Weighted Log-rank Statistic to Compare Shared-Path Adaptive Treatment Strategies

KELLEY M. KIDWELL and ABDUS S. WAHED

University of Pittsburgh,

Graduate School of Public Health, Biostatistics

kmk99@pitt.edu

Abstract

Adaptive treatment strategies more 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 paper, 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 or more two-stage, shared-path 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.

Keywords: Adaptive treatment strategy; Counting process; Proportional hazard; Survival function; Two-stage design; Weighted log-rank statistic

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 article focuses on time-to-event outcome data and hence the review of literature will mainly 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 analysis, 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). 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.

The goal of this article is to present methods for comparing two shared-path adaptive treatment strategies (strategies that share some of the same treatment paths, see Section 2). In addition, we would like to compare more than two adaptive treatment strategies 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 survival curves of two or more shared-path adaptive treatment strategies include: (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 Guo (2005), Lokhnygina and Helterbrand (2007), or in Feng and Wahed (2008), 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.

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. In this paper we first propose a weighted log-rank statistic to compare two shared-path adaptive treatment strategies and then extend it to compare the overall survival distributions of more than two shared-path adaptive treatment strategies.

2 Setup

2.1 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, receive maintenance treatment B, randomly allocated to levels B_1 or B_2 (see Figure 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.

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 as follows:

Definition 2. *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 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.

2.2 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 one can envision one outcome for each possible strategy for each individual, hence every individual has his/her own set of imaginary outcomes for every possible treatment strategy. This 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 the 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 received and responded to treatment A_j. For treatment strategy A_jB_k, a particular patient may respond to A_j or fail to respond to A_j while receiving at most one treatment at stage one or stage two. Since every patient only follows one path within a treatment strategy, we cannot observe R_{ji}, T_{ji}^{NR}, and T_{jki}^{R}, for all $j, k = 1, 2$, for each patient. Consequently, these variables are the counterfactuals or

potential random variables. 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 groups with a specific extension to compare all four strategies, A_1B_1, A_1B_2, A_2B_1 and A_2B_2.

2.3 Observed data & assumptions

The observed data for a two-stage design described in Figure 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 2.2.

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 is given as $T_i = \sum_{j=1}^{J} X_{ji}\{(1 - R_{ji})T_{ji}^{NR} + R_{ji}Z_iT_{j1i}^{R} + R_{ji}(1 - Z_i)T_{j2i}^{R}\}$. Thus, an individual's survival time is not related to others' treatment allocation. 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. 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.

3 Log-rank statistic for comparing two dependent strategies

3.1 The 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

$$Z_n(t) = \int_0^t \frac{Y_{11}(s)Y_{12}(s)}{Y_{11}(s) + Y_{12}(s)} \left\{ \frac{dN_{11}(s)}{Y_{11}(s)} - \frac{dN_{12}(s)}{Y_{12}(s)} \right\}, \tag{1}$$

where $N_{1ki}(s) = I(U_{1ki} \leq s, \delta_{ki} = 1)$, $Y_{1ki}(s) = I(U_{1ki} \geq s)$, $N_{1k}(s) = \sum_{i=1}^n N_{1ki}(s)$, 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 article. Specifically, we propose a weighted log-rank statistic to test the hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$ accounting for the fact that patients following A_1B_1 includes a group of patients who also follow A_1B_2.

We present the notation for time-dependent weights which is adapted from Guo and Tsiatis (2005). Explicitly, let $W_{11i}(s) = X_i\{1 - R_i(s) + \frac{R_i(s)Z_i}{\pi}\}/\phi$ 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 $R_i(s) = 1$ 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 probability of being assigned to A_1. Similarly, $W_{12i}(s) = X_i\{1 - R_i(s) + \frac{R_i(s)(1-Z_i)}{1-\pi}\}/\phi$ for estimating quantities related to the strategy A_1B_2. Note that if a patient randomized to A_1 has not responded by time s, $W_{11i}(s) = W_{12i}(s) = 1/\phi$, confirming that the non-responders are consistent with both strategies; if the patient has responded and is randomized to B_1 by time s, $W_{11i}(s) = 1/(\phi\pi)$ 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 $W_{12i}(s) = 1/\{\phi(1-\pi)\}$. This construction of weights is based on the fundamental principle of inverse-probability-of-treatment-weighting (Robins et al., 1994).

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 Table 1. The general at-risk process for all patients is $Y_i(s) = I(U_i \geq s)$, the at-risk process for those with initial treatment A_j, $j = 1, 2$, is $Y_{ji}(s) = I(U_i \geq s, X_i = 2 - j)$, the weighted at-risk process is $\bar{Y}_{jk}(s) = \sum_{i=1}^{n} W_{jki}(s) Y_{ji}(s)$, the at-risk process for only those who are non-responders to A_j is $Y_j^{NR}(s) = \sum_{i=1}^{n} \{1 - R_i(s)\} Y_{ji}(s)$, the overall at-risk process for patients treated with A_j is $Y_{j\cdot}(s) = \sum_{i=1}^{n} Y_{ji}(s)$ and the overall at risk-process for all patients is $Y(s) = \sum_{i=1}^{n} Y_i(s)$. Likewise, the general event process for any patient i is $N_i(s) = I(U_i \leq s, \delta_i = 1)$, the event process for those with first-line treatment A_j, $j = 1, 2$, is $N_{ji}(s) = I(U_i \leq s, \delta_i = 1, X_i = 2 - j)$, the weighted event process is $\bar{N}_{jk}(s) = \sum_{i=1}^{n} W_{jki}(s) N_{ji}(s)$, the event process for only those who are non-responders to A_j is $N_j^{NR}(s) = \sum_{i=1}^{n} \{1 - R_i(s)\} N_{ji}(s)$, the overall event process for patients treated with A_j is $N_{j\cdot}(s) = \sum_{i=1}^{n} N_{ji}(s)$, 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: $\Lambda_{11}(t) = \Lambda_{12}(t)$ is defined as

$$Z_n^W(t) = \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \left\{ \frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}_{12}(s)}{\bar{Y}_{12}(s)} \right\}. \tag{2}$$

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 (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 $\mathcal{F}(t) = \sigma\{R_i(s), Z_i R_i(s), I(C_i \leq s), N_{ji}(s), i = 1, \ldots, n;\ j = 1, 2;\ 0 \leq s \leq t\}$), the weighted log-rank statistic in equation (2) has expectation zero (see Section 3.2).

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) to test the null hypothesis H_0: $\Lambda_{11}(t) = \Lambda_{12}(t)$.

3.2 Asymptotic properties

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

$$G_n(t) \;=\; n^{-1/2}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}\left\{\frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{12}(s)}{\bar{Y}_{12}(s)}\right\} \tag{3}$$

and $R_n(t) = n^{-1/2}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}\{d\Lambda_{11}(s) - d\Lambda_{12}(s)\}$, since $\bar{M}_{jk}(t) = \bar{N}_{jk}(t) - \int_0^t \bar{Y}_{jk}(s)d\Lambda_{jk}(s)$.

Feng and Wahed (2008, p. 699) have showen that $d\bar{M}_{jk}(t) = \sum_{i=1}^n W_{jki}(t)dM_{jki}(t)$ and it is easy to show that $E\{d\bar{M}_{jk}(t)|\mathcal{F}(t_-)\} = 0$, where $M_{jki}(t)$ is the ith patient specific martingale, corresponding to $M_{jk}(t) = N_{jk}(t) - \int_0^t Y_{jk}(s)d\Lambda_{jk}(s)$, 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. Under the null hypothesis, $\Lambda_{11}(t) = \Lambda_{12}(t)$, so $n^{-1/2}Z_n^{WLR}(t) = G_n(t)$ in equation (3). Since 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 further expand $G_n(t)$. Using $d\bar{M}_{jk}(t) = \sum_{i=1}^n W_{jki}(t)dM_{jki}(t)$, $G_n(t)$ can be expressed as a difference of two martingale processes, $G_n(t) = G_n^{11}(t) - G_n^{12}(t) = n^{-1/2}\left\{\sum_{i=1}^n \int_0^t \frac{\bar{Y}_{12}(s)W_{11i}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}dM_{11i}(s) - \sum_{i=1}^n \int_0^t \frac{\bar{Y}_{11}(s)W_{12i}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}dM_{12i}(s)\right\}$. 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 $n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\bar{Y}^2_{1(3-k)}(s)W^2_{1ki}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}Y_{1i}(s)d\Lambda_{1k}(s)$, $k=1,2$.

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

$$\begin{aligned} & n^{-1}\sum_{k=1}^{2}\sum_{i=1}^{n}\int_0^t \frac{\bar{Y}^2_{1(3-k)}(s)W^2_{1ki}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}Y_{1i}(s)d\Lambda_{1k}(s) \\ & - 2n^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}\sum_{i=1}^{n} W_{11i}(s)W_{12i}(s)cov\{dM_{11i}(s), dM_{12i}(s)\}. \end{aligned} \tag{4}$$

First, note that $W_{11i}(s)W_{12i}(s) = \phi^{-2}\{1-R_i(s)\}$. 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 (4) can be shown to be equal to $\sum_{i=1}^{n} 1/\phi^2\{1-R_i(s)\}\{Y_{1i}(s)d\Lambda_0(s)\} = \phi^{-2}Y_1^{NR}(s)d\Lambda_0(s)$. Thus, a consistent variance estimator of $n^{-1/2}Z_n^W(t)$ is given by

$$\begin{aligned} \hat{\sigma}^2(t) = {} & n^{-1}\int_0^t \frac{\bar{Y}^2_{12}(s)\sum_{i=1}^{n} W^2_{11i}(s)Y_{1i}(s) + \bar{Y}^2_{11}(s)\sum_{i=1}^{n} W^2_{12i}(s)Y_{1i}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}^2}\left\{\frac{dN_{1.}(s)}{Y_{1.}(s)}\right\} \\ & - 2n^{-1}\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\left\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\right\}^2}\left\{\phi^{-2}Y_1^{NR}(s)\frac{dN_{1.}(s)}{Y_{1.}(s)}\right\}. \end{aligned} \tag{5}$$

The notation used in the above equation or elsewhere in this article can be reviewed in Table 1. The corresponding standardized weighted log-rank test statistic is given by $T_n^W(L)$, where

$$T_n^W(L) = n^{-1/2}Z_n^W(L)/\hat{\sigma}(L), \tag{6}$$

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.

4 LOG-RANK TESTS FOR MULTIPLE DEPENDENT ADAPTIVE TREATMENT STRATEGIES

In the setting described above, 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 hazards 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 3, an unbiased estimator of $\zeta(\mathbf{t})$ is given by $\hat{\zeta}(t) = \{\frac{d\bar{N}_{11}(t)}{\bar{Y}_{11}(t)} - \frac{d\bar{N}_{12}(t)}{\bar{Y}_{12}(t)}, \frac{d\bar{N}_{11}(t)}{\bar{Y}_{11}(t)} - \frac{d\bar{N}_{21}(t)}{\bar{Y}_{21}(t)}, \frac{d\bar{N}_{11}(t)}{\bar{Y}_{11}(t)} - \frac{d\bar{N}_{22}(t)}{\bar{Y}_{22}(t)}\}^T$. The corresponding weighted log-rank statistic for testing H_0 is the vector of the weighted martingale differences, $\mathbf{Z}_n^{MW}(t) = \{Z_n^{11.12}(t), Z_n^{11.21}(t), Z_n^{11.22}(t)\}^T$ where

$$Z_n^{jk.j'k'}(t) = \int_0^t \frac{\bar{Y}_{jk}(s)\bar{Y}_{j'k'}(s)}{\bar{Y}_{jk}(s) + \bar{Y}_{j'k'}(s)} \left\{ \frac{d\bar{N}_{jk}(s)}{\bar{Y}_{jk}(s)} - \frac{d\bar{N}_{j'k'}(s)}{\bar{Y}_{j'k'}(s)} \right\}. \tag{7}$$

Under the null hypothesis, the statistic $\mathbf{Z}_n^{MW}(t)$ has expectation zero. Since $\mathbf{Z}_n^{MW}(t)$ is a linear combination of weighted Z^W-statistics defined in equation (2), by the multivariate central limit theorem for martingales (Fleming and Harrington, 1991), $n^{-1/2}\mathbf{Z}_n^{MW}(t)$ follows a mean zero Gaussian process with asymptotic variance covariance matrix, $\mathbf{\Sigma}(t)$, that can be estimated by $\hat{\mathbf{\Sigma}}(t) = \{s_{ij}(t)\}^{3x3}$, where the elements of $\hat{\mathbf{\Sigma}}(t)$ are defined as follows.

The estimated variances of $\mathbf{Z}_n^{MW}(t)$ are given below, where the induction-treatment-specific processes, $N_{1.}(s)$ and $Y_{1.}(s)$ used in equation (5), have been substituted with the overall processes, $N(s)$ and $Y(s)$, to reflect that under the null, all strategies have equal hazards. Explicitly,

$$s_{11}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{12}^2(s) \sum_{i=0}^n W_{11i}^2(s) Y_{1i}(s) + \bar{Y}_{11}^2(s) \sum_{i=0}^n W_{12i}^2(s) Y_{1i}(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}^2} \left\{ \frac{dN(s)}{Y(s)} \right\}$$
$$- 2n^{-1} \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}^2} \left\{ \phi^{-2} Y_1^{NR}(s) \frac{dN(s)}{Y(s)} \right\}. \quad (8)$$

$$s_{22}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{21}^2(s) \sum_{i=0}^n W_{11i}^2(s) Y_{1i}(s) + \bar{Y}_{11}^2(s) \sum_{i=0}^n W_{21i}^2(s) Y_{2i}(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{21}(s)\}^2} \left\{ \frac{dN(s)}{Y(s)} \right\} \quad (9)$$

$$s_{33}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{22}^2(s) \sum_{i=0}^n W_{11i}^2(s) Y_{1i}(s) + \bar{Y}_{11}^2(s) \sum_{i=0}^n W_{22i}^2(s) Y_{2i}(s)}{\{\bar{Y}_{11}(s) + \bar{Y}_{22}(s)\}^2} \left\{ \frac{dN(s)}{Y(s)} \right\}. \quad (10)$$

Note that the last two formulas above do not contain a covariance term since $d\bar{N}_{jk}(s)/\bar{Y}_{jk}(s)$ and $d\bar{N}_{j'k'}(s)/\bar{Y}_{j'k'}(s)$, $j \neq j'$, are conditionally independent given $\mathcal{F}(s_-)$.

To obtain an expression for the estimated covariance terms in $\hat{\Sigma}(t)$, we first give the expressions for the covariances in the supplementary material. Following those expressions where $d\hat{\Lambda}_0(s) = dN(s)/Y(s)$, natural estimates of the covariances are given by

$$s_{12}(t) = n^{-1} \int_0^t \bar{Y}_{21}(s)[\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s) + \bar{Y}_{21}(s)\}]^{-1} \left\{ \bar{Y}_{12}(s) \sum_{i=1}^n W_{11i}^2(s) Y_{1i}(s) - \phi^{-2} \bar{Y}_{11}(s) Y_1^{NR}(s) \right\} d\hat{\Lambda}_0(s), \quad (11)$$

$$s_{13}(t) = n^{-1} \int_0^t [\bar{Y}_{22}(s)\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s) + \bar{Y}_{22}(s)\}]^{-1} \left\{ \bar{Y}_{12}(s) \sum_{i=1}^n W_{11i}^2(s) Y_{1i}(s) - \phi^{-2} \bar{Y}_{11}(s) Y_1^{NR}(s) \right\} d\hat{\Lambda}_0(s), \quad (12)$$

$$s_{23}(t) = n^{-1} \int_0^t [\{\bar{Y}_{11}(s) + \bar{Y}_{21}(s)\}\{\bar{Y}_{11}(s) + \bar{Y}_{22}(s)\}]^{-1} \left\{ \bar{Y}_{21}(s)\bar{Y}_{22}(s) \sum_{i=1}^n W_{11i}^2(s) Y_{1i}(s) + (1-\phi)^{-2} \bar{Y}_{11}^2(s) Y_2^{NR}(s) \right\} d\hat{\Lambda}_0(s). \quad (13)$$

The vector of weighted log-rank statistics, $n^{-1/2}\mathbf{Z}_n^{MW}(t)$, presented in Section 4, converges in distribution under the null hypothesis to a trivariate normal distribution with mean zero and variance covariance matrix $\mathbf{\Sigma}(t)$, where $\mathbf{\Sigma}(t)$ is estimated by $\hat{\mathbf{\Sigma}}(t) = \{s_{ij}(t)\}^{3x3}$. Using the unbiased and consistent estimators of $\mathbf{\Sigma}(t)$, by multivariate Slutsky's theorem, we have $n^{-1}\mathbf{Z}_n^{MW}(t)^T \hat{\mathbf{\Sigma}}^{-1}(t) \mathbf{Z}_n^{MW}(t)$ 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 treat-

ment strategies A_jB_k, $j,k=1,2$, is then expressed in the form

$$T_n^{MW}(L) = n^{-1}\mathbf{Z}_n^{MW}(L)^T\hat{\mathbf{\Sigma}}^{-1}(L)\mathbf{Z}_n^{MW}(L), \tag{14}$$

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.

5 SIMULATION RESULTS

5.1 Data generation

To evaluate the performance of the weighted log-rank statistics for comparing two or more (shared-path or separate-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. In our simulation study, 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 (6) 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. While testing four shared-path adaptive treatment strategies, we have compared the proposed weighted log-rank statistic (WLR), $T_n^{MW}(L)$ from equation (14), to the standard unweighted log-rank statistic (SLR).

We outline the data generation process here and provide specific parameters for each simulation in Sections 5.2 and 5.3. The initial treatment indicator, X_i, was generated from a Bernoulli distribution with $\phi = pr(X_i = 1) = 0.5$ 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 $T_{jki} = (1 - R_i)T_{ji}^{NR} + R_i T_{jki}^{*}$, $j, k = 1, 2$. These variables reflect the overall survival time under strategy $A_j B_k$, $(j, k = 1, 2)$. The observed survival time for the ith individual in the absence of censoring is defined as $T_i = X_i[R_i\{Z_i T_{11i}^{*} + (1-Z)T_{12i}^{*}\} + (1-R_i)T_{1i}^{NR}] + (1-X_i)[R_i\{Z_i T_{21i}^{*} + (1-Z_i)T_{22i}^{*}\} + (1-R_i)T_{2i}^{NR}]$. 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. The final observed time was then defined as $U_i = \min(T_i, C_i)$ with corresponding complete case indicator, $\delta_i = I(T_i \leq C_i)$.

For each generated dataset we conducted weighted log-tank tests described in Sections 3.2 and 4 to test the hypotheses $H_{0,1} : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_0(t)$ and $H_{0,2} : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t) = \Lambda_0(t)$, respectively. We report the estimated type I error (proportion of samples for which the hypothesis was falsely rejected) for all tests in Tables 2 and 3 when $H_{0,1}$ and $H_{0,2}$ were true, and the estimated power (proportion of samples for which the hypothesis was correctly rejected)

for all tests in Tables 4 and 5.

5.2 Simulation from the null distribution

To investigate the performance of the weighted log-rank statistics under $H_{0,1} : \Lambda_{11}(t) = \Lambda_{12}(t)$ and $H_{0,2} : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$, we generated 5000 datasets with the following parameters: $\mu_1^{NR} = \mu_2^{NR} = \theta_1^R = \theta_2^R = 1$ and $\theta_{11}^{RE} = \theta_{12}^{RE} = \theta_{21}^{RE} = \theta_{22}^{RE} = 5$. With a 40% response rate, the censoring paramter v was set to 8.4 and 3.5 and with a 60% response rate, v was set to 12 and 5.6 to produce about 30% and 50% censoring, respectively.

Table 2 presents the estimated type I error rates (proportion of samples for which the hypothesis was falsely rejected) for testing the null hypothesis $H_{0,1}$. For a sample size of 200, 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 0.01. 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 0.01. Preserving the response rate at 40%, but increasing censoring led to a decreased type I error rate, such that the WLR test had an estimated type I error rate of 0.04, the IWLR test and the SLR test had estimated error rates less than 0.001. Under this null distribution, increasing censoring decreased the percentage of observed responders. In this particular case for a sample size of 200, the true distribution specified 40% responders, but with 30% censoring the percentage of observed responders decreased to about 35%, while for 50% censoring, the percentage of observed responders dropped to about 29%. As the percentage of observed responders decreases, the statistic relies more on the information from non-responders. Since non-responders are consistent with both strategies they are weighted equally and thus the WLR test will behave more like the SLR test rejecting the null more often, however, since the WLR correctly accounts for the covariance between the groups, it is not as conservative as the IWLR or

SLR tests. In general, increasing censoring slightly decreased the estimated type I error rate for all tests, but the WLR test maintained the type I error rate of 0.05 in all scenarios. 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.

Table 3 presents the estimated type I error rates for testing the null hypothesis $H_{0,2}$ using the WLR and SLR tests. The type I error rates for both statistics were similar across all combinations. For a response rate of 40% and censoring of 30%, the WLR test maintained type I error for sample sizes greater than 100. Specifically, for a sample size of 200, the WLR test for 40% responders and 30% censoring produced an estimated type I error rate of 0.05 while the SLR test produced an estimated error rate of 0.04. 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 and the SLR test produced estimated type I error rates of 0.04. Increasing the response rate to 60% produced acceptable type I error rates for all sample sizes of 50 and greater, 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.

5.3 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 (proportion of samples for which the hypothesis was correctly rejected) of the weighted log-rank tests. Data were generated from populations under the alternative hypotheses, where four true survival distributions, designated as scenarios (a)-(d), were plotted in Figure 2 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 ($\mu_1^{NR} = \theta_1^R = \theta_{11}^{RE} = 1$,

$\mu_2^{NR} = 1.11$, $\theta_2^R = 1.67$, $\theta_{12}^{RE} = 5$, $\theta_{21}^{RE} = 3.33$, $\theta_{22}^{RE} = 0.25$). Scenario (c) respresents survival curves where one strategy, A_2B_1, dominates the other strategies ($\mu_1^{NR} = \mu_2^{NR} = 0.33$, $\theta_1^R = \theta_2^R = 2$, $\theta_{11}^{RE} = 1$, $\theta_{12}^{RE} = 0.14$, $\theta_{21}^{RE} = 3.33$, $\theta_{22}^{RE} = 0.5$). Finally, scenario (d) represents intersecting survival curves violating the proportional hazards assumption under which the log-rank statistic is optimal ($\mu_1^{NR} = 0.14$, $\mu_2^{NR} = \theta_{11}^{RE} = \theta_{22}^{RE} = 1$, $\theta_1^R = 2.5$, $\theta_2^R = 0.2$, $\theta_{12}^{RE} = 0.1$, $\theta_{21}^{RE} = 0.33$). The censoring parameter v was set to 5 for scenarios (a) and (b), 4 for scenario (c) and 4.7 for scenario (d) so that censoring ranged from 22-44%.

Table 4 presents the results for testing the null hypothesis $H_{0,1}$ versus the alternative hypothesis that the cumulative hazards for the two shared-path adaptive treatment strategies differ. 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 (c) where the curves begin together, but subsequently deviate. The WLR test maintained very large power in this situation, while the IWLR test and the SLR test failed to pick up the difference in the survival curves in almost all of the iterations.

Table 5 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), we see that for a 40% response rate and about 35% censoring, the WLR test had a high power at 0.996 unlike the SLR test which had power of 0.296, 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. In one of the scenarios (not presented here) with parameters $\mu_1^{NR} = 0.1$, $\mu_2^{NR} = \theta_2^R = \theta_{11}^{RE} = \theta_{22}^{RE} = 1$, $\theta_1^R = 2.5$, $\theta_{12}^{RE} = 0.5$, and $\theta_{21}^{RE} = 0.33$, the WLR test had less power than the SLR test for 40% and 45% responders. This may be due to a relatively higher

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 than the SLR test with higher power. Power of the WLR test also increased and dominated that of the SLR test for increasing percentage of censoring.

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

6 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. (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, refer to Figure 1 and let A_1 represent chemotherapy, A_2 represent ABMT, B_1 represent cis-RA and B_2 represent no further treatment. Therefore, we are interested in comparing the following four treatment strategies: (i) CC: 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) AC: 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 CC and CN and 91 patients who do not respond to first-stage treatment of ABMT and are therefore consistent with shared-path adaptive treatment strategies AC and AN. The goal was to compare survival distributions under these four adaptive treatment strategies. Survival distributions in Figure 3 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 alone and the superiority of cis-RA with no further therapy after chemotherapy over transplantation. 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.

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 CC to CA), the WLR statistic was 0.12 with $p = 0.90$. For comparing treatment strategies which share the same initial treatment of ABMT (shared-path treatment strategies AC to AN) the WLR statistic was -1.16 with $p = 0.25$, showing that the two strategies that start on ABMT are not significantly different. To test if there was a difference in overall survival across the four strategies (CC, CN, AC, AN), the

weighted log-rank statistic from equation (14) 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 2.04 with $p = 0.56$.

Note that from Figure 3, the overall survival curves of the four adaptive treatment strategies look similar to that from simulation under alternative scenario (d). In this dataset of 379 patients, 54% of patients respond to the initial treatment and about 31% of patients are censored. In this setting, the WLR test does have more power to reject the null hypothesis of equal hazards when it is false.

7 Discussion

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 the literature. These shared-path adaptive treatment strategies share a common path of treatment such that there is a common group of patients who are consistent 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 generally maintains type I error rates and has greater power than naive methods of analysis in most cases. Future research in this area 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.

Supplementary Material

Derivation of the variance-covariance matrix of the weighted log-rank statistic comparing four adaptive treatment strategies

We have presented the estimated variance-covariance matrix of $n^{-1/2}\mathbf{Z}_n^{MW}(t)$ in Section 4. Here we give the expressions for the asymptotic covariances, that led to the estimated covariances in equations 4.13-4.15. We derive the covariance specifically for $\sigma_{12}(t) = cov\{Z_n^{11.12}(t), Z_n^{11.21}(t)\}$ corresponding to the estimated covariance $s_{12}(t)$; the derivations of $\sigma_{13}(t)$ and $\sigma_{23}(t)$ follow similarly. To begin, we define the covariance under the null hypothesis, $\sigma_{12}(t) = cov\{Z_n^{11.12}(t), Z_n^{11.21}(t)\}$

$$
\begin{aligned}
&= cov\left[\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}\left\{\frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}_{12}(s)}{\bar{Y}_{12}(s)}\right\}, \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)}\left\{\frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}_{21}(s)}{\bar{Y}_{21}(s)}\right\}\right] \\
&= cov\left[\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)}\left\{\frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{12}(s)}{\bar{Y}_{12}(s)}\right\}, \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)}\left\{\frac{d\bar{M}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{M}_{21}(s)}{\bar{Y}_{21}(s)}\right\}\right].
\end{aligned}
\tag{15}
$$

Distributing the terms and further simplifying equation (15) using martingale properties,

$$
\begin{aligned}
&\sigma_{12}(t) \\
=&cov\left\{\int_0^t \frac{\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} d\bar{M}_{11}(s) - \int_0^t \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} d\bar{M}_{12}(s),\right. \\
&\left.\int_0^t \frac{\bar{Y}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} d\bar{M}_{11}(s) - \int_0^t \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} d\bar{M}_{21}(s)\right\} \\
=&cov\left\{\int_0^t \frac{\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} d\bar{M}_{11}(s), \int_0^t \frac{\bar{Y}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} d\bar{M}_{11}(s)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&- cov\left\{\int_0^t \frac{\bar{Y}_{12}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} d\bar{M}_{11}(s), \int_0^t \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} d\bar{M}_{21}(s)\right\} \\
&- cov\left\{\int_0^t \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} d\bar{M}_{12}(s), \int_0^t \frac{\bar{Y}_{21}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} d\bar{M}_{11}(s)\right\} \\
&+ cov\left\{\int_0^t \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)} d\bar{M}_{12}(s), \int_0^t \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)} d\bar{M}_{21}(s)\right\} \\
=&E\int_0^t \frac{\bar{Y}_{12}(s)\bar{Y}_{21}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}} cov\{d\bar{M}_{11}(s), d\bar{M}_{11}(s) \mid \mathcal{F}(s_-)\} \\
&- E\int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}} cov\{d\bar{M}_{12}(s), d\bar{M}_{11}(s) \mid \mathcal{F}(s_-)\}. \qquad (16)
\end{aligned}
$$

In the intermediate steps to reach equation (16), we have used the fact that $cov\{d\bar{M}_{11}(s), d\bar{M}_{21}(s)\} = cov\{d\bar{M}_{12}(s), d\bar{M}_{21}(s)\} = 0$, due to the separate-path nature of the pairs of strategies (A_1B_1, A_2B_1) and (A_1B_2, A_2B_1). 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 $cov\{dM_{11i}(s), dM_{11i}(s)|\mathcal{F}(s_-)\} = Y_{11i}(s)d\Lambda_0(s)$ and $cov\{dM_{12i}(s), dM_{11i}(s)|\mathcal{F}(s_-)\} = \{1 - R_i(s)\}Y_{1i}(s)d\Lambda_0(s)$. 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 3.2, we find $\sigma_{12}(t)$

$$
\begin{aligned}
=&E\left[\int_0^t \frac{\bar{Y}_{12}(s)\bar{Y}_{21}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}} \sum_{i=1}^n W_{11i}^2(s)Y_{1i}(s)d\Lambda_0\right. \\
&\left. - \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{21}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}} \phi^{-2} \sum_{i=1}^n \{1 - R_{1i}(s)\}Y_{1i}(s)d\Lambda_0(s)\right] \\
=&E\left[\int_0^t \frac{\bar{Y}_{21}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}} \left\{\bar{Y}_{12}(s)\sum_{i=1}^n W_{11i}^2(s)Y_{1i}(s) - \phi^{-2}\bar{Y}_{11}(s)Y_1^{NR}(s)\right\} d\Lambda_0(s)\right]
\end{aligned}
$$

Similarly,

$$
\sigma_{13}(t) = E\left[\int_0^t \frac{\bar{Y}_{22}(s)}{\{\bar{Y}_{11}(s)+\bar{Y}_{12}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{22}(s)\}} \left\{\bar{Y}_{12}(s)\sum_{i=1}^n W_{11i}^2(s)Y_{1i}(s) - \phi^{-2}\bar{Y}_{11}(s)Y_1^{NR}(s)\right\} d\Lambda_0(s)\right] \qquad (17)
$$

$$
\sigma_{23}(t) = E\left[\int_0^t \frac{1}{\{\bar{Y}_{11}(s)+\bar{Y}_{21}(s)\}\{\bar{Y}_{11}(s)+\bar{Y}_{22}(s)\}} \left\{\bar{Y}_{21}(s)\bar{Y}_{22}(s)\sum_{i=1}^n W_{11i}^2(s)Y_{1i}(s) + (1-\phi)^{-2}\bar{Y}_{11}^2(s)Y_2^{NR}(s)\right\} d\Lambda_0(s)\right]. \qquad (18)
$$

By substituting $d\Lambda_0(s)$ with its estimate $d\hat{\Lambda}_0(s) = dN(s)/Y(s)$, we have the consistent estimators s_{12}, s_{13}, s_{23} given in Section 4.

References

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.

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.

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.

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.

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,** S4S13.

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.

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.

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,**.

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.

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.

Thall, P. F., Millikan, R. E., and Sung, H.-G. (2000). Evaluting multiple treatment courses in clincial 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.

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). *Blood* **107,** 4207–13.

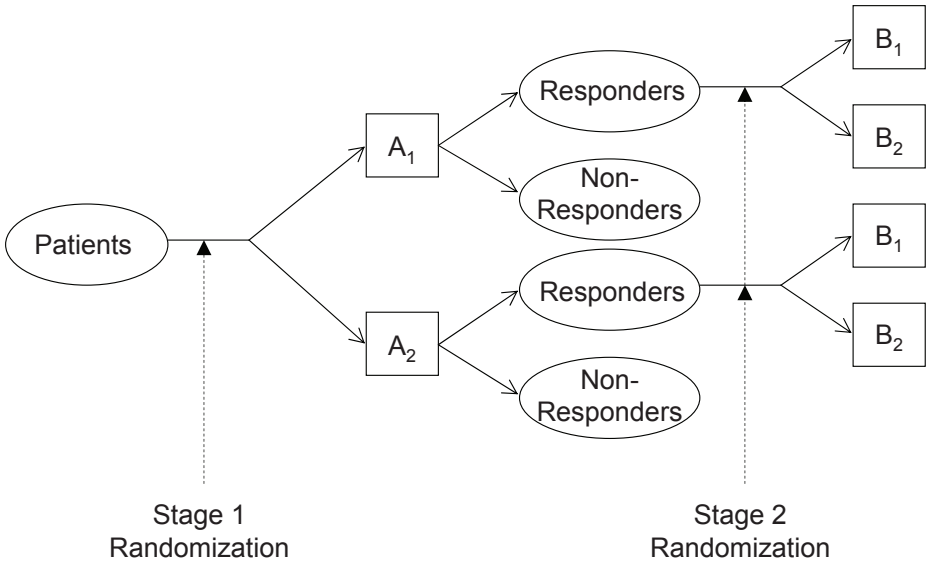

Figure 1: An example of a two-stage SMART design where only responders receive maintenance therapy

Table 1: At-risk and Event Process Notation

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

Table 2: Type I Error Rate under Null Hypotheses $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$. Target type I error rate is $\alpha = 0.05$; WLR is the proposed weighted log-rank statistic in equation (6); IWLR is the independent weighted log-rank statistic; SLR is the standard unweighted log-rank statistic.

		30% censoring			50% censoring		
Response Rate	n	WLR	IWLR	SLR	WLR	IWLR	SLR
40	50	0.056	0.005	0.004	0.033	0.001	0.001
	100	0.053	0.008	0.008	0.033	<0.001	<0.001
	200	0.047	0.007	0.009	0.037	0.001	<0.001
	300	0.041	0.005	0.008	0.033	0.001	<0.001
	400	0.044	0.009	0.008	0.039	<0.001	<0.001
	500	0.038	0.008	0.006	0.033	0.001	<0.001
60	50	0.053	0.015	0.014	0.047	0.008	0.006
	100	0.048	0.022	0.015	0.045	0.009	0.005
	200	0.042	0.017	0.014	0.041	0.010	0.005
	300	0.047	0.020	0.015	0.040	0.011	0.007
	400	0.039	0.014	0.011	0.035	0.009	0.005
	500	0.042	0.017	0.014	0.036	0.009	0.005

Table 3: Type I Error Rate under the Null Hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$. Target type I error rate is $\alpha = 0.05$; WLR is the weighted log-rank statistic in equation (14); SLR denotes the standard unweighted log-rank statistic.

		30% censoring		50% censoring	
Response Rate	n	WLR	SLR	WLR	SLR
40	50	0.058	0.047	0.044	0.047
	100	0.056	0.043	0.038	0.042
	200	0.045	0.038	0.040	0.043
	300	0.041	0.035	0.033	0.044
	400	0.047	0.040	0.034	0.039
	500	0.047	0.037	0.031	0.043
60	50	0.048	0.047	0.047	0.049
	100	0.046	0.040	0.040	0.038
	200	0.048	0.038	0.037	0.037
	300	0.049	0.041	0.041	0.038
	400	0.042	0.042	0.038	0.038
	500	0.043	0.045	0.040	0.043

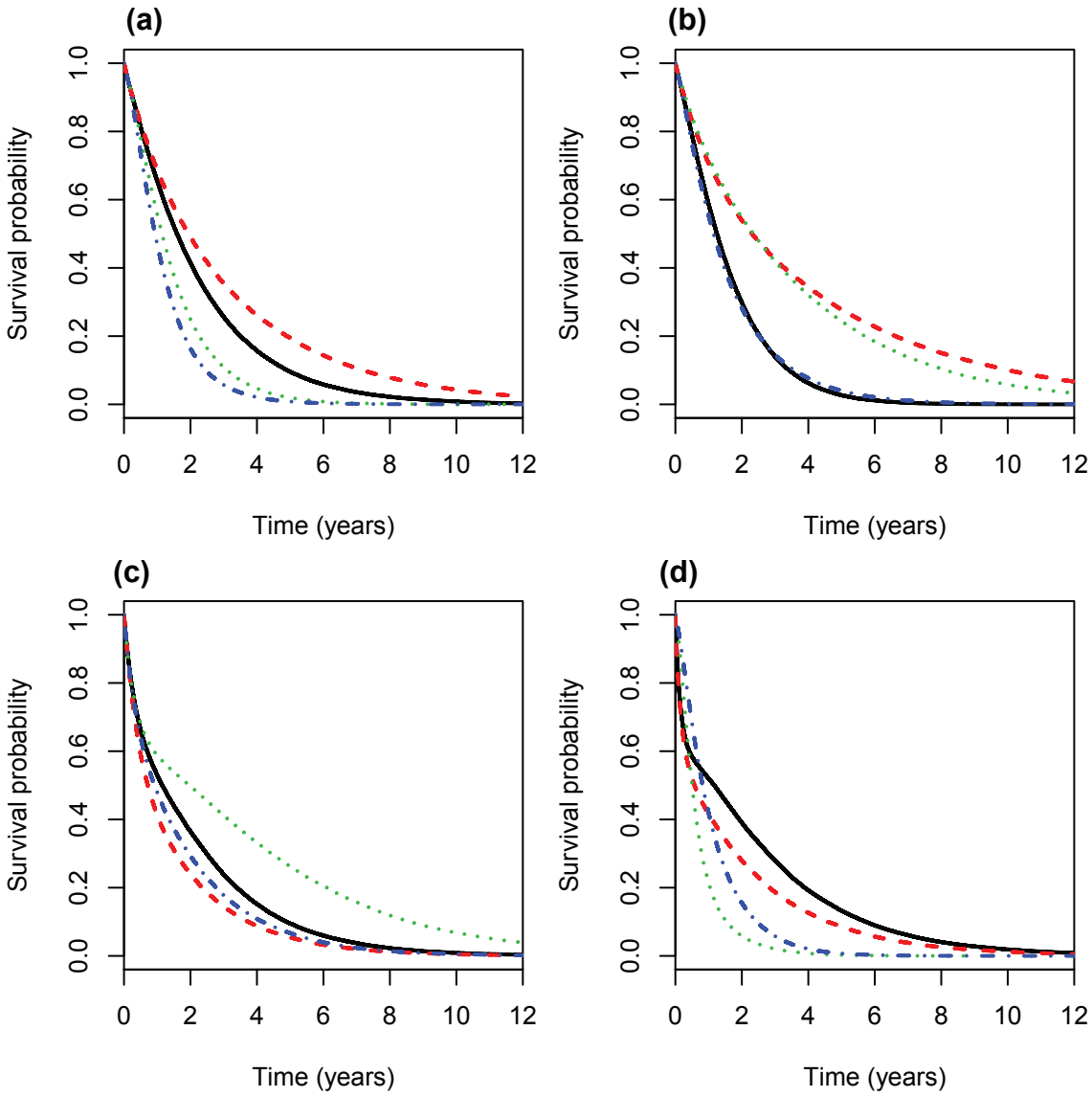

Figure 2: Survival curves for treatment strategies A_1B_1 (solid), A_1B_2 (dashes), A_2B_1 (dots), A_2B_2 (dot-dash), under different alternative hypotheses scenarios for 60% responders.

Table 4: Power against Alternative Survival Curves under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t)$. See Figure 2 and Section 5.3 for description of alternative survival curves (a)-(d); WLR is the proposed weighted log-rank statistic in equation (6); IWLR is the independent weighted log-rank statistic; SLR is the standard unweighted log-rank statistic.

				Power		
Scenario	Response Rate	Censoring Rate	n	WLR	IWLR	SLR
(a)	40	36	200	0.167	0.026	0.005
	60	44	200	0.237	0.096	0.028
(b)	40	35	200	0.886	0.542	0.130
	60	42	200	0.978	0.897	0.536
(c)	40	27	200	0.703	0.053	$<$0.001
	60	36	200	0.801	0.285	0.004
(d)	40	23	200	0.651	0.029	$<$0.001
	60	34	200	0.767	0.193	$<$0.001

Table 5: Power against Alternatives under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$. See Figure 2 and Section 5.3 for description of alternative survival curves (a)-(d); WLR is the weighted log-rank statistic in equation (14); SLR denotes the standard unweighted log-rank statistic.

				Power	
Scenario	Response Rate	Censoring Rate	n	WLR	SLR
(a)	40	31	200	0.503	0.162
	60	35	200	0.921	0.657
(b)	40	35	200	0.997	0.296
	60	42	200	1.000	0.796
(c)	40	29	200	0.921	0.084
	60	40	200	0.989	0.174
(d)	40	22	200	0.654	0.434
	60	27	200	0.964	0.366

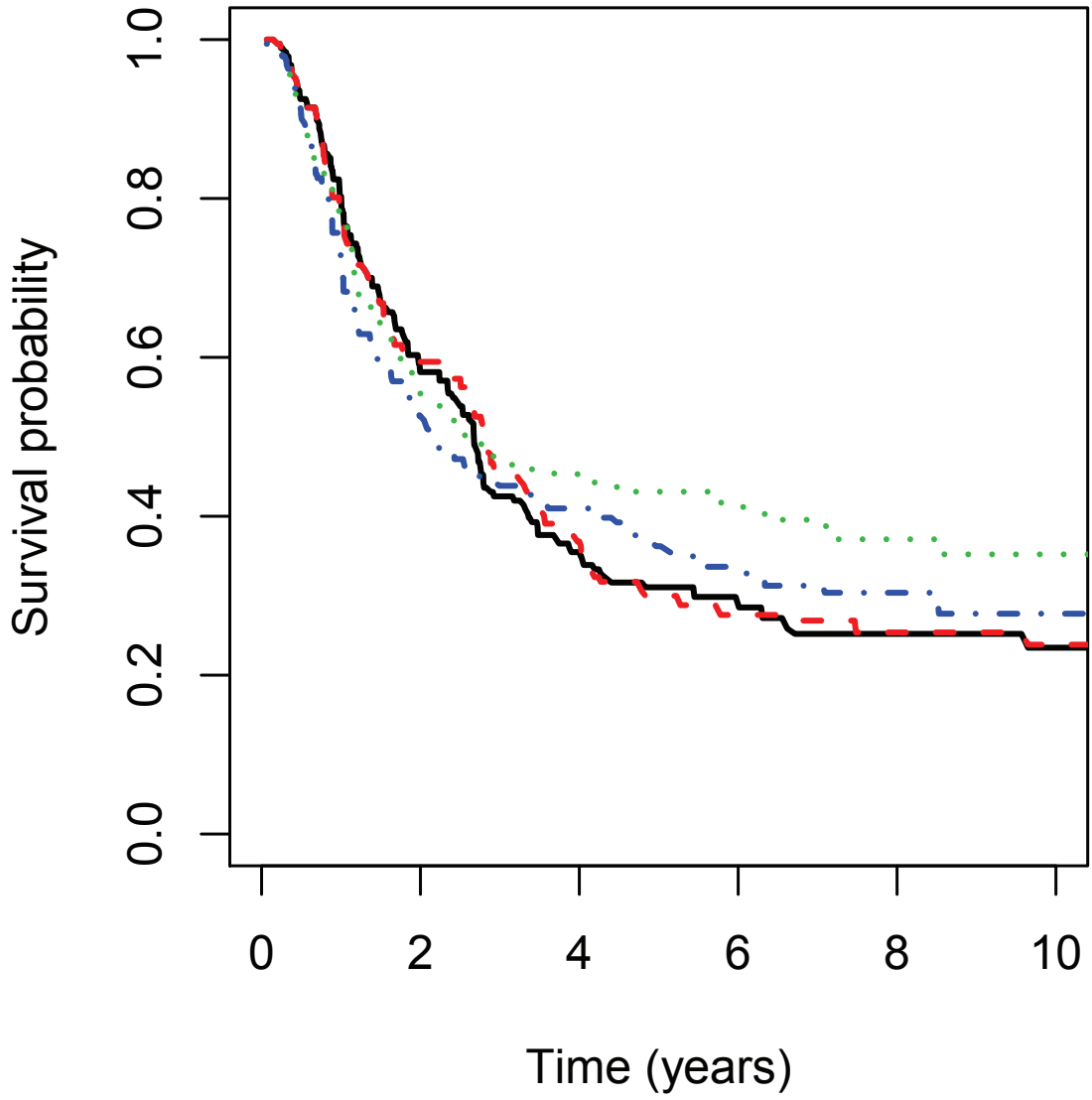

Figure 3: Weighted survival curves under four treatment strategies in the neuroblastoma study. CC (solid): 'Treat with chemotherapy followed by cis-RA if there is no disease progression'; CN (dashes): 'Treat with chemotherapy and if there is no disease progression, do not continue treatment'; AC (dots): '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'