bar{Y}_{11}(s) + \bar{Y}_{j'k'}(s)\}$.

By substituting $d\Lambda_0(s)$ and $d\Lambda_j^{NR}(s)$ with their estimates $d\hat{\Lambda}_0(s) = dN(s)/Y(s)$ and $d\hat{\Lambda}_j^{NR}(s) = dN_j^{NR}(s)/Y_j^{NR}(s)$, we have the consistent estimators s_{12}, s_{13}, s_{23} given below:

$$s_{12}(t) = n^{-1}\int_0^t \frac{\bar{Y}_{21}(s)}{\omega_{12.21}(s)}\left\{\bar{Y}_{12}(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s)d\hat{\Lambda}_0(s) - \frac{1}{\phi^2}\bar{Y}_{11}(s)Y_1^{NR}(s)d\hat{\Lambda}_1^{NR}(s)\right\} \tag{3.10}$$

$$s_{13}(t) = n^{-1}\int_0^t \frac{\bar{Y}_{22}(s)}{\omega_{12.22}(s)}\left\{\bar{Y}_{12}(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s)d\hat{\Lambda}_0(s) - \frac{1}{\phi^2}\bar{Y}_{11}(s)Y_1^{NR}(s)d\hat{\Lambda}_1^{NR}(s)\right\} \tag{3.11}$$

$$s_{23}(t) = n^{-1}\int_0^t \frac{1}{\omega_{21.22}(s)}\left\{\bar{Y}_{21}(s)\bar{Y}_{22}(s)\sum_{i=1}^{n} W_{11i}^2(s)Y_i(s)d\hat{\Lambda}_0(s) + \frac{1}{(1-\phi)^2}\bar{Y}_{11}^2(s)Y_2^{NR}(s)d\hat{\Lambda}_2^{NR}(s)\right\}, \tag{3.12}$$

where $\omega_{jk.j'k'}(s) = \{\bar{Y}_{11}(s) + \bar{Y}_{jk}(s)\}\{\bar{Y}_{11}(s) + \bar{Y}_{j'k'}(s)\}$.

The vector of weighted log-rank statistics, $n^{-1/2}Z_n^{MW}(t)$, presented in Section 3.2, converges in distribution under the null hypothesis to a trivariate normal distribution with mean zero and variance covariance matrix $\Sigma(t)$, where $\Sigma(t)$ is estimated by equation (3.2). Using the unbiased and consistent estimators of $\Sigma(t)$, by multivariate Slutsky's theorem, we have $n^{-1}Z_n^{MW}(t)^T\hat{\Sigma}^{-1}(t)Z_n^{MW}(t)$ converges in distribution under the null hypothesis to a chi-square distribution with three degrees of freedom.

The weighted log-rank test statistic comparing overall survival distributions for adaptive treatment strategies A_jB_k, $j,k = 1,2$, is then expressed in the form

$$T_n^{MW}(L) = n^{-1} Z_n^{MW}(L)^T \hat{\Sigma}^{-1}(L) Z_n^{MW}(L), \tag{3.13}$$

where L is some time less than the maximum follow-up time. The level α weighted log-rank test rejects the overall equality of adaptive treatment strategies' cumulative hazards when $T_n^{MW}(L) \geq \chi^2_{\alpha;\,3}$ where $\chi^2_{\alpha;\,3}$ is the $(1-\alpha)^{th}$ quantile of a chi-square distribution with three degrees of freedom.

3.4 SIMULATION STUDIES

Please refer to Section 2.8.1 for details on the data generation process. For each generated dataset we conducted the weighted log-rank test described in Section 3.2 to test the hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t) = \Lambda_0(t)$. We report the estimated type I error (proportion of samples for which the hypothesis was falsely rejected) for all tests in Table 3.1 when H_0 was true, and the estimated power (proportion of samples for which the hypothesis was correctly rejected) for all tests in Table 3.3. We have compared the proposed weighted log-rank statistic (WLR), $T_n^{MW}(L)$ from equation (3.13), to the standard unweighted log-rank statistic (SLR).

3.4.1 Simulation from Null Distribution

To investigate the performance of the weighted log-rank statistic under the null hypotheses, we generated 5000 datasets with the following parameters: $\theta_1^R = \theta_2^R = 0.5$ and $\theta_{11}^{RE} = \theta_{12}^{RE} = \theta_{21}^{RE} = \theta_{22}^{RE} = 1$. With a 40% response rate, $\mu_1^{NR} = \mu_2^{NR} = 0.91$ and the censoring paramter v was set to 3.80 and 2.00 and with a 60% response rate, $\mu_1^{NR} = \mu_2^{NR} = 0.56$, v was set to 3.60 and 1.85 to produce about 30% and 50% censoring, respectively.

Table 3.1 presents the estimated type I error rates for testing the null hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t) = \Lambda_0(t)$ under several combinations of sample size,

Table 3.1: Type I error rate under the null hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$

Response		30% censoring		50% censoring	
Rate (%)	n	WLR	SLR	WLR	SLR
40	250	0.056	0.045	0.060	0.044
	500	0.052	0.042	0.053	0.047
	1000	0.053	0.039	0.055	0.042
50	250	0.057	0.038	0.058	0.040
	500	0.056	0.042	0.053	0.044
	1000	0.055	0.039	0.052	0.039
60	250	0.059	0.040	0.057	0.042
	500	0.049	0.040	0.062	0.043
	1000	0.053	0.035	0.051	0.039

The target type I error rate is $\alpha = 0.05$. WLR is the weighted log-rank statistic in equation (3.13). SLR denotes the standard unweighted log-rank statistic.

response rates and censoring for the proposed weighted log-rank test and the standard unweighted log-rank test. The type I error rates for both statistics were similar across all combinations and around the 0.05 nominal level. Specifically, for a sample size of 500, the WLR test for 40% responders and 30% censoring produced an estimated type I error rate of 0.052 while the SLR test produced an estimated error rate of 0.042. We note that the SLR test was not as conservative when comparing four adaptive treatment strategies as when comparing only two. When censoring was increased to 50%, the WLR test produced an estimated type I error rate of 0.053 and the SLR test yielded 0.047. Increasing the response rate to 60% produced acceptable type I error rates for all sample sizes of greater than 250, but with similar results of about equal estimated type I error rates for all sample sizes and censoring combinations for both the WLR and the SLR tests.

3.4.1.1 Comparison of Estimated Variance-Covariance with Monte Carlo Results Under the null hypothesis, we have compared the mean estimated variance-covariance matrix to the Monte Carlo variance-covariance matrix for response rates of 40% and 60%

Table 3.2: Mean of the estimated variance-covariance matrix versus the Monte Carlo estimated variance-covariance matrix

Censoring Rate	Response Rate	n	Estimated Var-Cov			Monte Carlo Var-Cov		
30	40	250	0.447	0.217	0.217	0.447	0.199	0.207
				0.894	0.669		0.892	0.669
					0.895			0.886
		500	0.456	0.225	0.225	0.452	0.227	0.218
				0.910	0.681		0.971	0.706
					0.910			0.898
		1000	0.462	0.229	0.229	0.455	0.243	0.238
				0.918	0.686		0.942	0.710
					0.917			0.938
	60	250	0.653	0.318	0.318	0.658	0.320	0.320
				0.985	0.653		0.995	0.683
					0.985			1.040
		500	0.668	0.329	0.329	0.663	0.344	0.342
				1.004	0.665		1.044	0.694
					1.004			1.031
		1000	0.675	0.335	0.335	0.674	0.313	0.312
				1.014	0.674		1.000	0.666
					1.014			1.021
50	40	250	0.275	0.134	0.134	0.273	0.127	0.126
				0.625	0.488		0.630	0.491
					0.626			0.622
		500	0.278	0.138	0.138	0.284	0.146	0.141
				0.634	0.494		0.678	0.522
					0.633			0.645
		1000	0.281	0.140	0.140	0.285	0.146	0.143
				0.637	0.496		0.645	0.505
					0.637			0.643
	60	250	0.385	0.189	0.189	0.387	0.192	0.195
				0.664	0.470		0.684	0.493
					0.664			0.704
		500	0.391	0.193	0.193	0.397	0.200	0.201
				0.674	0.476		0.703	0.497
					0.673			0.698
		1000	0.393	0.196	0.196	0.388	0.185	0.179
				0.677	0.479		0.675	0.475
					0.677			0.673

Var-Cov denotes Variance-Covariance.

with censoring of 30% and 50% for sample sizes 250, 500, and 1000 in Table 3.2. The variance-covariance matrices in Table 3.2 are as follows:

$$\begin{pmatrix} var(n^{-1/2}Z_{11.12}) & cov(n^{-1/2}Z_{11.12}, n^{-1/2}Z_{11.21}) & cov(n^{-1/2}Z_{11.12}, n^{-1/2}Z_{11.22}) \\ & var(n^{-1/2}Z_{11.21}) & cov(n^{-1/2}Z_{11.21}, n^{-1/2}Z_{11.22}) \\ & & var(n^{-1/2}Z_{11.22}) \end{pmatrix}$$

From this table we can see that the under the null hypothesis, the mean estimates of s_{12} and s_{13} are the same and the mean estimates of s_{22} and s_{33} are almost identical as expected. The largest absolute difference occurs when there are 40% responders, 30% censoring, and a sample size of 500. This difference in the mean estimated variance, s_{22}, as compared to the Monte Carlo variance is 0.061. Interestingly, the biggest differences (all less than 0.061) in the variance-covariance estimates occur in the variance calculations, not in the covariance calculations. This is strong evidence to support our covariance derivation and the use of the weighted log-rank statistic to compare two shared-path or multiple ATS.

3.4.2 Simulation from Alternative Distributions

Since the type I error rates were upheld, we explored a variety of scenarios performing 5000 iterations to test the power of the weighted log-rank test at a sample size of 250. Data were generated from populations under the alternative hypotheses where at least one of the cumulative hazards of the four adaptive treatment strategies was not equal to the others. The four true survival distributions under alternative hypotheses, designated as scenarios (a)-(d), are plotted in Figure 3.1 when 60% of the population respond to A_j, $j = 1, 2$. Scenario (a) represents a typical alternative distribution of survival curves where all four curves differ ($\mu_1^{NR} = \theta_1^R = 1$, $\mu_2^{NR} = 1.25$, $\theta_2^R = 0.5$, $\theta_{11}^{RE} = 2$, $\theta_{12}^{RE} = 3.33$, $\theta_{21}^{RE} = 1.11$, $\theta_{22}^{RE} = 0.67$). Scenario (b) represents four survival curves where the shared-path strategies have vastly different survival ($\mu_1^{NR} = \theta_1^R = \theta_{11}^{RE} = 1$, $\mu_2^{NR} = 1.11$, $\theta_2^R = 1.67$, $\theta_{12}^{RE} = \theta_{21}^{RE} = 3.33$, $\theta_{22}^{RE} = 0.5$). Scenario (c) respresents survival curves where one strategy, A_2B_1, dominates the other strategies ($\mu_1^{NR} = \mu_2^{NR} = 1$, $\theta_1^R = \theta_2^R = \theta_{21}^{RE} = 2$, $\theta_{11}^{RE} = 0.67$, $\theta_{12}^{RE} = 0.5$, $\theta_{22}^{RE} = 0.4$). Finally, scenario (d) represents intersecting survival curves violating the proportional hazards assumption under which the log-rank statistic is optimal ($\mu_1^{NR} = \theta_2^R = \theta_{22}^{RE} = 1.43$,

$\mu_2^{NR} = 0.33$, $\theta_1^R = 0.2$, $\theta_{11}^{RE} = 1$, $\theta_{12}^{RE} = 1.67$, $\theta_{21}^{RE} = 0.2$). The censoring parameter v was set to 5 for scenarios (a) and (b), 6 for scenario (c), and 5.5 for scenario (d) so that censoring ranged from 23-41%.

Table 3.3 presents the power for comparing the survival distributions of the four adaptive treatment strategies. The WLR was compared to the SLR test. Again, in all cases, increasing the response rate from 40% to 60% increased the power of both statistics. In almost all of the scenarios tested, the WLR test had greater power to correctly reject the null hypothesis. Specifically, in scenario (b) where there is a pairing of curves, we see that for a 40% response rate and about 34% censoring, the WLR test had a high power at 0.985 unlike the SLR test which had power of 0.311, even though the survival distributions of A_1B_1 and A_2B_2 were very similar and so were the survival distributions of A_1B_2 and A_2B_1. $\mu_1^{NR} = 0.1$, $\mu_2^{NR} = \theta_2^R = \theta_{11}^{RE} = \theta_{22}^{RE} = 1$, $\theta_1^R = 2.5$, $\theta_{12}^{RE} = 0.5$, $\theta_{21}^{RE} = 0.33$, and $v = 4$, the WLR test had less power than the SLR test for 40% and 45% responders. This may be due to the unequal percentage of responders being censored compared to non-responders. For this and similar scenarios, as the percentage of responders increased above 50%, the WLR test almost always performed better with higher power than the SLR test.

In conclusion, the proposed weighted log-rank statistic maintained type I error in sample sizes as small as 250 with 30-50% patients censored. It also exhibited greater power when comparing multiple adaptive treatment strategies in most situations, including cases where the proportional hazards assumption was violated.

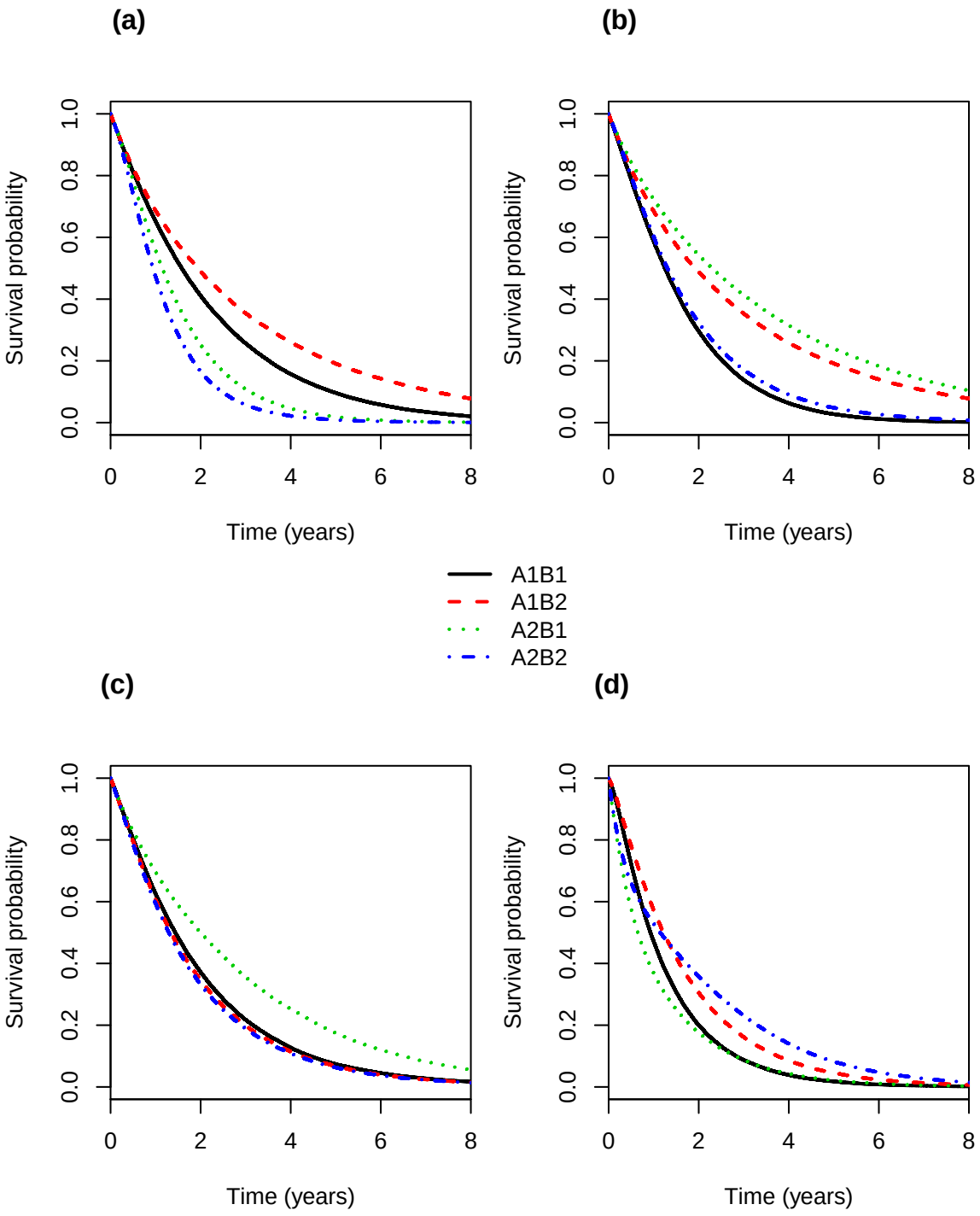


Figure 3.1: Survival curves for treatment strategies A_1B_1 (solid), A_1B_2 (dashes), A_2B_1 (dots), A_2B_2 (dot-dash), under different alternative hypotheses scenarios for 60% responders.

Table 3.3: Power against alternatives under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$ for a sample size of n=250

Scenario	Response Rate (%)	Censoring Rate (%)	Power WLR	Power SLR
(a)	40	31	0.613	0.197
	50	33	0.857	0.457
	60	35	0.973	0.770
(b)	40	34	0.985	0.311
	50	38	0.995	0.516
	60	41	0.999	0.752
(c)	40	28	0.663	0.085
	50	30	0.764	0.118
	60	33	0.833	0.178
(d)	40	23	0.993	0.980
	50	24	0.987	0.855
	60	26	0.980	0.614

See Figure 3.1 and Section 3.4.2 for a description of alternative survival scenarios (a)-(d). WLR is the weighted log-rank statistic in equation (3.13). SLR denotes the standard unweighted log-rank statistic.

4.0 DATA ANALYSIS

We applied the weighted log-rank test statistic to compare overall survival of the adaptive treatment strategies from the Children's Cancer Group high-risk neuroblastoma study reported by Matthay et al. (1999, 2009). This two-stage randomized trial began in 1991 and ended in 1996 with 539 eligible children ages 1-18 years with newly diagnosed high-risk neuroblastoma (the most common extracranial solid tumor of childhood). All of the patients were initially treated with chemotherapy and 379 patients without progressive disease participated in the first-stage randomization. Patients were assigned to chemotherapy (n=190) or to ABMT, a combination of myeloablative chemotherapy, total-body irradiation, and transplantation of autologous bone marrow purged of cancer cells (n=189). Patients without disease progression (and who consented to further treatment) participated in the second-stage randomization. Of the 203 patients who were eligible for the second-stage randomization, 102 were assigned to receive treatment of 13-cis-retinoic acid (cis-RA) and the other 101 patients were assigned not to receive any further treatment.

To clarify the treatment strategies and SMART design utilized in this trial, refer to Figure 4.1. We are interested in comparing the following four treatment strategies: (i) CR: Treat with chemotherapy followed by cis-RA if there is no disease progression; (ii) CN: Treat with chemotherapy and if there is no disease progression, do not continue treatment; (iii) AR: Treat with ABMT followed by cis-RA if there is no disease progression; (iv) AN: Treat with ABMT and if there is no disease progression, do not continue treatment. Notice that there are 85 patients who do not respond to the first-stage treatment of chemotherapy and are therefore consistent with shared-path adaptive treatment strategies CR and CN and 91 patients who do not respond to first-stage treatment of ABMT and are therefore consistent with shared-path

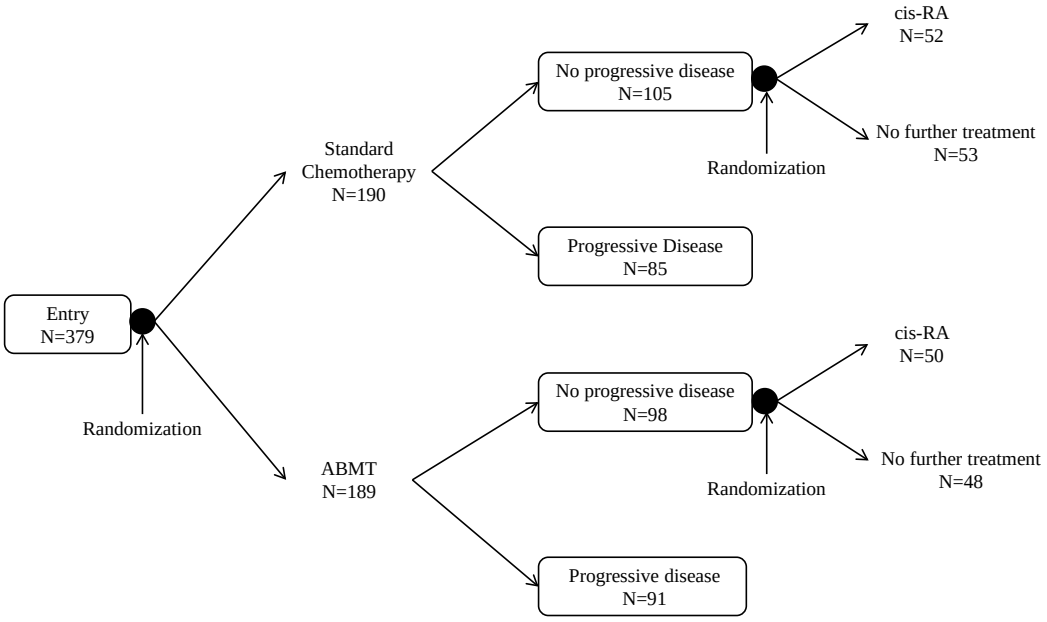


Figure 4.1: Diagram of the SMART design in the Children's Cancer Group high-risk neuroblastoma study

adaptive treatment strategies AR and AN. The goal was to compare survival distributions under these four adaptive treatment strategies. Survival distributions in Figure 4.2 were created using the weighted risk set estimator for the survival function from Guo and Tsiatis (2005).

In the main findings of the study, separate analyses for the first- and second-stage treatments were reported, ignoring the induction or maintenance treatments while conditioning on patients who were eligible to receive second-stage treatments. Initially, for three-year event-free survival, Matthay et al. (1999) reported the superiority of ABMT over chemotherapy for the first stage treatment and the superiority of cis-RA over no further treatment in the second stage. In 2009, Matthay et al. reported that ABMT significantly improved the five year event-free and overall survival compared to non-myeloablative chemotherapy, and cis-RA or transplantation improved overall survival compared to no further therapy. Analyzing this data by considering second-stage randomization and using adaptive treatment strategies, however, demonstrated no significant improvements in overall survival.

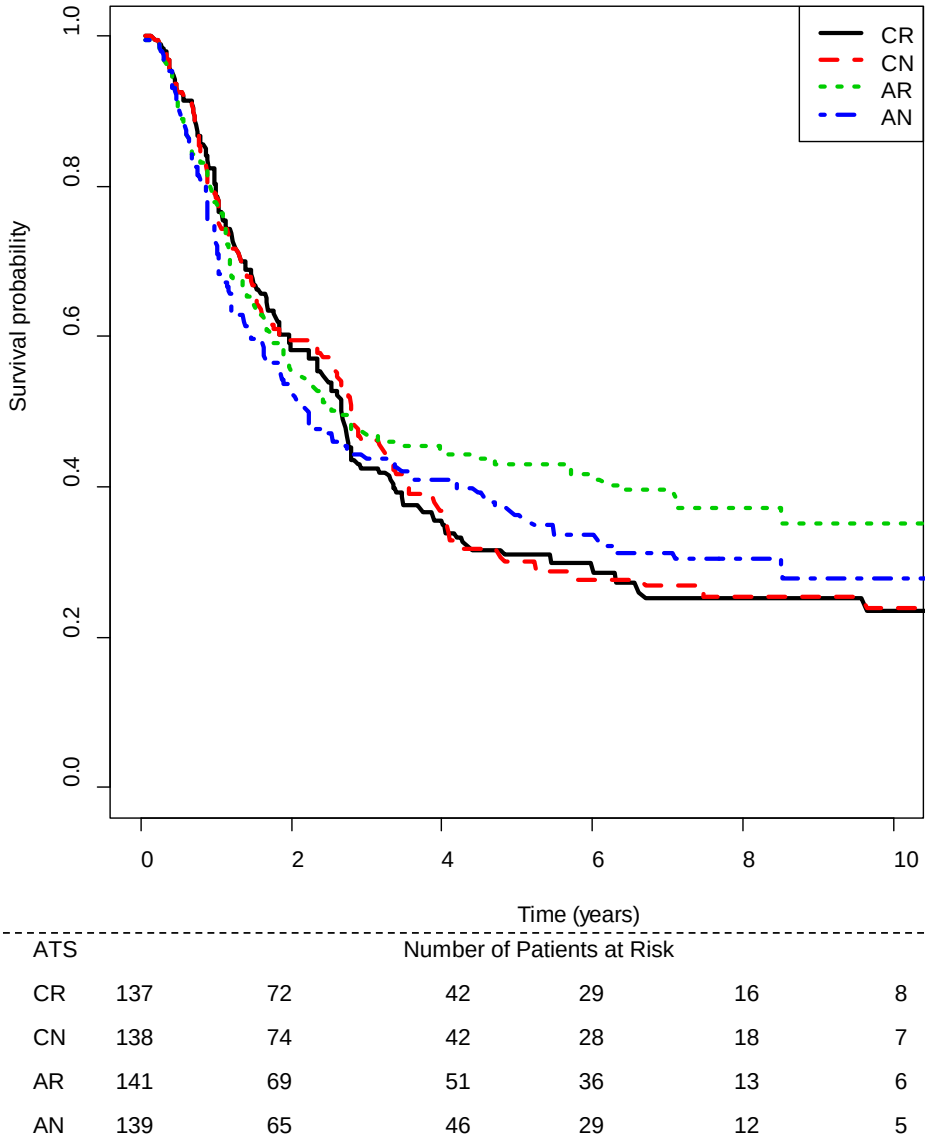


Figure 4.2: Weighted survival curves under four treatment strategies in the neuroblastoma study with the number of patients at-risk for each strategy. CR (solid): 'Treat with chemotherapy followed by cis-RA if there is no disease progression'; CN (dashes): 'Treat with chemotherapy and if there is no disease progression, do not continue treatment'; AR (dots): 'Treat with ABMT followed by cis-RA if there is no disease progression'; AN (dot-dash): 'Treat with ABMT and if there is no disease progression, do not continue treatment'

First we note that there was an overall response rate of 53.6%, where 55.3% of patients who received chemotherapy responded and 51.9% of patients who received ABMT responded. Overall, 31.4% of patients were censored, 27.4% of patients randomized to chemotherapy and 35.4% of patients randomized to ABMT. While censoring was similar among the groups by induction treatment, it differed substantially between non-responders and responders. Only 19.9% of non-responders to chemotherapy or ABMT were censored, whereas 41.4% of responders to chemotherapy or ABMT were censored.

To test if there was a significant difference in the hazards of treatment strategies which share the same initial treatment of chemotherapy (shared-path treatment strategies CR to CN), the WLR statistic was 0.12 with $p = 0.90$. This results agrees with Figure 4.2 as the weighted survival curves for those following CR and CN appear to be almost exactly the same. For comparing treatment strategies which share the same initial treatment of ABMT (shared-path treatment strategies AR to AN) the WLR statistic was -1.07 with $p = 0.29$, showing that the two strategies that start on ABMT are not significantly different. Figure 4.2 shows separation between these two curves; this separation, however, occurs after five years, where the number of patients following each strategy has dropped to less than 36 (14 or almost 50% of the patients in AR and AN were non-responders to ABMT). Thus, if there is a true difference in those treated with ABMT followed by cis-RA as compared to those only treated with ABMT, this study did not have enough patients (high enough power) to find this difference. To test if there was a difference in overall survival across the four strategies (CR, CN, AR, AN), the weighted log-rank statistic from equation (3.13) was computed. There was no significant difference in the overall survival of the four adaptive treatment strategies as the WLR test produced a chi-square statistic of 2.00 with $p = 0.57$. We note that the difference in censoring between those who responded to the first-line treatment and those who did not is a case of informative censoring. Alike many other survival methods, the weighted log-rank test may not be valid when censoring is informative.

5.0 SAMPLE SIZE

5.1 INTRODUCTION

Treatment is inherently a dynamic process, a conversation between the physician and patient, adapting to the patient's needs. Treatment often changes due to the patient's adherence, response, or side effects such that a patient often receives a string of treatments or dosages based on his or her individual behaviors and/or characteristics. Sequences of individually tailored treatments are referred to as adaptive treatment strategies (ATS) or dynamic treatment regimes. ATS are especially relevant in the treatment of chronic diseases where treatment goals include decreasing symptoms, preventing progression, and improving function. These diseases, such as cancer, AIDS, substance abuse, depression, and other mental health disorders, are generally complicated and heterogeneous, where a single treatment that "cures" all patients does not exist. Instead, conversations between physicians and their patients are imperative to find the best series of individualized treatments.

In order to advance treatment options for patients with chronic diseases and find optimal ATS, we must operationalize the construction and comparison of ATS. Strategies can be constructed and compared using observational data, but in order to more easily control for confounding and evaluate ATS, randomized trials are necessary. The sequential multiple assignment randomized trial (SMART) (Lavori et al., 2000, 2004, Murphy, 2005) was developed to include the dynamic aspect of treatment prescription into randomized trials investigating sequences of time-varying treatments subject to modification based on individuals' intermediate behaviors, characteristics, or responses. In this dissertation, we focus on two-stage SMART designs.

SMARTs have been implemented throughout the recent past (Strecher et al., 2008, Pelham and Fabiano, 2008, Auyeung et al., 2009, Kasari, 2009, Jones, 2010, Tannir, 2010, Mitchell et al., 2011, Wang et al., 2011), especially in areas of substance abuse, chronic diseases, and mental health disorders. Earlier trials were not equipped with the statistical techniques to compare adaptive treatment strategies and instead compared stage-specific treatments. More recent trials have been designed and analyzed with the goal of comparing ATS since statistical literature has increased in this area to address issues related to the design, comparison, and estimation of ATS (Thall et al., 2000, Murphy, 2003, 2005, Dawson and Lavori, 2004, Wahed and Tsiatis, 2006, Wahed, 2010, Orellana et al., 2010, Zhao et al., 2011, Almirall et al., 2012). Such trials and statistical theory have focused on analyzing continuous, binary, and/or time-to-event outcomes, with most statistical tests comparing two two-stage ATS with different initial treatments (allowing the groups to be independent). We refer to these independent strategies, as separate-path ATS. On the contrary, strategies are shared-path if patients following one strategy share a common path of treatment with patients following another strategy and, hence, the groups share the same first-stage treatment and are not statistically independent (for detailed definitions, please refer to Section 2.2).

Many have focused on the analysis aspect of ATS in observational and randomized settings, yet, we have previously noted a gap in statistical analysis literature focusing on the overall survival comparison of shared-path ATS (Section 1.4). Comparatively, relatively less development occurred in the design aspect of SMARTs, and thus, we find such a gap also existing in the statistical design literature. Specifically, there does not exist sample size determination techniques for trial designs interested in the overall survival comparison of multiple ATS. As SMARTs are increasingly implemented, it is essential that the necessary methods for designing these trials are developed. Chakraborty (2011, p. 42) declares the importance of sample size research, "As is the case with any study, sample size calculation is a crucial part of SMART design," and notes that "there are still open questions relating to sample size issues in a SMART design that warrant further research." Therefore, the goal of this section is to address an open sample size question in SMART design. Explicitly, we will present a sample size formula to adequately power a two-stage SMART with the aim of comparing the overall survival of multiple ATS, including shared-path ATS.

5.2 RELATED WORK

Most sample size formulations (Murphy, 2005, Feng and Wahed, 2008, Wolbers and Helterbrand, 2008, Dawson and Lavori, 2010, Oetting et al., 2011, Li and Murphy, 2011) have focused on the comparison of two two-stage separate-path ATS since most analyses have focused on this comparison. Apart from these comparisons, Feng and Wahed (2009), Dawson and Lavori (2010) and Li and Murphy (2011), have addressed methods for the comparison of two shared-path ATS in the presence of continuous and time-to-event outcomes. These methods, however, did not address the overall survival comparison for multiple ATS, including those which are shared-path. We provide a brief overview of sample size methods to power SMARTs. First, we introduce sample size methods for trials with binary outcomes, followed by trials with continuous outcomes, and finally, trials with time-to-event outcomes. Our interest lies in the latter trials; to extend the methodology to power SMARTs with time-to-event outcomes to compare multiple ATS.

Often, mental health trials will have binary outcomes such that the patient responds successfully or unsuccessfully to the treatment or treatment strategy. For example, the outcome for a depression trial may be dichotomous such that a successful outcome is defined by a depression score on the Beck Depression Inventory (BDI) of 12 or under, whereas, an unsuccessful outcome is defined by a BDI score over 12. Dawson and Lavori (2004) focused on these binary outcomes in psychopharmacology trials and developed sample size equations to compare the outcomes from patients following an adaptive treatment strategy to patients who only received a single treatment (the single treatment corresponding to a treatment chosen from a single stage of the adaptive treatment strategy). This comparison is similar to equivalence testing of two treatments.

Alternatively, outcomes could be continuous, e.g., using the depression score itself, CD4 counts for AIDS, or school performance scores for children with ADHD. Murphy (2005) focused on such outcomes, powering SMARTs primarily to develop ATS, not to confirm the superiority of one strategy over another. Murphy introduced a sample size formula to compare two separate-path ATS. Simple upper bounds of the variance calculation were given, so that the user could input this bound, the desired type I and II error, and effect

size (signal to noise ratio) to calculate the necessary sample size. Oetting et al. (2011) expanded on Murphy's work focusing on hypothesis testing and the estimation of stage-specific and conditional treatment effects. Among the methods in this paper were two sample size equations for comparing two separate-path ATS based on Z-statistics. For both equations the user must provide the desired effect size, but one also demands the initial treatment response rate, whereas the other does not.

Dawson and Lavori (2010) developed sample size methods for pairwise comparisons of both separate-path and shared-path ATS with continuous outcomes. For separate-path ATS, the method utilized a variance inflation factor to account for the loss of precision due to missingness created by the sequential randomization in a SMART design. Dawson and Lavori (2010) denoted shared-path ATS as "overlapping strategies," explaining that "any overlap between a pair of ATS (created by sequential treatment assignment) not only diminishes causal difference but also introduces positive between-strategy covariance." They recognized that a "significant challenge is the development of methods for sample size determination because of the sequential and adaptive nature of both the strategies under study and treatment assignment mechanism used to assign subjects to the adaptive treatment strategy." To address this challenge, they provided sample size formulae for multi-stage or k-stage designs using a version of the G-Computational Algorithm (Robins et al., 1994). For a multi-stage design, they suggest that the required sample size be set to the maximum sample size needed from any pairwise comparison.

Other sample size methodologies for ATS have focused on time-to-event outcomes, for example, time to death, time to a certain number of alcoholic drinks, or time to a first school disciplinary event. Wolbers and Helterbrand (2008) attempted to make methods as simple and applicable as possible, providing sample size equations to analyze two-stage designs based on comparisons using Cox regressions or log-rank tests. Instead of analyzing the overall survival of the entire ATS, they suggested comparing second-stage treatments directly applying Schoenfeld's formula, and then modified the formula by using simulation results to compare independent induction treatments for those who responded and received the same second-stage treatment.

In order to compare entire ATS with time-to-event outcomes, Li and Murphy (2011) proposed sample size formulae, focusing on the comparison of two separate-path ATS in both point-wise and overall survival settings. Their comparisons used the weighted Kaplan Meier and weighted log-rank tests. Conservative sample size calculations replaced the variance calculations with upper bounds to lessen distributional assumptions. The user must provide the first- and second-stage randomization probabilities, the probability of observing an event before the end of the study for patients following the first (reference) ATS, the hazard ratio of survival times for those following the second, independent ATS compared to the first (reference) ATS, the type I error, and power. An applet is available to calculate the sample size to power SMART with the goal of comparing two separate-path ATS at `http://methodologymedia.psu.edu/logranktest/samplesize`. In the supplementary material for this article, Li and Murphy also provide sample size formulae for more general two-stage randomized trials, including the test statistic to test the equality of survival probabilities at one time point using the weighted Kaplan-Meier estimator of two shared-path ATS. The covariance term in this statistic is based on the joint asymptotic distribution of the weighted Kaplan-Meier estimators of the shared-path ATS' survival functions, and it is estimated using an empirical estimator.

Feng and Wahed (2009) also provided a sample size equation to compare point-wise survival of two shared-path ATS using a weighted sample proportion estimator. This formula required parametric specifications of the survival times for non-responders and responders, as well as, the censoring distribution. While addressing the comparison of shared-path ATS, Li and Murphy (2011) and Feng and Wahed (2009) did not provide methods for the overall survival comparison of shared-path ATS. Feng and Wahed (2008) addressed the overall survival comparison, but for two separate-path ATS. They provided a sample size equation based on the supremum weighted log-rank test. This formula did not include parametric working assumptions, but rather depended on the expected proportions of death among responders and non-responders at the end of the study and the hazard ratio between independent strategies.

While methodological research is focused mostly on the analysis aspect of SMARTs, several methods exist for the design aspect of these trials. Specifically, there are sample

size formulae to power SMARTs with binary, continuous, or time-to-event outcomes, but most methods focus on comparing two ATS, usually separate-path strategies. Thus, besides simulation, a method does not exist to calculate the sample size of a SMART with the goal of comparing overall survival of multiple ATS, some of which may be shared-path, in the time-to-event outcome setting.

In order to derive such a sample size equation, not only are methods from ATS literature relevant, but so are methods comparing overall survival by using multivariate log-rank tests from multi-arm randomized trial literature. A pivotal paper by Ahnn and Anderson (1995) presented sample size equations based on the Tarone-Ware and Harrington and Fleming test statistics comparing $k \geq 2$ survival distributions. The follow-up paper (Ahnn and Anderson, 1998) provided extensions to more complex designs, including non-proportional hazards, time-dependent loss to follow-up, non-compliance, drop-in, and varying accrual, drop-in and drop-out rates, using Markov models. Jung and Hui (2002) improved upon Ahnn and Anderson's formula showing that it underestimated sample size. Generalizing Ahnn and Anderson's sample size formula for unequal allocation, Halabi and Singh (2004) presented sample size calculations for stratified and unstratified log-rank tests. Here assumptions included the contiguous time-varying proportional hazards alternative, hazard ratios which do not depend on time, and equal censoring among treatment groups. They also used power series approximations to the non-central chi-square distribution of the log-rank test. Barthel et al. (2006) presented a general framework for sample size calculation, including staggered entry and loss to follow-up, implemented in the freely available program ART for STATA. The sample size calculation assumed local alternatives, but supplementary material approximated the formula for more distant alternatives using iterative methods.

Borrowing the local alternative framework and non-central chi-square distribution from the multi-arm randomized trial literature, and combining it with the sequential and adaptive nature of methods from ATS literature, we develop a sample size formula to compare multiple ATS, some of which may be shared-path, in simple two-stage SMART designs. We derive the sample size formula based on the statistic presented in Section 3. We then present simulation studies employing the sample size formula testing its empirical power and discuss the results, limitations, and planned work.

5.3 SAMPLE SIZE CALCULATION

We develop the sample size equation under the scenario discussed in Section 1.1 to design a trial similar to that shown in Figure 2.1. Thus, we are interested in constructing a trial with enough power to find a difference in the overall survival distributions of four adaptive treatment strategies A_1B_1, A_1B_2, A_2B_1, and A_2B_2. We have shown in Section 3.2 that under the null hypothesis of four equal cumulative hazards, the weighted log-rank test is asymptotically chi-square distributed with three degrees of freedom.

We define the alternative hypothesis H_1 to be a sequence of alternatives converging toward the null hypothesis as $n \to \infty$. We base the sample size equation on this contiguous time-varying proportional hazards alternative (Kosorok and Lin, 1999) wherein the hazard functions for the four ATS are

$$\begin{aligned}
\lambda_{11}^n(t) &= \lambda_0(t)e^{\gamma_{11}^*/n^{1/2}}, \\
\lambda_{12}^n(t) &= \lambda_0(t)e^{\gamma_{12}^*/n^{1/2}}, \\
\lambda_{21}^n(t) &= \lambda_0(t)e^{\gamma_{21}^*/n^{1/2}}, \\
\text{and } \lambda_{22}^n(t) &= \lambda_0(t)e^{\gamma_{22}^*/n^{1/2}}.
\end{aligned}$$

Note that $\lambda_0(t)$ is a continuous baseline hazard and γ_{jk}^* is a scalar constant.

We express the z-statistics, $Z_n^{11.jk}(t)$, $j,k = 1,2$, in terms of weighted martingales as follows:

$$\begin{aligned}
Z_n^{11.jk}(t) =& G_n(t) + R_n(t) \\
=& n^{-1/2} \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{jk}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{jk}(s)} \left\{ \frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{jk}(s)}{\bar{Y}_{jk}(s)} \right\} \qquad (5.1) \\
&+ n^{-1/2} \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{jk}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{jk}(s)} \{d\Lambda_{11}(s) - d\Lambda_{jk}(s)\}, \qquad (5.2)
\end{aligned}$$

since $\bar{M}_{jk}(s) = \bar{N}_{jk}(s) - \int_0^s \bar{Y}_{jk}(u)d\Lambda_{jk}(u)$. From the martingale central limit theorem, the difference in weighted martingales, $G_n(t)$, shown in equation (5.1), converges to mean 0. Then, under the alternative hypothesis, $Z_n^{11.jk}(t)$ reduces to $R_n(t)$ in equation (5.2). Under

the contiguous alternative hypothesis and using a Taylor's series expansion of $n^{1/2}\{d\Lambda_{11}^n(s) - d\Lambda_{jk}^n(s)\}$, $j,k = 1,2$, we have:

$$\begin{aligned} & n^{1/2}\{d\Lambda_{11}^n(s) - d\Lambda_{jk}^n(s)\} \\ &= n^{1/2}\{e^{\gamma_{11}^*/(n^{1/2})} - e^{\gamma_{jk}^*/(n^{1/2})}\}\lambda_0(s)ds \\ &= (\gamma_{11}^* - \gamma_{jk}^*)\lambda_0(s)ds\{1 + o(1/\sqrt{n})\}, \end{aligned} \tag{5.3}$$

where $o(1)$ is an error term that converges uniformly in s to 0 as $n \to \infty$.

Let X_{ji} be the first-stage treatment indicator such that $X_{1i} = 1$ if individual i receives treatment A_1 with probability ϕ_1, 0 otherwise, and $X_{2i} = (1 - X_{1i}) = 1$ if individual i receives treatment A_2 with probability $\phi_2 = (1 - \phi_1)$, 0 otherwise. Let π_{z_k} be the randomization probability to second-stage treatment, $\pi_{z_1} = \pi_z$ and $\pi_{z_2} = 1 - \pi_z$ and let π_j^R be the expected proportion of individuals who have responded to A_j by the end of the trial (time L). Then, the limiting values of the weighted at-risk process for strategy A_jB_k, $\bar{Y}_{jk}(s)$, found in the Z-statistics in equation (3.1) and variance equations (3.3)-(3.5), as well as, $\sum_{i=1}^n W_{jki}^2(s)Y_i(s)$ and $Y_j^{NR}(s)$ found in the variance equations (3.3)-(3.5) and covariance equations (3.10)-(3.12) can be approximated by their population counterparts. We define $\psi_j^{NR}(s)$ to be the limiting distribution of $\sum_{i=1}^n X_{ji}\{1 - R_i(s)\}Y_i(s)/\{n\phi_j(1 - \pi_j^R)\}$ and $\psi_{jk}^R(s)$ to be the limiting distribution of $\sum_{i=1}^n X_{ji}R_i(s)Z_{ki}Y_i(s)/(n\phi_j\pi_j^R\pi_{z_k})$. Then, we have

$$\begin{aligned} E[\bar{Y}_{jk}(s)/n] =& E\left[\sum_{i=1}^n W_{jki}Y_i(s)/n\right] \\ =& E\left[\frac{1}{n\phi_j}\sum_{i=1}^n X_{ji}\{1 - R_i(s)\}Y_i(s) + \frac{1}{n\phi_j\pi_{z_k}}\sum_{i=1}^n X_{ji}R_i(s)Z_{ki}Y_i(s)\right] \\ \approx& (1 - \pi_j^R)\psi_j^{NR}(s) + \pi_j^R\psi_{jk}^R(s). \end{aligned}$$

Similarly, we have

$$\begin{aligned} E\left[\sum_{i=1}^n W_{jki}^2(s)Y_i(s)/n\right] =& E\left[\frac{1}{n\phi_j^2}\sum_{i=1}^n X_{ji}\{1 - R_i(s)\}Y_i(s) + \frac{1}{n\phi_j^2\pi_{z_k}^2}\sum_{i=1}^n X_{ji}R_i(s)Z_{ki}Y_i(s)\right] \\ \approx& \frac{1}{\phi_j}\left[(1 - \pi_j^R)\psi_j^{NR}(s) + \frac{\pi_j^R}{\pi_{z_k}}\psi_{jk}^R(s)\right], \end{aligned}$$

and

$$E\left[Y_j^{NR}(s)/n\right] = E\left[\sum_{i=1}^{n} X_{ji}\{1-R_i(s)\}Y_i(s)/n\right]$$
$$\approx \phi_j(1-\pi_j^R)\psi_j^{NR}(s).$$

For simplicity, we assume that response rates and censoring are similar in all strategies, such that $\pi_1^R = \pi_2^R = \pi_0^R$, and $\psi_1^{NR}(s) = \psi_2^{NR}(s) = \psi_0^{NR}(s)$ and $\psi_{11}^R(s) = \psi_{12}^R(s) = \psi_{21}^R(s) = \psi_{22}^R(s) = \psi_0^R(s)$. We also assume that randomization probabilities to A_1 and A_2 and to B_1 and B_2 are equal, such that $\phi = \phi_1 = \phi_2 = 0.5$ and $\pi_z = \pi_{z_1} = \pi_{z_2} = 0.5$, respectively. Under these simplified assumptions,

$$E[\bar{Y}_{jk}(s)/n] \approx (1-\pi_0^R)\psi_0^{NR}(s) + \pi_0^R\psi_0^R(s),$$
$$E\left[\sum W_{jki}^2(s)Y_i(s)/n\right] \approx 2\{(1-\pi_0^R)\psi_0^{NR}(s) + 2\pi_0^R\psi_0^R(s)\},$$
$$\text{and } E\left[Y_j^{NR}(s)/n\right] \approx 0.5(1-\pi_0^R)\psi_0^{NR}(s).$$

Then, using equation (5.3), $R_n(t)$ in equation (5.2) can be written as

$$R_n(t) = n^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{jk}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{jk}(s)} n^{1/2}\{d\Lambda_{11}(s) - d\Lambda_{jk}(s)\}$$
$$\approx n^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{jk}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{jk}(s)}(\gamma_{11}^* - \gamma_{jk}^*)\lambda_0(s)ds,$$

which converges in probability to

$$\int_0^t 0.5\{\pi_0^R\psi_0^R(s) + (1-\pi_0^R)\psi_0^{NR}(s)\}(\gamma_{11}^* - \gamma_{jk}^*)\lambda_0(s)ds$$
$$= 0.5(\gamma_{11}^* - \gamma_{jk}^*)\{\pi_0^R D^R(t) + (1-\pi_0^R)D^{NR}(t)\}$$

where

$$D^R(t) = \int_0^t dD^R(s) = \int_0^t \psi_0^R(s)\lambda_0(s)ds,$$
$$\text{and } D^{NR}(t) = \int_0^t dD^{NR}(s) = \int_0^t \psi_0^{NR}(s)\lambda_0(s)ds.$$

Here, under the null distribution, $D^R(t)$ is the probability of observing an event by time t from individuals who have responded to A_j and received B_k and $D^{NR}(t)$ is the probability of observing an event by time t from individuals who have not responded to A_j.

We assume that the asymptotic arguments apply for reasonably large sample size n and fixed alternative, $\gamma_{11.jk}$, when comparing strategy A_1B_1 with strategy A_jB_k, so like many sample size formulas, power is based on a fixed alternative. We approximate power calculations under the contiguous alternative using $\gamma^*_{11} - \gamma^*_{jk} = n^{1/2}\gamma_{11.jk}$. Then, our vector of Z-statistics, $n^{-1/2}Z_n^{WLR}(t) = n^{-1/2}\{Z_n^{11.12}(t), Z_n^{11.21}(t), Z_n^{11.22}(t)\}^T$, each with expressions from equation (3.1), can be approximated as $n^{1/2}\mu = n^{1/2}\{\mu^{11.12}(t), \mu^{11.21}(t), \mu^{11.22}(t)\}$

$$\mu^{11.12}(t) = 0.5\gamma_{11.12}\{\pi_0^R D^R(t) + (1-\pi_0^R)D^{NR}(t)\} \tag{5.4}$$

$$\mu^{11.21}(t) = 0.5\gamma_{11.21}\{\pi_0^R D^R(t) + (1-\pi_0^R)D^{NR}(t)\} \tag{5.5}$$

$$\text{and } \mu^{11.22}(t) = 0.5\gamma_{11.22}\{\pi_0^R D^R(t) + (1-\pi_0^R)D^{NR}(t)\}. \tag{5.6}$$

Note that we can also express this fixed alternative as the negative log hazard ratio of the strategy A_jB_k to A_1B_1. Here, $-n^{1/2}log\{\lambda_{jk}/\lambda_{11}\} = \gamma^*_{11} - \gamma^*_{jk}$, thus we can approximate the Z-statistics from the user inputs of the hazard ratio of strategies A_jB_k as compared to A_1B_1, the expected proportion of responders, and the expected proportions of death among the non-responders and responders by the end of the study.

The variance equations from $\hat{\mathbf{\Sigma}}(t) = \{s_{pq}(t)\}^{3x3}$, where the elements are defined in equations (3.3)-(3.5) and (3.10)-(3.12), are similarly estimated by the matrix $\xi = \{\xi_{pq}(t)\}^{3x3}$ with elements estimated as follows:

$$\xi_{11}(t) = 2\pi_0^R D^R(t) + (1-\pi_0^R)\pi_0^R D^{NR}(t), \tag{5.7}$$

$$\xi_{22}(t) = 2\pi_0^R D^R(t) + (1-\pi_0^R)D^{NR}(t), \tag{5.8}$$

$$\xi_{33}(t) = 2\pi_0^R D^R(t) + (1-\pi_0^R)D^{NR}(t), \tag{5.9}$$

$$\xi_{12}(t) = \pi_0^R D^R(t) + 0.5(1-\pi_0^R)\pi_0^R D^{NR}(t), \tag{5.10}$$

$$\xi_{13}(t) = \pi_0^R D^R(t) + 0.5(1-\pi_0^R)\pi_0^R D^{NR}(t), \tag{5.11}$$

$$\text{and } \xi_{23}(t) = \pi_0^R D^R(t) + 0.5(1-\pi_0^R)(2-\pi_0^R)D^{NR}(t). \tag{5.12}$$

Note that we have approximated $d\Lambda_j^{NR}(s)$, $j = 1, 2$, from the covariance estimations, with $(1-\pi_0^R)d\Lambda_0(s)$.

Assuming the above conditions, it can be shown that under a sequence of alternatives converging toward the null hypothesis, the weighted log-rank statistic has an asymptotic chi-square distribution with 3 degrees of freedom and non-centrality parameter $v = n\mu\xi^{-1}\mu^T$. Approximations of μ and ξ depend on equal censoring and randomization probabilities. Users must provide parameter inputs using a priori knowledge. Specifically, users of this sample size calculation must provide:

- The desired type I error: α
- The desired power level: $1 - \beta$
- The expected proportion of responders by the end of the study (time L): π_0^R
- The expected proportion of death in non-responders by the end of the study: $D^{NR}(L)$
- The expected proportion of death in responders by the end of the study: $D^R(L)$
- The hazard ratios between strategies: $\lambda_{jk}/\lambda_{11}$

Using this information, $\mu\xi^{-1}\mu^T$ is calculated to obtain the non-centrality parameter of the corresponding chi-square distribution with specified type I error and power. Then, the sample size n is solved for, where $n = v/(\mu\xi^{-1}\mu^T)$. This sample size, n, is the total number of patients in all strategies. Using the randomization probabilities to A_j and B_k (assumed to be 0.5), we can find the strategy-specific sample sizes.

5.4 NUMERICAL STUDY

To verify the utility of our sample size formula, we have calculated sample sizes for various alternative hypotheses defined by the user-specified input as described in the previous section. We present the results from the sample size calculation in Table 5.1. This table presents the total sample sizes for comparing all four ATS, A_jB_k, $j, k = 1, 2$ under different alternatives. The desired type I error rate was set to be 0.05 and power to be 0.80. We have specified the hazard ratio between strategies A_1B_2 and A_1B_1 to be 1.1, between A_2B_1 and A_1B_1 to be 1.3, and we have varied the hazard ratio between A_2B_2 and A_1B_1 to be 1.2, 1.3, and 1.5. We have varied the response rate such that either 40% or 60% of patients respond to the initial

Table 5.1: Results from Sample Size Calculation

Response Rate	$D^R(L)$	$D^{NR}(L)$	n $\lambda_{22}/\lambda_{11} = 1.2$	$\lambda_{22}/\lambda_{11} = 1.5$	$\lambda_{22}/\lambda_{11} = 1.7$
40%	0.2	0.3	3178	1308	723
		0.5	2005	836	451
		0.7	1447	608	324
	0.4	0.3	1849	758	422
		0.5	1589	654	362
		0.7	1390	575	316
60%	0.2	0.3	3801	1526	885
		0.5	2656	1072	615
		0.7	2026	822	467
	0.4	0.3	2123	851	495
		0.5	1901	763	443
		0.7	1718	691	400

n is the total sample size for all strategies. Desired type I error was set to 0.05 and power set to 0.80. $\lambda_{jk}/\lambda_{11}$ is the hazard ratio of strategy A_jB_k to A_1B_1. $\lambda_{12}/\lambda_{11} = 1.1$ and $\lambda_{21}/\lambda_{11} = 1.3$. $D^R(L)$ and $D^{NR}(L)$ are the expected proportions of events among the responders and non-responders respectively, by the end of the study.

treatment A_j, $j = 1, 2$ and we've varied the proportions of events among the responders ($D^R(L)$) and non-responders ($D^{NR}(L)$).

The sample size formula does behave as expected such that the closer the hazard ratios are to each other and to the null, the larger the sample size. As the expected proportions of events among responders and non-responders increases, the sample size decreases. Also, as the difference between the expected proportions of events among the responders and non-responders increases, the sample size decreases. Finally, we see that as the response rate increases from 40% to 60%, the sample size increases. At first glance, this may seem counterintuitive, however, we can illustrate this easily as follows. Consider a total sample size of 200 where 50% are randomized to A_1 and the other 50% are randomized to A_2. Thus, there are 100 patients in each initial treatment arm. If the response rate is 40%, 40 patients from each A_j, $j = 1, 2$, are randomized among B_1 and B_2 and the other 120 patients, split equally among A_1 and A_2, are non-responders. Then each strategy has a working sample

size of 20+60=80. If, however, the response rate to the first-stage treatment is 60%, keeping all randomization probabilities equal at 0.5, the working sample size for each strategy drops to 30+40=70. Thus, when there are more responders, the strategy-specific working sample size decreases, so more patients are needed overall to achieve the same power. Since the weighted log-rank test takes all of the non-responders into consideration, utilizing this group of patients who share common treatment paths, it is advantageous and leads to a smaller sample size when there are fewer responders. This test statistic lends itself to trials for chronic diseases, especially, late-stage or late-onset diseases where a high response rate is not expected, but rather, the goal is to find any measure to elongate survival.

5.5 SIMULATION STUDIES

Since the sample size equation produces appropriate sample sizes and behaves as expected, we generated data based on the sample sizes calculated under different alternatives and tested the empirical power using the weighted log-rank statistic presented in Section 3.3. In order to satisfy the proportional hazards assumption given the user provided inputs, we generated the population survival data from exponential distributions with rates given by $\lambda_{12}/\lambda_{11}, \lambda_{21}/\lambda_{11}$, and $\lambda_{22}/\lambda_{11}$, setting $\lambda_{11} = 1$. We used a trial and error method to find the end of the study time L and the censoring parameter v, such that the user-specified values of $D^R(L)$ and $D^{NR}(L)$ were upheld (we could not solve for these parameters in closed form). Details of the data generating process is shown in the algorithm provided in the appendix.

We had varying degrees of success employing our sample size calculation. Results are presented in Table 5.2. Most notably, we see that five of these cases reach 80% power, and several other cases come close, however, not all of the alternatives shown in Table 5.2 reach this desired level of power. In the first block of results, where $\lambda_{12}/\lambda_{11} = 1.4, \lambda_{21}/\lambda_{11} = 1.5, \lambda_{22}/\lambda_{11} = 1.2$, all of the cases reach or are very close to 80% power, except for one case. When $\pi_0^R = 0.6$ and $v = 2$, we exceed our desired power with an empirical power of 83%, but when $v = 4$, the empirical power of 74% is 6% lower than the desired level. The highest power reached in the second block of results, where $\lambda_{12}/\lambda_{11} = 1.1, \lambda_{21}/\lambda_{11} = 1.3, \lambda_{22}/\lambda_{11} = 1.5$,

is 0.78, but the other alternatives in this block underestimate the power by 4-9%. In the third block of results, where $\lambda_{12}/\lambda_{11} = 1.2, \lambda_{21}/\lambda_{11} = 1.3, \lambda_{22}/\lambda_{11} = 1.4$, the sample size of 1322 delivers 80% power, but the other sample sizes underestimate power, again, by 4-9%. It remains unclear why some alternatives lead to underestimating the power, while other alteratives achieve the desired power.

While some alternatives achieve the desired power level, it appears that most simulations underestimate power. We believe this is due to underestimating the variance-covariance parameters in equations (5.7)-(5.12). We have approximated the instantaneous hazard rate of the non-responders, $d\Lambda_j^{NR}$, $j = 1, 2$, using $(1 - \pi_0^R)d\Lambda_0(s)$. Such an approximation is appropriate for survival (proportions), however, this does not hold for hazard rates. Currently, we are thinking of considering some variance inflation factor.

We suggest further simulations to explore the discrepancy between the empirical and expected power levels from the sample size equation. It is ideal to find situations with uninformative censoring, however, given the user inputs, it is difficult to find the L and v to satisfy this and even more difficult to expect that these parameters be the ones followed for the trial. Ideally, we would like the sample size formula to remain robust to the inputs. It appears, however, that the sample size equation is quite sensitive to the specifications of $D^R(L)$ and $D^{NR}(L)$ which then determine L and v. Therefore, more studies are needed to assess the power and robustness of this statistic and perhaps modifications to ensure the formula not only behaves as expected, but produces robust results which will allow the SMARTs designed using this formula to have the desired power.

Table 5.2: Testing the Empirical Power of the Sample Size Formula

$\lambda_{12}/\lambda_{11}$	$\lambda_{21}/\lambda_{11}$	$\lambda_{22}/\lambda_{11}$	π_0^R	$D^R(L)$	$D^{NR}(L)$	L	v	n	Power
1.4	1.5	1.2	0.5	0.30	0.43	0.4	1	727	0.808
			0.5	0.42	0.60	0.6	2	521	0.808
			0.6	0.43	0.61	0.6	2	612	0.826
			0.6	0.46	0.65	0.6	4	574	0.737
			0.7	0.34	0.47	0.4	2	916	0.803
			0.7	0.44	0.65	0.6	4	682	0.782
1.1	1.3	1.5	0.5	0.31	0.56	0.5	5	795	0.707
			0.5	0.22	0.37	0.3	5	1182	0.755
			0.6	0.36	0.57	0.6	2	819	0.761
			0.6	0.41	0.66	0.7	3	712	0.761
			0.7	0.32	0.44	0.4	2	1078	0.733
			0.7	0.29	0.40	0.4	1	1188	0.780
1.2	1.3	1.4	0.5	0.28	0.47	0.4	4	1322	0.803
			0.5	0.37	0.62	0.6	4	1022	0.711
			0.6	0.30	0.46	0.4	3	1134	0.761
			0.6	0.39	0.60	0.6	3	1478	0.764
			0.7	0.33	0.44	0.4	2	1601	0.759
			0.7	0.40	0.54	0.5	3	1313	0.745

n is the total sample size for all strategies. Desired type I error was set to 0.05. $\lambda_{jk}/\lambda_{11}$ is the hazard ratio of strategy A_jB_k to A_1B_1. π_0^R is the expected proportion of responders by the end of the study. $D^R(L)$ and $D^{NR}(L)$ are the expected proportions of events among the responders and non-responders, respectively, by the end of the study. L denotes the end of the study in years and v is the censoring parameter.

6.0 DISCUSSION AND FUTURE WORK

Adaptive treatment strategies have become more prevalent in clinical research, especially in the treatment of chronic diseases, where management of the disease is more important than a cure. Two-stage randomization designs (or more generally SMART designs) are, therefore, commonly being used in clinical trials to compare adaptive treatment strategies with two decision points. Since many clinical trials focus on a time-to-event endpoint, the development of statistical methods for survival analysis in two-stage randomized designs is essential. While others have developed statistics to estimate point-wise survival or compare overall survival distributions of separate-path adaptive treatment strategies, methods for comparing the overall survival distributions of adaptive treatment strategies that share common paths are not available in current literature.

These shared-path adaptive treatment strategies share a common path of treatment such that there is a common group of patients who could be considered as treated with more than one adaptive treatment strategy in the data collected through SMART designs. To address this, we have proposed a weighted log-rank statistic which takes into account both the two-stage randomized design and the statistical dependence among groups of patients who follow each strategy. We have provided the asymptotic properties of these tests and we have shown that the proposed weighted log-rank statistic comparing two or more adaptive treatment strategies which may share common treatment paths generally maintains type I error rates and has greater power than naive methods of analysis in most cases. Our derivation of the asymptotic properties of the statistic is based on the assumption that the censoring is non-informative. Like many other survival methods, this method may not be valid when the censoring is informative, e.g., when the censoring rate differs by response status. More research is needed to incorporate informative censoring into the weighted log-rank statistic.

Future research in the area of adaptive treatment strategies also includes the extension of the weighted log-rank statistic to compare survival distributions of patients who follow adaptive treatment strategies in general (multi-stage) SMART designs and further verification of the sample size equation. Increasing the number of stages in a SMART design, increases the number of dependent groups as more patients follow a common path and therefore are shared-path strategies. While most practical SMART designs include only two-stages, we expect SMART designs to include more stages and treatments in the future as the usage of SMARTs and the statistical techniques to design and analyze these trials increase. Thus, extending the WLR test could prove very useful to compare the overall survival of multiple ATS in multi-stage designs.

To encourage the use of SMARTs and their analysis using the weighted log-rank test, practical sample size equations are needed. We have presented the sample size formulation with the goal of comparing the overall survival of multiple ATS, some of which may be shared-path. More work is needed, however, to validate this formula and test its robustness to assumptions. Specifically, simulation studies are needed to continue to test the empirical power of the formula, especially the effect of informative censoring and assuming exponential survival distributions. Simulations are also needed to test the formulation's robustness to assuming equal randomization probabilities to first- and second-stage treatments, equal proportions of responders to each of the initial treatments, and equal censoring rates in all strategies.

Methods for the design and analysis of SMARTs are on the rise as patients and doctors demand personalized medicine. By employing our sample size formula, clinics can set up a SMART to compare multiple ATS using information from pilot or previous studies. Utilizing the SMART design allows prospective clinical trials to study sequences of treatments dependent on patient's characteristics and/or behaviors. Analyzing the data from these trials provides evidence for the best treatment strategy for patients suffering from chronic diseases. In the treatment of chronic diseases, such as cancer or AIDS, we're interested in elongating the patients' survival, thus we must be able to compare patients' overall survival between different adaptive treatment strategies. The weighted log-rank test compares the overall survival of two shared-path or multiple ATS, allowing SMARTs to find an overall

difference in ATS which can lead to confirmatory trials of treatment sequences and optimal adaptive treatment strategies.

APPENDIX

ALGORITHM TO GENERATE DATA FOR SAMPLE SIZE SIMULATIONS

1. Generate a large number (we used 50,000) of observations from an exponential distribution with parameter λ_{11} (we took $\lambda_{11} = 1$) to represent the population of failure times for those following A_1B_1.
2. For the given hazard ratios, $\lambda_{jk}/\lambda_{11}$, $j, k = 1, 2$, generate similar population observations for the other three strategies.
3. Create responder and non-responder populations within each of these 4 populations. Specifically,
   a. Choose a cut-off t_0 for non-responders survival time in the A_1B_1 population (we chose 75^{th} percentile of the distribution) such that at least $100(1 - \pi_0^R)\%$ observations are below the cut-off.
   b. Randomly designate $100(1 - \pi_0^R)\%$ of this population with survival less than t_0 to be non-responders (call this subpopulation Ω_{10}), and remaining observations as responders (call them Ω_{11}).
   c. Find failure times in the A_1B_2 population that closely match the observations of non-responders in the A_1B_1 population (Ω_{10}) and designate them as non-responders. Replace these failure times from non-responders in A_1B_2 with those identified as non-responders in A_1B_1 population. The remaining observations will be responders in the A_1B_2 population (Ω_{12}).
   d. Follow similar procedures (Steps 3a-3c) to create a common group of non-responders between strategies A_2B_1 and A_2B_2 (Ω_{20}), and corresponding responders (Ω_{21}, Ω_{22}).

4. Combine the two non-responders population ($\Omega_{NR} = \Omega_{10} \cup \Omega_{20}$) and the four responders population ($\Omega_R = \Omega_{11} \cup \Omega_{12} \cup \Omega_{21} \cup \Omega_{22}$).
5. Determine the maximum follow-up time L and the censoring parameter v to match the desired proportion of events ($D^R(L)$ and $D^{NR}(L)$) using the observations in Ω_{NR} and Ω_{NR}.
6. Given the overall sample size n, sample $0.5n(1-\pi_0^R)$ observations from Ω_{10} as observations not responding to A_1, and sample $0.5n(1-\pi_0^R)$ observations from Ω_{20} as observations not responding to A_2, rounding up to the largest integer.
7. Given the overall sample size n, sample $0.25n\pi_0^R$ observations from each Ω_{11}, Ω_{12}, and Ω_{21}. To ensure the sample size n and to account for the rounding, set one of the group sizes, we have chosen those from Ω_{22}, equal to the difference between n and the other group sizes.
8. Assign the response and treatment indicators. Assign the response indicator such that $R = 1$ for all those sampled from Ω_R and $R = 0$ for all those from Ω_{NR}, the initial treatment indicator such that $X = 1$ for those sampled from Ω_{10}, Ω_{11}, and Ω_{12}, and $X = 0$ for those sampled from Ω_{20}, Ω_{21}, and Ω_{22}, and the second-stage treatment indicator such that $Z = 1$ for those sampled from Ω_{11}, and Ω_{21}, and $Z = 0$ for those sampled from Ω_{12}, and Ω_{22}.
9. Apply censoring to the survival times. We have applied uniform censoring from zero to v. Define the observed survival time to be the minimum of the sampled failure time and censoring time with complete-case indicator equal to 1 if the sampled time is less than the censoring time.

BIBLIOGRAPHY

Ahnn, S. and Anderson, S. J. (1995). Sample size determination for comparing more than two survival distributions. *Statistics in Medicine*, 14:2273–2282.

Ahnn, S. and Anderson, S. J. (1998). Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. *Statistics in Medicine*, 17:2525–2534.

Almirall, D., Compton, S. N., Gunlicks-Stoessel, M., Duan, N., and Murphy, S. A. (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy.

Auyeung, S. F., Long, Q., Royster, E. B., Murthy, S., McNutt, M. D., Lawson, D., Miller, A., Manatunga, A., and Musselman, D. L. (2009). Sequential multiple-assignment randomized trial design of neurobehavioral treatment for patients with metastatic malignant melanoma undergoing high-dose interferon-alpha therapy. *Clinical Trials*, 6:480–490.

Barthel, F., Babiker, A., Royston, P., and Parmar, M. (2006). Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Statistics in Medicine*, 25:2521–42.

Chakraborty, B. (2011). Dynamic treatment regimes for managing chronic health conditions: A statisticial perspective. *American Journal of Public Health*, 101:40–45.

Collins, L. M., Murphy, S. A., and Bierman, K. A. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science*, 5:185–196.

Dawson, R. and Lavori, P. W. (2004). Placebo-free designs for evaluating new mental health treatments: The use of adaptive treatment strategies. *Statistics in Medicine*, 23:3249–3262.

Dawson, R. and Lavori, P. W. (2010). Sample size calculations for evaluating treatment policies in multi-stage designs. *Clinical Trials*, 7:643–652.

Feng, W. and Wahed, A. S. (2008). Supremum weighted log-rank test and sample size for comparing two-stage adaptive treatment strategies. *Biometrika*, 95:695–707.

Feng, W. and Wahed, A. S. (2009). Sample size for two-stage studies with maintenance therapy. *Statistics in Medicine*, 28:2028–2041.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley and Sons, Inc., New York.

Guo, X. (2005). *Statistical analysis in two-stage randomization designs in clinical trials*. PhD thesis, Department of Statistics, North Carolina State University.

Guo, X. and Tsiatis, A. A. (2005). A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics*, 1:1–15.

Halabi, S. and Singh, B. (2004). Sample size determination for comparing several survival curves with unequal allocations. *Statistics in Medicine*, 23:1793–1815.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69:553–566.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–60.

Jones, H. (2010). Reinforcement-based treatment for pregnant drug abusers (HOME II). http://clinicaltrials.gov/ct2/show/NCT01177892?term=jones+pregnant{&}rank=9.

Jung, S.-H. and Hui, S. (2002). Sample size calculation for rank tests comparing K survival distributions. *Lifetime Data Analysis*, 8:361–373.

Kasari, C. (2009). Developmental and augmented intervention for facilitating expressive language (CC-NIA). http://clinicaltrials.gov/ct2/show/NCT01013545?term=kasari{&}rank=5.

Kosorok, M. R. and Lin, C. Y. (1999). The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association*, 94:320–332.

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: Biased individually tailored within-subject randomization. *Journal of the Royal Statistical Society*, A 163:29–38.

Lavori, P. W., Dawson, R., and J., R. A. (2000). Flexible treatment strategies in chronic disease: Clinical research implications. *Biological Psychology*, 48:605–614.

Lavori, P. W., Dawson, R., and J., R. A. (2004). Dynamic treatment regimes: Practical design considerations. *Clinical Trials*, 1:9–20.

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

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

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*, 58:48–57.

Marlowe, D. B., Festinger, D. S., Dugosh, K. L., Lee, P. A., and Benasutti, K. M. (2007). Adapting judicial supervision to the risk level of drug offenders: Discharge and 6-month outcomes from a prospective matching study. *Drug and Alcohol Dependence*, 88:S4–S13.

Matthay, K. K., Reynolds, C. P., Seeger, R. C., Shimada, H., Adkins, E. S., Haas-Kogan, D., Gerbing, R. B., London, W. B., and Villablanca, J. G. (2009). Long-term results for children with high-risk neuroblastoma treated on a randomized trial of myeloablative therapy followed by 13-cis-retinoic acid: A children's oncology group study. *Journal of Clinical Oncology*, 27:1007–1013.

Matthay, K. K., Villablanca, J. G., Seeger, R. C., Stram, D. O., Harris, R. E., Ramsay, N. K., Swift, P., Shimada, H., Black, C. T., Brodeur, G. M., Gerbing, R. B., and Reynolds, C. P. (1999). Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. *The New England Journal of Medicine*, 341:1165–1173.

Mitchell, J. E., Agras, S., Crow, S., Halmi, K., Fairburn, C. G., Bryson, S., and Kraemer, H. (2011). Stepped care and cognitive-behavioural therapy for bulimia nervosa: randomised trial. *The British Journal of Psychiatry*, 198:391–397.

Murphy, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society*, 65:331–66.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistical Methods*, 24:1455–81.

Murphy, S. A. and McKay, J. R. (2004). Adaptive treatment strategies: An emerging approach for improving treatment effectiveness). *Clinical Science*, Winter2003/Spring2004:7–13.

Oetting, A. I., Levy, J. A., Weiss, R. D., and Murphy, S. A. (2011). Statistical methodology for a SMART design in the development of adaptive treatment strategies. In Shrout, P., Keyes, K., and Ornstein, K., editors, *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, American Psychopathological Association, chapter 8, pages 179–205. American Psychiatric Publishing, Inc., Arlington, VA.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *The International Journal of Biostatistics*, 6.

Pelham, W. and Fabiano, G. (2008). Evidence-based psychosocial treatments for attention-deficit/hyperactivity disorder. *Journal of Clinical Child and Adolescent Psychology*, 37:184–214.

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

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

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., and Kashner, T. M. (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR*D): Rationale and design. *Controlled Clinical Trials*, 25:119–42.

Stone, R. M., Berg, D. T., George, S. L., Dodge, R. K., Paciucci, P. A., Schulman, P. P., Lee, E. J., Moore, J. O., Powell, B. L., Baer, M. R., Bloomfield, C. D., and Schiffer, C. A. (2001). Postremission therapy in older patients with de novo acute myeloid leukemia: A randomized trial comparing mitoxantrone and intermediate-dose cytarabine with standard-dose cytarabine. *Blood*, 98:548–53.

Stone, R. M., Berg, D. T., George, S. L., Dodge, R. K., Paciucci, P. A., Schulman, P. P., Lee, E. J., Moore, J. O., Powell, B. L., Baer, M. R., 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*, 332:1671–1677.

Strecher, V., McClure, J., Alexander, G., Chakraborty, B., Nair, V., Konkel, J., Greene, S., Collins, L., Carlier, C., Wiese, C., Little, R., Pomerleau, C., and Pomerleau, O. (2008). Web-based smoking cessation programs: Results of a randomized trial. *American Journal of Preventive Medicine*, 34:373–381.

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., McGee, M. F., Simpson, G. M., Stevens, M. C., and Lieberman, J. A. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin*, 29:15–31.

Tannir, N. M. (2010). Sequential two-agent assessment in renal cell carcinoma therapy. http://clinicaltrials.gov/ct2/show/NCT01217931.

Thall, P. F., Millikan, R. E., and Sung, H.-G. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, 30:1011–1128.

Wahed, A. S. (2010). Inference for two-stage adaptive treatment strategies using mixture distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:1–18.

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.

Wahed, A. S. and Tsiatis, A. A. (2006). Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, 93:163–177.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2011). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. http://amstat.tandfonline.com/doi/full/10.1080/01621459.2011.641416.

Winter, J. N., Weller, E. A., Horning, S. J., Krajewska, M., Variakojis, D., Habermann, T. M., Fisher, R. I., Kurtin, P. J., Macon, W. R., Chhanabhai, M., Felgar, R. E., Hsi, E. D., Medeiros, L. J., Weick, J. K., Reed, J. C., and Gascoyne, R. D. (2006). Prognostic significance of Bcl-6 protein expression in DLBCL treated with CHOP or R-CHOP: A prospective correlative study. *Blood*, 107:4207–13.

Wolbers, M. and Helterbrand, J. D. (2008). Two-stage randomization designs in drug development. *Statistics in Medicine*, 27:4161–4174.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, Michael, R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67:1422–1433.