Weighted log-rank statistic to compare shared-path adaptive treatment strategies

KELLEY M. KIDWELL*,
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
kidwell@umich.edu

ABDUS S. WAHED
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

SUMMARY
Adaptive treatment strategies (ATSs) 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 designs, are useful to assess ATSs 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 ATSs and propose a weighted log-rank statistic to compare overall survival distributions of multiple two-stage ATSs, some of which may be shared-path. Large sample properties of the statistic are derived and the type I error rate and power of the test are compared with the standard log-rank test through simulation.

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

1. INTRODUCTION
Physicians rarely choose a treatment for a patient randomly from competing treatments; rather they prescribe treatments based on their clinical experience in treating patients with similar characteristics and those patients’ responses 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 like treatment strategies that are adopted by physicians in practice. The SMART design allows for the assessment and comparison of adaptive treatment strategies (ATSs, 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 ATSs.

To better illustrate the emerging paradigm of ATS, 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, for example, “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 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 second-stage treatment are compared in order 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 and others, 2001; Stroup and others, 2003; Rush and others, 2004; Winter and others, 2006; Marlowe and others, 2007; Matthay and others, 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 and others, 2000; Murphy, 2003, 2005; Dawson and Lavori, 2010; Wahed and Tsiatis, 2004; Wahed, 2010; Orellana and others, 2010). This article focuses on time-to-event outcome data and hence the review of literature will emphasize statistical methods for survival analysis in SMART designs.

The first valid attempt in developing a test comparing overall survival curves under two ATSs was taken by Guo (2005) in his unpublished PhD dissertation, where an inverse-weighted version of the log-rank test for comparing two separate-path ATSs (strategies that do not share the same treatment paths; see Section 2) was developed. Lokhnygina and Helterbrand (2007) extended the idea of Lunceford and others (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. For the same comparison, Feng and Wahed (2008) presented a supremum weighted log-rank statistic, while Li and Murphy (2011) formally presented the statistic proposed in Guo (2005) with both time-dependent and time-independent weights. These tests, however, only compare the overall survival of two separate-path ATSs.

The goal of this article is to present methods for comparing overall survival curves under multiple ATSs, some of which may be shared-path (strategies that share some of the same treatment paths; see Section 2), using test statistics similar to \( k \)-sample log-rank tests (Harrington and Fleming, 1982). A naive approach to comparing survival curves of multiple ATSs is to form groups where each group includes all of the patients who follow each ATS and then to apply the standard unweighted log-rank (SLR) test. This approach, however, forms groups that contain some of the same patients violating the standard log-rank assumption that groups are statistically independent.

Comparison of shared-path ATSs is challenging since the correlation between survival curves needs to be accounted for in the estimation process. Accounting for this correlation allows us to compare treatment strategies that share the same initial treatment. When comparing only two shared-path ATSs, a standard log-rank test can be used where patients sharing the common treatment path are excluded from the comparison, making the two groups independent. However, our goal is to develop a log-rank test for comparing the
overall survival between multiple ATSs, all of which may not have a common shared path. Thus, in this paper, we first motivate our goal by presenting the theoretical foundation of a WLR statistic to compare the overall survival distributions of two shared-path ATSs and then extend this methodology to compare multiple ATSs.

2. Setup

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 initial treatment $A$, receive maintenance treatment $B$, and randomly allocated to levels $B_1$ or $B_2$ (see Figure 1). For simplicity, we 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 various treatment strategies $A_j B_k$, $j, k = 1, 2$, where strategy $A_j B_k$ is defined as follows.

**Definition 1** ATS $A_j B_k$: Treat with $A_j$ followed by $B_k$ if the patient is eligible and consents to subsequent second-stage therapy.

Furthermore, we classify strategies into shared-path and separate-path as follows.

**Definition 2** Shared-path ATSs: Two-stage ATSs 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 $A_1 B_1$ and $A_1 B_2$. Strategy $A_1 B_1$ dictates that a patient be treated with $A_1$ and then by $B_1$ only if the patient responds to $A_1$. Similarly, strategy $A_1 B_2$ 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 $A_1 B_1$ but did not respond to $A_1$ will receive exactly the same sequence of treatment as a patient who is treated under strategy $A_1 B_2$ but did not respond. Therefore, strategies $A_1 B_1$ and $A_1 B_2$ are shared-path ATSs. Similarly, the pair $(A_2 B_1, A_2 B_2)$ 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 $A_1 B_1$ and $A_2 B_1$ are separate-path ATSs since patients treated with $A_1 B_1$ cannot receive a treatment sequence received by patients treated with $A_2 B_1$. Similarly, pairs $(A_1 B_1, A_2 B_2)$, $(A_1 B_2, A_2 B_1)$, and $(A_1 B_2, A_2 B_2)$ are also separate-path.
2.1 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 \( A_j B_k \) and one can envision one outcome for each possible strategy for each individual, hence every individual has his/her own set of possible outcomes for every possible treatment strategy. This entire set of possible outcomes for an individual is referred to as his/her counterfactual outcomes.

In order to define patients’ counterfactual outcomes, which in this setting are the potential survival times, we introduce the following notation. Let \( R_{ji} = 1 \) if the \( i \)th patient responded to the initial treatment \( A_j \) and \( R_{ji} = 0 \) if the \( i \)th patient did not respond. 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 \( T_{ji}^R \) denote the survival time for patient \( i \) if he/she responded to treatment \( A_j \) and received treatment \( B_k \). For treatment strategy \( A_j B_k \), 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 \((R_{1i}, R_{2i}, T_{1i}^{NR}, T_{2i}^{NR}, T_{1i}^R, T_{2i}^R, T_{1i}^{21}, T_{2i}^{21}, T_{2i}^{22})\) 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 \( A_j B_k \), the potential survival time, \( T_{ji} \), can be expressed in terms of his/her counterfactual outcomes as \( T_{ji} = (1-R_{ji})T_{ji}^{NR} + R_{ji}T_{ji}^R \). Let us define \( \Lambda_{jk}(t) \) to be the cumulative hazard associated with \( T_{jk} \).

2.2 Observed data and assumptions

The observed data for a two-stage design described in Figure 1 can be represented as a set of random vectors \( \{X_i, R_i, R_i T_i^R, R_i Z_i, U_i, \delta_i\} \), for \( i = 1, \ldots, n \), where \( X_i = 2 - j \) if the \( i \)th 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 \( i \)th 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 \( R_i = X_i R_{ji} I(C_i > T_i^R) + (1-X_i) R_{ji} I(C_i > T_i^R) \), where \( R_{ji} \) is the counterfactual response defined in Section 2.1.

First, we make the stable unit treatment value or consistency assumption (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 \( T_i = \sum_{j=1}^{2} X_{ji} [(1 - R_{ji}) T_{ji}^{NR} + R_{ji} (Z_i T_{ji}^R + (1-Z_i) T_{ji}^{21})] \), 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 the random assignment of treatments (Orellana and others, 2010).

3. Theoretical foundation: log-rank statistic for comparing two shared-path strategies

In terms of the notation adopted in Section 2, our goal is to test the hypothesis \( H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t) \). Note that the four groups of individuals treated with these four strategies overlap, but the overlap is not common across all four groups. Therefore, it is not possible to apply the standard multiple-arm log-rank test to address this problem. Moreover, since the overlap is not common (e.g. the non-responders to \( A_1 \) are common to strategies \( A_1 B_1 \) and \( A_1 B_2 \), non-responders to \( A_2 \) are common to
strategies $A_2B_1$ and $A_2B_2$), it is not possible to form independent randomized groups representative of the corresponding populations. Thus, we will first develop a WLR test to test the equality of two shared-path ATSs (e.g. $H_0: \Lambda_{11}(t) = \Lambda_{12}(t)$) and then generalize it to serve our purpose. Even though a pair of shared-path strategies (e.g. $A_1B_1$ and $A_1B_2$) can be tested by applying the standard log-rank test to patients who responded to $A_1$, this does not extend to comparisons of strategies involving more than one initial treatment.

The SLR 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 $A_1B_2$ (test the null hypothesis of no difference between the two survival distributions) based on the observed data \{$U_{iki} = \min(1_{ki}, C_i)$, $\delta_{iki} = I(1_{ki} \leq C_i)$, $k = 1, 2$; $i = 1, \ldots, n$\}, we would use the standard two-sample unweighted log-rank test statistic

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

(3.1)

where $N_{1ki}(s) = I(U_{1ki} \leq s, \delta_{1ki} = 1)$, $Y_{1ki}(s) = I(U_{1ki} > s)$, $N_{1k}(s) = \sum_{i=1}^n N_{1ki}(s)$, and $Y_{1k}(s) = \sum_{i=1}^n Y_{1ki}(s)$ for $k = 1, 2$ to test $H_0: \Lambda_{11}(t) = \Lambda_{12}(t)$. Under the null hypothesis, $n^{-1/2}Z_n(t)$ is asymptotically normally distributed with mean zero and variance that can be consistently estimated from observed event times.

The SLR, however, does not account for the second randomization in a two-stage SMART design. In such a 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 $B_1$. The SLR also assumes independence, but non-responders to $A_1$ are consistent with both ATS $A_1B_1$ and $A_1B_2$, thus the non-responders to $A_1$ are common to both groups. Hence, the two groups of patients following ATS $A_1B_1$ and $A_1B_2$ are not statistically independent and the SLR fails to account for this.

Accounting for the second randomization has been addressed by Guo (2005), where he proposed a weighted version of the log-rank statistic. This statistic weights the at-risk and event processes according to the response status and randomization probability for each individual. Guo’s WLR statistic and the corresponding supremum version (Feng and Wahed, 2008), however, are only applicable to testing separate-path strategies (e.g. $A_1B_1$ versus $A_2B_1$). To account for the statistical dependency between strategies with overlap, we modify Guo’s log-rank statistic and present the univariate WLR statistic to compare two-shared path ATSs as motivation for the multivariate WLR statistic to be presented in Section 4.

### 3.1 Weighted two-sample statistic

Here, we first present the notation for time-dependent weights which is adapted from Guo and Tsiatis (2005). Explicitly, let $W_{11i}(s) = (X_i/\phi)[1 - R_i(s) + R_i(s)Z_i/\pi]$ be the weight assigned to the $i$th patient at time $s$ for the purpose of estimating quantities related to the strategy $A_1B_1$, where $R_i(s) = R_i(T_i^R \leq s)$, such that $R_i(s) = 1$ if the $i$th patient responded to $A_1$ by time $s$, 0 otherwise; $\pi$ is the known probability of a patient being assigned to maintenance therapy $B_1$; and $\phi$ is the known probability of a patient being assigned to first-stage therapy $A_1$. Similarly, $W_{12i}(s) = (X_i/\phi)[1 - R_i(s) + R_i(s)(1 - Z_i)/(1 - \pi)]$ for estimating quantities related to the strategy $A_1B_2$. Note that if a patient has not responded by time $s$, $W_{11i}(s) = W_{12i}(s) = 1/\phi$, 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 $s$, $W_{11i}(s) = 1/(\phi\pi)$ and $W_{12i}(s) = 0$; if the patient has responded and is randomized to $B_2$ by
The quantity \( d \) time \( s \), \( W_{11}(s) = 0 \) and \( W_{12}(s) = 1/\phi(1 - \pi) \). This construction of weights is based on the fundamental principle of inverse-probability-of-treatment-weighting (Robins and others, 1994).

We introduce further notation which is included in Table 1 for quick reference. The general at-risk process for all patients is \( Y_i(s) = I(U_i \geq s) \), the weighted at-risk process is \( \tilde{Y}_{jk}(s) = \sum_{i=1}^{n} W_{jki}(s) Y_i(s) \), the at-risk process for only those who are non-responders to \( A_j \) is \( Y_{NR}^{j}(s) = \sum_{i=1}^{n} I(X_i = 2 - j) \{1 - R_i(s)\} Y_i(s) \), the overall at-risk process for those 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 for any patient \( i \) is \( N_i(s) = I(U_i \leq s, \delta_i = 1) \), the weighted event process is \( \tilde{N}_{jk}(s) = \sum_{i=1}^{n} W_{jki}(s) N_i(s) \), the event process for those who are non-responders to \( A_j \) is \( N_{NR}^{j}(s) = \sum_{i=1}^{n} I(X_i = 2 - j) \{1 - R_i(s)\} N_i(s) \), the overall event process for patients treated with \( A_j \) is \( N_j(s) = \sum_{i=1}^{n} I(X_i = 2 - j) N_i(s) \), 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 WLR statistic for testing \( H_0: \Lambda_{11}(t) = \Lambda_{12}(t) \) is defined as

\[
Z_n^W(t) = \int_0^t \left\{ \frac{\tilde{Y}_{11}(s) \tilde{Y}_{12}(s)}{\tilde{Y}_{11}(s) + \tilde{Y}_{12}(s)} - \frac{d\tilde{N}_{11}(s)}{\tilde{Y}_{11}(s)} - \frac{d\tilde{N}_{12}(s)}{\tilde{Y}_{12}(s)} \right\} ds.
\]  

(3.2)

The rationale behind this formulation of the test statistic is given in Feng and Wahed (2008). In short, the quantity \( d\tilde{N}_{1k}(s) / \tilde{Y}_{1k}(s) \) is an unbiased estimator of the instantaneous cumulative hazard at time \( s \),
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 WLR statistic in equation (3.2) has expectation zero (see Section 3.2).

While the WLR statistic looks almost identical to that of the SLR, 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 SLR. The variance calculation will change substantially in order to account for this correlation between these two terms and address the remaining inadequacy of the standard and supremum log-rank tests.

3.2 Asymptotic properties

First, we note that $n^{-1/2}Z_n^W(t)$ in equation (3.2) 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 $G_n(t) = n^{-1/2} \int_0^t \bar{Y}_{11}(s)\bar{Y}_{12}(s)/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\{d\bar{M}_{11}(s)/\bar{Y}_{11}(s) - d\bar{M}_{12}(s)/\bar{Y}_{12}(s)\}$ and $R_n(t) = n^{-1/2} \int_0^t \bar{Y}_{11}(s)\bar{Y}_{12}(s)/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\{d\Lambda_{11}(s) - d\Lambda_{12}(s)\}$, since $\bar{M}_{jki} = \bar{N}_{jki} - \int_0^t \bar{Y}_{jki}(u)\ d\Lambda_{jki}(u)$.

Feng and Wahed (2008, p. 699) have shown that $d\bar{M}_{jki} = \sum_{i=1}^n W_{jki}(s)\ dM_{jki}(s)$ and it is easy to show that $E\{d\bar{M}_{jki}(s)/\mathcal{F}(s-\cdot)\} = 0$, where $M_{jki}(s)$ is the $i$th patient specific martingale, corresponding to $M_{jki} = N_{jki} - \int_0^t Y_{jki}(u)\ d\Lambda_{jki}(u)$, 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. Under the null hypothesis, $\Lambda_{11}(t) = \Lambda_{12}(t)$, so $n^{-1/2}Z_n^W(t) = G_n(t)$, and since martingale increments have mean zero, $E\{Z_n^W(t)\} = 0$.

To derive the variance of $n^{-1/2}Z_n^W(t)$, we further expand $G_n(t)$. Using $d\bar{M}_{jki} = \sum_{i=1}^n W_{jki}(s)\ dM_{jki}(s)$, $G_n(t)$ can be expressed as a difference of two martingale processes, $G_n(t) = G_n^{11}(t) - G_n^{12}(t) = n^{-1/2}\sum_{i=1}^n \int_0^t \bar{Y}_{11}(s)W_{11i}(s)/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\ dM_{11i}(s) - \sum_{i=1}^n \int_0^t \bar{Y}_{11}(s)W_{12i}(s)/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}\ dM_{12i}(s)$. By the martingale central limit theorem (Fleming and Harrington, 1991, Chapter 5), $G_n(t)$ converges to a Gaussian process with mean zero. Therefore, $G_n(t)$ converges to a Gaussian process with mean zero and variance equal to $\text{var}[G_n^{11}(t)] + \text{var}[G_n^{12}(t)] - 2\text{cov}[G_n^{11}(t), G_n^{12}(t)]$. The variances of $G_n^{11}(t)$ and $G_n^{12}(t)$ can be calculated the same way as the variance for the supremum WLR statistic in Feng and Wahed (2008).

To find the covariance between two martingale processes, $\text{cov}[G_n^{11}(t), G_n^{12}(t)]$, we use a 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 $\int H_1H_2\ d\text{cov}(M_1, M_2)$. Then, the asymptotic variance of $G_n(t)$ can be expressed as the limiting value of $n^{-1}\sum_{k=1}^n \sum_{i=1}^n [\bar{Y}_{11}(s)W_{11i}(s)/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}]^2 Y_i(s)\ d\Lambda_{1k}(s) - 2n^{-1}\sum_{k=1}^n \sum_{i=1}^n [\bar{Y}_{11}(s)\bar{Y}_{12}(s)/\{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}]^2 |\bar{Y}_{11}(s)|^2 Y_i(s)\ d\Lambda_{NR}(s)/\phi^2 = Y_i^{NR}(s)\ d\Lambda_{NR}(s)/\phi^2$. Thus, a consistent variance estimator of $n^{-1/2}Z_n^W(t)$ is given by

$$\hat{\sigma}^2(t) = n^{-1} \int_0^t \frac{\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \left\{ \frac{dN_{11}(s)}{Y_{11}(s)} + \frac{dN_{12}(s)}{Y_{12}(s)} \right\} - 2(n\phi^2)^{-1} \int_0^t \left\{ \frac{\bar{Y}_{11}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \right\}^2 \left\{ \frac{Y_1^{NR}(s)}{Y_1^{NR}(s)} \frac{dN_{11}^{NR}(s)}{Y_1^{NR}(s)} \right\}. \quad (3.3)$$

The notation used in the above equation or elsewhere in this article can be reviewed in Table 1. With the results derived in this section, we are now ready to present the WLR statistic to compare multiple ATSs.
4. Log-rank Tests for Multiple ATSSs

In the setting described above, we would now like to compare multiple adaptive strategies, explicitly all four ATSSs, \( A_j B_k \), \( j, k = 1, 2 \), and test the overall null hypothesis of no treatment effect, \( H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t) \), against the alternative hypothesis, \( H_1 \): at least one cumulative hazard differs.

To derive the multivariate WLR statistic, we first note that \( H_0 \) can be cast as a vector of differences between cumulative hazards such that \( H_0 : \vec{\zeta}(t) = 0 \) where \( \vec{\zeta}(t) = \{ \Lambda_{11}(t) - \Lambda_{12}(t), \Lambda_{11}(t) - \Lambda_{21}(t), \Lambda_{11}(t) - \Lambda_{22}(t) \}^T \). Following Section 3, an unbiased estimator of \( \vec{\zeta}(t) \) is given by \( \hat{\vec{\zeta}}(t) = \int_0^t \{ d\tilde{N}_{11}(s)/\tilde{Y}_{11}(s) - d\tilde{N}_{12}(s)/\tilde{Y}_{12}(s), d\tilde{N}_{11}(s)/\tilde{Y}_{11}(s) - d\tilde{N}_{21}(s)/\tilde{Y}_{21}(s), d\tilde{N}_{11}(s)/\tilde{Y}_{11}(s) - d\tilde{N}_{22}(s)/\tilde{Y}_{22}(s) \}^T \). The corresponding WLR statistic is the vector of weighted martingale differences, \( \hat{Z}_n^{MW} = \{ \hat{Z}_{11,12}^{WR}(t), \hat{Z}_{11,21}^{WR}(t), \hat{Z}_{11,22}^{WR}(t) \}^T \) where

\[
\hat{Z}_n^{jk, j'k'}(t) = \int_0^t \frac{\tilde{Y}_{jk}(s)}{\tilde{Y}_{jk}(s) + \tilde{Y}_{j'k'}(s)} \left\{ \frac{d\tilde{N}_{jk}(s)}{\tilde{Y}_{jk}(s)} - \frac{d\tilde{N}_{j'k'}(s)}{\tilde{Y}_{j'k'}(s)} \right\}. \tag{4.1}
\]

Under the null hypothesis, the statistic \( \hat{Z}_n^{MW} \) has expectation zero. Since \( \hat{Z}_n^{MW} \) is a linear combination of weighted \( Z_n^{W} \)-statistics defined in equation (3.2), by the multivariate central limit theorem for martingales (Fleming and Harrington, 1991), \( n^{-1/2} \hat{Z}_n^{MW} \) follows a mean zero Gaussian process with asymptotic variance–covariance matrix, \( \Sigma(t) \), that can be estimated by \( \hat{\Sigma}(t) = \{ s_{pq}(t) \}_{3 \times 3} \) where the elements of \( \hat{\Sigma}(t) \) are defined as follows.

The diagonal elements of \( \hat{\Sigma}(t) \) are given below, where \( d\hat{\Lambda}_0(s) \), previously estimated in equation (3.3) by using the induction-treatment-specific processes, \( \hat{N}_1(s) \) and \( \hat{Y}_1(s) \), is now estimated by using the overall processes, \( N(s) \) and \( Y(s) \), to reflect that under \( H_0 \), all strategies have equal hazards. Explicitly,

\[
s_{11}(t) = n^{-1} \int_0^t \frac{\tilde{Y}_{12}(s) \sum_{i=0}^n W_{11i}(s) Y_i(s) + \tilde{Y}_{11}(s) \sum_{i=0}^n W_{11i}(s) Y_i(s)}{\{ \tilde{Y}_{11}(s) + \tilde{Y}_{12}(s) \}^2} d\hat{\Lambda}_0(s),
\]

\[
s_{22}(t) = n^{-1} \int_0^t \frac{\tilde{Y}_{21}(s) \sum_{i=0}^n W_{11i}(s) Y_i(s) + \tilde{Y}_{21}(s) \sum_{i=0}^n W_{11i}(s) Y_i(s)}{\{ \tilde{Y}_{11}(s) + \tilde{Y}_{21}(s) \}^2} d\hat{\Lambda}_0(s), \tag{4.3}
\]

and

\[
s_{33}(t) = n^{-1} \int_0^t \frac{\tilde{Y}_{22}(s) \sum_{i=0}^n W_{11i}(s) Y_i(s) + \tilde{Y}_{22}(s) \sum_{i=0}^n W_{11i}(s) Y_i(s)}{\{ \tilde{Y}_{11}(s) + \tilde{Y}_{22}(s) \}^2} d\hat{\Lambda}_0(s). \tag{4.4}
\]

where \( d\hat{\Lambda}_j^{NR}(s) = d\tilde{N}_j^{NR}(s)/\tilde{Y}_j^{NR}(s) \). Note that the last two formulas above do not contain a covariance term since \( d\tilde{N}_{jk}(s)/\tilde{Y}_{jk}(s) \) and \( d\tilde{N}_{j'k'}(s)/\tilde{Y}_{j'k'}(s) \), \( j \neq j' \), in \( \hat{Z}_{11,12}^{WR}(t) \) and \( \hat{Z}_{11,22}^{WR}(t) \), are conditionally independent given \( F(s) \). To obtain the estimated covariance terms, we first derive the expressions for the covariances (see supplementary material available at Biostatistics online). Following those expressions, the off-diagonal elements...
of $\hat{\Sigma}(t)$ are given by

\[
s_{12}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{21}(s)}{\omega_{1221}(s)} \left\{ \bar{Y}_{12}(s) \sum_{i=1}^n W_{11i}(s) Y_i(s) \, d\hat{A}_0(s) - \frac{1}{\phi^2} \bar{Y}_{11}(s) Y_{1R}(s) \, d\hat{A}_{1R}(s) \right\}, \quad (4.5)
\]

\[
s_{13}(t) = n^{-1} \int_0^t \frac{\bar{Y}_{22}(s)}{\omega_{1222}(s)} \left\{ \bar{Y}_{12}(s) \sum_{i=1}^n W_{11i}(s) Y_i(s) \, d\hat{A}_0(s) - \frac{1}{\phi^2} \bar{Y}_{11}(s) Y_{1R}(s) \, d\hat{A}_{1R}(s) \right\}, \quad (4.6)
\]

\[
s_{23}(t) = n^{-1} \int_0^t \frac{1}{\omega_{2122}(s)} \left\{ \bar{Y}_{22}(s) \sum_{i=1}^n W_{11i}(s) Y_i(s) \, d\hat{A}_0(s) + \frac{1}{(1-\phi)^2} \bar{Y}_{22}(s) Y_{2R}(s) \, d\hat{A}_{2R}(s) \right\}, \quad (4.7)
\]

where $d\hat{A}_0(s) = dN(s)/Y(s)$, $d\hat{A}_{jR}(s) = dN_{jR}(s)/Y_{jR}(s)$, and $\omega_{j,k,j'k'}(s)$ is the $j$th row and $k$th column of the matrix $\Omega(s)$. The vector of WLR statistics, $n^{-1/2}Z_n^{MW}(t)$ converges in distribution under the null hypothesis to a trivariate normal distribution with mean zero and variance–covariance matrix $\Sigma(t)$, consistently estimated by $\hat{\Sigma}(t)$. By multivariate Slutsky’s theorem, we have that $n^{-1/2}Z_n^{MW}(t) \mathcal{N}(t) Z_n^{MW}(t)$ converges in distribution under the null hypothesis to a chi-square distribution with three degrees of freedom. The WLR test statistic comparing overall survival distributions for ATSs $A_jB_k$, $j, k = 1, 2$, is then expressed in the form

\[
T_n^{MW}(L) = n^{-1/2}Z_n^{MW}(L) \mathcal{N}(L) Z_n^{MW}(L),
\]

where $L$ is some time less than the maximum follow-up time. The level $\alpha$ WLR test rejects the overall equality of the ATSs’ cumulative hazards when $T_n^{MW}(L) \geq \chi_{\alpha/3}^2$, where $\chi_{\alpha/3}^2$ is the $(1-\alpha)$th quantile of a chi-square distribution with three degrees of freedom.

5. Simulation Results

To evaluate the performance of the WLR statistics for comparing multiple ATSs, 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 WLR statistics under various alternative scenarios. The performance of our WLR statistic, $T_n^{MW}(L)$ from equation (4.8), is compared with the SLR test applied to groups of patients who followed each strategy. The groups for the SLR test were formed by combining those who did not respond to $A_j$ with those who responded to $A_j$ and received treatment $B_k$.

We outline the data generation process here and provide specific parameters for each simulation in supplementary material available at Biostatistics online. The initial treatment indicator $X_t$ was generated from a Bernoulli distribution with $\phi = \text{pr}(X_t = 1) = 0.5$ so that there were about an equal number of patients initially treated with $A_1$ and $A_2$. We took $R_t$, the response indicator, to be Bernoulli with $\text{pr}(R_t = 1) = \pi_R$, $\pi_R \in (0.4, 0.6)$, so that there were 40% or 60% of patients who responded to the initial treatment. When $R_t = 0$, a survival time $T_{ji}^{NR}$, $j = 1, 2$, was generated from an exponential distribution with mean $\mu_{ji}^{NR}$. When $R_t = 1$, the treatment $B_1$ indicator $Z_t$ was generated from a Bernoulli ($\pi = 0.5$) distribution. Also when $R_t = 1$, time to response $T_{ji}^{RE}$, $j = 1, 2$, was generated from an exponential distribution with mean $\theta_j^{RE}$ and time from response to an event $T_{jki}^{RE}$, $j, k = 1, 2$, was generated from an exponential distribution with mean $\theta_{jki}^{RE}$. The total survival time for those who responded to $A_j$ and were randomized to $B_k$ is, thus, $T_{jki} = T_{ji}^{NR} + R_t T_{jki}^{RE}$, for $j, k = 1, 2$. The variables of interest here are the time-to-events $T_{jki}$, where $T_{jki} = (1 - R_t) T_{ji}^{NR} + R_t T_{jki}^{RE}$, $j, k = 1, 2$. These variables reflect the overall survival
Table 2. The type I error rate under null hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$

<table>
<thead>
<tr>
<th>Censoring rate</th>
<th>Response rate</th>
<th>$n$</th>
<th>WLR</th>
<th>SLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>40</td>
<td>250</td>
<td>0.056</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.052</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.053</td>
<td>0.039</td>
</tr>
<tr>
<td>60</td>
<td>250</td>
<td>0.054</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.058</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.056</td>
<td>0.038</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>250</td>
<td>0.053</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.053</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.055</td>
<td>0.042</td>
</tr>
<tr>
<td>60</td>
<td>250</td>
<td>0.051</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.052</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.054</td>
<td>0.041</td>
</tr>
</tbody>
</table>

The target type I error rate is $\alpha = 0.05$; WLR is the proposed WLR statistic in equation (4.8); SLR is the standard unweighted log-rank statistic.

Table 3. Power against alternative survival curves under $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$ for a sample size of 250

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Response rate</th>
<th>Censoring rate</th>
<th>WLR</th>
<th>SLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>40</td>
<td>31</td>
<td>0.613</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>35</td>
<td>0.973</td>
<td>0.770</td>
</tr>
<tr>
<td>(b)</td>
<td>40</td>
<td>34</td>
<td>0.985</td>
<td>0.311</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>41</td>
<td>0.999</td>
<td>0.752</td>
</tr>
<tr>
<td>(c)</td>
<td>40</td>
<td>28</td>
<td>0.663</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>33</td>
<td>0.833</td>
<td>0.178</td>
</tr>
<tr>
<td>(d)</td>
<td>40</td>
<td>23</td>
<td>0.993</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>26</td>
<td>0.980</td>
<td>0.614</td>
</tr>
</tbody>
</table>

See Section 5 for a description of alternative survival curves (a)–(d); WLR is the proposed WLR statistic in equation (4.8); SLR is the standard unweighted log-rank statistic.

time under strategy $A_j B_k$ ($j, k = 1, 2$). The observed survival time for the $i$th individual in the absence of censoring is defined as $T_i = X_i [R_i (T_{i11}^* + (1 - Z_i) T_{i12}^*)] + (1 - R_i) T_{iNR}^*$. Additionally, a right censored time $C_i$ was generated independently from a uniform distribution from zero to $v$, such that 30% or 50% of the population were censored. Censoring was independent and uninformative of the response and the survival time. The final observed time was then defined as $U_i = \min(T_i, C_i)$ with corresponding complete case indicator $\delta_i = I(T_i \leq C_i)$. If an individual was censored before the response, the individual was recoded as a non-responder.

For each generated dataset, we conducted the WLR test described in Section 4 to test the hypothesis $H_0 : \Lambda_{11}(t) = \Lambda_{12}(t) = \Lambda_{21}(t) = \Lambda_{22}(t)$. We report the estimated type I error (proportion of samples for which the hypothesis was falsely rejected) in Table 2 when $H_0$ was true, and the estimated power (proportion of samples for which the hypothesis was correctly rejected) in Table 3.

Table 2 presents the estimated type I error rates for the WLR and SLR tests. The type I error rates for both statistics were similar across all combinations. For a response rate of 40%, censoring of 30%, and a sample size of 500, the WLR test 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 combines and equally weights all patients who follow a strategy regardless of their response status. When censoring was increased to 50%, the WLR test and the SLR test produced estimated type I error rates of 0.053 and 0.045, respectively. Increasing the response rate to 60% produced acceptable type I error rates for sample sizes >500, with similar results of estimated type I error rates for both the WLR and the SLR tests around the nominal 0.05 level.

Since the type I error rates were upheld, we explored a variety of scenarios performing 5000 iterations to test the power of the WLR test. Data were generated from populations under the alternative hypotheses (see supplementary material available at Biostatistics online), where four true survival distributions, designated as scenarios (a)–(d), were plotted and presented in Figure 2 when 60% of the population respond to $A_j$, $j = 1, 2$.

Table 3 presents the power for comparing the survival distributions of four ATSs. In all cases, increasing the response rate from 40% to 60% increased the power of both the WLR and SLR. In almost all of the scenarios tested, the WLR test had greater power to correctly reject the null hypothesis. Specifically, in scenario (c), where the survival of strategy $A_2B_1$ dominates the other strategies, we see that, for a 40% response rate and about 28% censoring, the WLR test had power of 0.663, whereas the SLR test had power of only 0.085. In one of the scenarios (not presented here), the WLR test had less power than the SLR test for 40% and 45% responders. This may be due to the disparate percentage of censoring among responders.
and 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 WLR 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 ATSs in most situations, including cases where the proportional hazards assumption was violated.

6. Discussion

ATSs 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 ATSs 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 two separate-path ATSs, methods for comparing the overall survival distributions of multiple ATSs that can share common paths are not available in the literature. These shared-path ATSs share a common path of treatment such that there is a common group of patients who are consistent with more than one ATS in the data collected through SMART designs. To address this, we have proposed a WLR statistic that 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 this test and we have shown that the proposed WLR statistic comparing multiple ATSs generally maintains type I error rates and has greater power than the naive method 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 the response status. More research is needed to incorporate informative censoring into the WLR statistic. Future research in this area also includes the extension of the WLR statistic to compare survival distributions of patients who follow ATSs in general (multi-stage) SMART designs.

Supplementary material

The derivation of the covariance expressions from Section 4 and the parameters for simulations under the null and alternative hypotheses discussed in Section 5 are contained in the supplementary material, which is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

Conflict of Interest: None declared.

References


Weighted log-rank to compare ATS


[Received November 3, 2011; revised October 8, 2012; accepted for publication October 8, 2012]