

Sample Size Calculation for Comparing Two-Stage Treatment Strategies with Censored Data

Zhiguo Li^{1*} and Susan Murphy^{1,2}

¹*Institute for Social Research,
University of Michigan, Ann Arbor, MI 48106, U.S.A.*

²*Department of Statistics,
University of Michigan, Ann Arbor, MI 48109, U.S.A.*

Abstract

In two-stage randomization trials subjects are first randomized to an initial treatment, and then if and when a response (or nonresponse) criterium is met, re-randomized to a second-stage treatment. When the primary outcome is a time to event, sample size calculation is challenging because the variances of common estimators depend in a complex manner on the joint distribution of primary time to event outcome and time to response (or nonresponse) to the initial treatment. To circumvent this problem we construct simple upper bounds of the variances; the use of these upper bounds allows us to use the same working assumptions needed to size a single stage trial. The use of the upper bounds necessarily yields conservative sample sizes when all working assumptions are correct, but has the advantage of robustness to incorrect working assumptions. The tests we consider for sample size calculation include a test based on the weighted Kaplan-Meier estimator of survival probabilities at a fixed time point and the weighted log rank test. We also consider several variants of the two-stage randomized design.

KEY WORDS: censored data; sample size calculation; two-stage randomization; weighted Kaplan-Meier estimator; weighted log rank test

*Correspondence to: Zhiguo Li, Survey Research Center, ISR 2051, Ann Arbor, MI 48106, U.S.A.

†E-mail: zhiguo@umich.edu

1. INTRODUCTION

Sequential, multiple assignment randomized or two-stage randomized trials [1, 2] are utilized in developing adaptive treatment strategies (also called treatment policies or dynamic treatment regimes) for treating chronic diseases and conditions. In these clinical trial designs each subject may be randomized multiple times. For example, in a recently concluded clinical trial involving children with ADHD (Pelham, personal communication) each child is first randomized to one of two initial treatments (a behavioral treatment or a medication) and then if the child experiences early signs of nonresponse, the child is rerandomized to one of two more intensive treatments (intensification of current treatment or a combined treatment). An interesting outcome in this trial is time until a major disciplinary event. A more well-known example of this type of trial is CATIE [3]; this 18-month trial involving schizophrenics initially randomized each subject to one of five drugs. If the subject did not respond (nonresponse due to either tolerance or lack of efficacy), the subject was rerandomized to one of several further medications, while responders stay on the assigned medication for the duration of 18-month treatment period. The primary outcome is time until all-cause treatment discontinuation. In both Pelham's study and in CATIE, nonresponders are further randomized to another treatment, but responders stay on the initial treatment. In trials in cancer, typically responders to the initial induction treatment are randomized to a maintenance treatment, but nonresponders are given a fixed further treatment. See [4] for an example. In contrast, a soon-to-begin trial involving pregnant women who are drug dependent will randomize both responders and nonresponders to a second treatment (Jones, personal communication). One of the primary outcomes in this trial is time until treatment discontinuation.

An important goal in two-stage randomized trials is to compare different adaptive treatment strategies. An adaptive treatment strategy (or treatment policy or dynamic treatment regime) is a sequence of decision rules one per stage of the treatment [1, 5, 6]. For example, one possible adaptive treatment strategy is: offer drug A initially, and if there is insufficient response, then offer drug B , however if the patient responds, continue to provide the initial drug A . In cases where the variable of interest is time to an event, a comparison of adaptive treatment strategies can be based on survival probabilities at a certain time point, for example, survival probabilities at the end of the study period. The test statistic for this comparison can be based on estimators of the survival functions. Various methods have been proposed to estimate the survival function of the time to event for an adaptive treatment strategy. Lunceford et al. [1] proposed several versions of a weighted sample proportion estimator, Wahed and Tsiatis [7] derived a semiparametric efficient estimator and a less efficient but easier to implement estimator, [8] proposed a weighted Nelson-Aalen estimator of the cumulative hazard function, and recently, Miyahara and Wahed (personal communication) proposed a weighted Kaplan-Meier estimator. These estimators provide the basis for different tests for comparing adaptive treatment strategies using data from a two-stage ran-

domized trial. Another test for comparing two adaptive treatment strategies is a weighted version of the usual log rank test [9]. All of the work mentioned above considers a two-stage randomized design in which there are two options for the first stage treatment and only responders (or nonresponders) to the first-stage treatment are further randomized to one of two options for second-stage treatment. However, these test statistics can be adapted for the more general designs that may involve randomization of both responders and nonresponders and permit the number of treatment options at each stage to vary across stages.

An important issue in the design of a two stage randomized trial is sample size calculation. One approach to sizing a two stage randomized trial is to guarantee a given power to detect a difference between two predetermined adaptive treatment strategies (these two may be the most extremal in terms of dose and intensity). In this paper we investigate sample size formulae for designing a two-stage randomized trial to permit the comparison of two predetermined adaptive treatment strategies with a given power. In particular we focus on the development of relatively simple, easy to use and robust sample size formulae that require similar information to that needed to size a two group randomized trial. We provide sample size formulae based on two different tests: a test of the equality of survival probabilities at one time point based on a weighted Kaplan-Meier estimator and a test of the equality of hazard functions based on a weighted log rank test. The major challenge to developing easy to use sample size formulae is that the variances involved in the test statistics are complex and in addition these variances depend on the joint distribution of time to (non) response to the initial treatment and time to the primary event; these two times are likely to be dependent. For example in CATIE the time to non-response to initial treatment and the time to treatment discontinuation are likely dependent.

To achieve the goal of simple, easy to use, sample size formulae that require no more working assumptions that would be required in sizing a two group randomized trial, we use three tools. First we express the variances involved in the sample size formulae in terms of potential outcomes; this makes it easier to identify the quantities for which we need working assumptions. Second we use time independent weights instead of time dependent weights to construct the test statistics that will be used to produce sample size formulae. This simplifies the sample size formula somewhat. We further simplify the sample size formulae by replacing the variances by upper bounds. This means that the proposed formulae will be conservative. In fact the sample size formulae will be conservative for two reasons: first we recommend using the more efficient, time dependent weights (see the next sections for details) in the data analysis and second conservatism is induced by the use of upper bounds for the variances.

The development of sample size formulae for two stage randomized trials is not new. When there is no censored data, [2] studied sample size calculation for two-stage designs with continuous primary outcomes. In the case of time to event outcomes sample size

formulae have been proposed and studied by [10] and [11]. Feng and Wahed [10] proposed using the supremum weighted log rank test for sample size calculation; in [11], the sample size is calculated using a test statistic based on a weighted sample proportion estimator of survival probabilities from [1].

The paper is organized as follows. For simplicity of presentation, in Sections 2 through 4, we discuss the test statistics and sample size formula methods in the setting of the typical two-stage designs. In Section 2, we describe the test statistic based on weighted Kaplan-Meier estimator and the weighted version of the log rank test. Here we generalize the time independent weighted Kaplan-Meier estimator of Miyahara and Wahed to a weighted Kaplan-Meier estimator using time dependent weights and we provide a formula for variance of the asymptotic distribution of the weighted Kaplan-Meier estimator. In Section 3, we discuss sample size calculation based on time independent weighted versions of the two test statistics. In Section 4 we compare, via a simulation study, the proposed sample size formula methods with the most promising methods presently in the literature. In Section 5, we show how our methods can be used for the more general class of designs and provide specific sample size formulae for two examples. We conclude with a discussion in Section 6. All proofs are in the Appendix.

2. TEST STATISTICS

For simplicity of notation, from this section through Section 4, we specialize to the two-stage randomized trial in which only responders to initial treatment are rerandomized. Suppose that there are n subjects who are initially randomized to one of two treatments, denoted by $A_1 = 1, 2$. Responders are further randomized to one of two second-stage treatments, denoted by $A_2 = 1, 2$; for now, we assume that the two second stage treatments do not vary by the initial treatment. Nonresponders are offered a fixed second-stage treatment or stay on the initial treatment. Denote the randomization probabilities by $p = P(A_1 = 1)$ and $q = P(A_2 = 1 | R = 1)$. On each subject we observe $\{A_1, R, RA_2, S', \Delta, U\}$ where $R = I(S \leq \min(T, C))$, $S' = \min(T, S, C)$, $\Delta = I(T \leq C)$, $U = \min(T, C)$, T is the primary time to event, S is the time to response to the initial treatment and C is the censoring time. We assume throughout that the censoring time C is independent of all other variables including A_1 and A_2 . The duration of the study is τ .

Denote “ jk ” to be the treatment strategy for which $A_1 = j$ and if there is response then $A_2 = k$, for $j = 1, 2$ and $k = 1, 2$. Note data from this trial provides information on 4 adaptive treatment strategies (strategies “11,” “12,” “21,” “22”). Denote T_{jk} to be the potential time to event if a subject is offered strategy jk . Let $\bar{F}_{jk}(t) = 1 - F_{jk}(t)$, $\Lambda_{jk}(t)$ and $\lambda_{jk}(t)$ be the survival function, cumulative hazard function and the hazard function of T_{jk} , respectively, for $j = 1, 2$ and $k = 1, 2$. The observable time to event T is related to the

potential times to events through relationships described, for example, in [1] and [7], and are omitted here. In particular we assume that if a subject is offered adaptive treatment strategy “ jk ” then $T = T_{jk}$.

We consider two test statistics and associated sample size formulae for comparing two strategies with different first stage treatments since the comparison of two strategies with the same first stage treatment can be reduced to the comparison of the two second-stage treatments only. Suppose we want to power the two-stage randomized trial to detect a difference, if any, between strategies “11” and “22” (the comparisons between strategies “12” and “21”, “11” and “21”, and “12” and “22” are similar). Since only strategies “11” and “22” are considered in the following we simplify the notation and from here on we denote the survival function, cumulative hazard function and hazard function of T_{jj} as $\bar{F}_j(t)$, $\Lambda_j(t)$ and $\lambda_j(t)$, respectively, for $j = 1, 2$.

A variety of statistics have been proposed for data analysis in two-stage randomization trials. All of these statistics involve weighting, which comes in two forms: time independent weights or time dependent weights. Weights are required because for any given adaptive treatment strategy, nonresponding subjects who have a treatment sequence consistent with this strategy are over represented by this design and responding subjects who have a treatment sequence consistent with this strategy are under represented by this design. Specifically, a nonresponding subject who is not randomized in the second stage has a treatment sequence that is consistent with the two strategies beginning with the same first stage treatment as the subject. To adjust for this we use “inverse probability weights” (see e.g., [12], [13]):

$$W_1 = \frac{I(A_1 = 1)}{p} \left[1 - R + \frac{RI(A_2 = 1)}{q} \right] \quad \text{and} \quad W_2 = \frac{I(A_1 = 2)}{1 - p} \left[1 - R + \frac{RI(A_2 = 2)}{1 - q} \right], \quad (1)$$

for strategies “11” and “22”, respectively, where $I(\cdot)$ denotes the indicator function. These weights, which are independent of time, are similar to the weights used by Lunceford et al. [1] and Miyahara and Wahed (the difference is that they consider a fixed first-stage treatment and hence the factors $I(A_1 = 1)/p$ and $I(A_1 = 2)/(1 - p)$ are not necessary).

Alternately one could use time dependent weights:

$$W_1(t) = \frac{I(A_1 = 1)}{p} \left[1 - I(S \leq t) + \frac{I(A_2 = 1)}{q} I(S \leq t) \right], \quad (2)$$

and

$$W_2(t) = \frac{I(A_1 = 2)}{1 - p} \left[1 - I(S \leq t) + \frac{I(A_2 = 2)}{1 - q} I(S \leq t) \right]. \quad (3)$$

These time dependent weights are used in [8] and [10]. Both time independent weights and time dependent weights adjust for the trial design. By using time dependent weights, at

time t , for example, all subjects who receive $A_1 = 1$ as the initial treatment and have not remitted nor censored are considered consistent with strategy “11” and are counted in the (weighted) risk set, while by using constant weights, subjects who receive $A_1 = 1$ and then $A_2 = 2$ are never included in the risk set for strategy “11”. Hence, intuitively the time dependent weights results in efficiency gains compared with the time independent weights. Note that even though (2), (3) are expressed in terms of the unobservable time to response to initial treatment, S , the use of the weights is valid since in the formulae below the weights will only be used in products which are observable (see (4), (5) below).

Test statistic based on the weighted Kaplan-Meier estimator: This test statistic is used for testing $H_0 : \bar{F}_1(t) = \bar{F}_2(t)$ versus $H_1 : \bar{F}_1(t) \neq \bar{F}_2(t)$ for some t satisfying $0 < t \leq \tau$. It is based on the weighted Kaplan-Meier estimator of $\bar{F}_j(t)$ as proposed by Miyahara and Wahed, which is defined as

$$\hat{\bar{F}}_{Kj}(t) = \prod_{u \leq t} \left[1 - \frac{\sum_{i=1}^n W_{ji}(u) dN_i(u)}{\sum_{i=1}^n W_{ji}(u) Y_i(u)} \right], \quad j = 1, 2, \quad (4)$$

where $N(u) = I(U \leq u, \Delta = 1)$ is the counting process for the time to event, and $Y(u) = I(U \geq u)$ is the “at risk” process. (4) uses time dependent weights; however as in Miyahara and Wahed (personal communication) we can use time independent weights (replace $W_{ji}(u)$ by W_{ji} in the above). As will be discussed in Section 3, we will use the test statistic based on weighted Kaplan-Meier estimator with time independent weights for sample size calculation, while the test statistic with time dependent weights is used for data analysis. The asymptotic properties of the weighted Kaplan-Meier estimators are derived in the Appendix. It is worth mentioning that another estimator of $\bar{F}_j(t)$ can be defined as $\exp[-\hat{\Lambda}_j(t)]$, where the estimator of $\Lambda_j(t)$ is the weighted Nelson-Aalen estimator proposed in [8]. Theorem 1 in the Appendix shows that the asymptotic distribution of the weighted Kaplan-Meier estimator is the same as that of this estimator, the latter of which is provided in [8].

By Theorem 1 in the Appendix, the asymptotic distribution of $\sqrt{n}[\hat{\bar{F}}_{Kj}(t) - \bar{F}_j(t)]$ is $N(0, \sigma_{Kj}^2(t))$, where

$$\sigma_{Kj}^2(t) = \bar{F}_j^2(t) E \left\{ \int_0^t \frac{W_j(u)}{\bar{F}_j(u) \bar{F}_C(u)} d[N(u) - Y(u) d\Lambda_j(u)] \right\}^2.$$

Let $\hat{\bar{F}}_C(u)$ be the usual Kaplan-Meier estimator of $\bar{F}_C(u)$. From this result, the test statistic for testing $H_0 : \bar{F}_1(t) = \bar{F}_2(t)$ for some $0 < t \leq \tau$ can be constructed as

$$T_K = \frac{\sqrt{n}[\hat{\bar{F}}_{K1}(t) - \hat{\bar{F}}_{K2}(t)]}{\sqrt{\hat{\sigma}_{K1}^2(t) + \hat{\sigma}_{K2}^2(t)}}.$$

where

$$\hat{\sigma}_{Kj}^2(t) = \frac{\hat{\bar{F}}_{Kj}^2(t)}{n} \sum_{i=1}^n \left\{ \int_0^t \frac{W_{ji}(u)}{\hat{\bar{F}}_{Kj}(u) \hat{\bar{F}}_C(u)} d[N_i(u) - Y_i(u) d\hat{\Lambda}_{Nj}(u)] \right\}^2$$

is a consistent estimator of $\sigma_{Kj}^2(t)$, with $\hat{\Lambda}_{Nj}(t)$ being the estimator of $\Lambda_j(t)$ in [8], for $j = 1, 2$. Note that the variances of the numerators of the test statistics are the sum of variances of the two components since \hat{F}_{K1} and \hat{F}_{K2} are independent. Also note that when time independent weights are used, the $W_j(u)$ and $W_{ji}(u)$ in the above expressions for $\sigma_{Kj}^2(t)$ and $\hat{\sigma}_{Kj}^2(t)$ are replaced by W_j and W_{ji} , respectively. Under the null hypothesis, the test statistic has an asymptotic $N(0, 1)$ distribution. For simplicity, in the following, we call the test based on the weighted Kaplan-Meier estimator with time independent weights cWKM and the test based on the weighted Kaplan-Meier estimator with time dependent weights tWKM.

The weighted log rank statistic: The second test statistic is the weighted version of the log rank test, which is used for testing $H_0 : \bar{F}_1 \equiv \bar{F}_2$ versus $H_0 : \bar{F}_1(t^*) \neq \bar{F}_2(t^*)$ for some $t^* \leq \tau$. The log rank test is the most commonly used test for comparing the distributions of two times to event. It is also commonly used to calculate sample sizes in classical survival analysis [14]. Weighting each subject as above, the following statistic is an analogue of the usual log rank statistic for testing $H_0 : \bar{F}_1 \equiv \bar{F}_2$:

$$G_n = \int_0^\tau \frac{\bar{Y}_{W2}(t)}{\bar{Y}_{W1}(t) + \bar{Y}_{W2}(t)} d\bar{N}_{W1}(t) - \int_0^\tau \frac{\bar{Y}_{W1}(t)}{\bar{Y}_{W1}(t) + \bar{Y}_{W2}(t)} d\bar{N}_{W2}(t), \quad (5)$$

where $\bar{Y}_{Wj}(t) = \sum_{i=1}^n W_{ji}(t)Y_i(t)/n$ and $d\bar{N}_{Wj}(t) = \sum_{i=1}^n W_{ji}(t)dN_i(t)/n$. This test statistic, which was proposed in [8], uses time dependent weights. One can use time independent weights as well, as in the weighted Kaplan-Meier estimator, just by replacing $W_{ji}(t)$ in the definition with W_{ji} . By Theorem 2 in the Appendix, the asymptotic distribution of $\sqrt{n}G_n$ under the null hypothesis is $N(0, \bar{\sigma}^2)$, where $\bar{\sigma}^2 = (\bar{\sigma}_1^2 + \bar{\sigma}_2^2)/4$, and

$$\bar{\sigma}_j^2 = E \left\{ \int_0^\tau W_j(t)[dN(t) - Y(t)d\Lambda_j(t)] \right\}^2.$$

Let

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^\tau W_{ji}(t)[dN_i(t) - Y_i(t)d\hat{\Lambda}_{Nj}(t)] \right\}^2, \quad j = 1, 2.$$

We can use the following test statistic to test $H_0 : \bar{F}_1 \equiv \bar{F}_2$:

$$T_L = \frac{2\sqrt{n}G_n}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}.$$

This test statistic also has an asymptotic $N(0, 1)$ distribution under H_0 . Again, when time independent weights are used, we simply replace $W_j(u)$ and $W_{ji}(u)$ in the above expressions by W_j and W_{ji} , respectively. We call the test using time independent weights cWLR and the test with time dependent weights tWLR hereafter. We will use cWLR for sample size calculation and use tWLR for data analysis. Note that the weighted log rank test here is

different from the test with the same name in classical survival analysis. We used this name because it is consistent with the terminology used for the weighted Kaplan-Meier and with the literature in this area. Also note that [15] proposed a pseudo score test of log hazard ratios in a Cox proportional hazards model for comparing two strategies. In the pseudo score function, each subject is weighted by a time independent weight. The statistic G_n defined above is equivalent to the pseudo score function defined there, but they express the asymptotic variance with a different formula.

3. SAMPLE SIZE CALCULATION

As mentioned in the introduction, we propose to use time independent weights in the test statistics for sample size calculation. This is because the time independent weights make it easier to obtain simple upper bounds on variances involved in the test statistics. These upper bounds are crucial in obtaining relatively simple sample size formulae (see below). On the other hand we suggest that test statistics using time dependent weights be used in the data analyses as these test statistics are potentially more powerful.

First suppose that the cWKM is used for sample size calculation. For definiteness, we suppose that the survival probabilities at the end of the study are to be compared, i.e, $t = \tau$. Under a significance level α , the rejection region of a two-sided test of $H_0 : \bar{F}_1(\tau) = \bar{F}_2(\tau)$ is $\{|T_K| > Z_{1-\frac{\alpha}{2}}\}$. By Theorem 1, the distribution of T_K is approximately normal with mean $\sqrt{n}[\bar{F}_2(\tau) - \bar{F}_1(\tau)]/\sqrt{\sigma_{K1}^2(\tau) + \sigma_{K2}^2(\tau)}$ and variance 1. To achieve a power $1 - \beta$ for the test, we set

$$P\{T_K > Z_{1-\frac{\alpha}{2}} \text{ or } T_K < -Z_{1-\frac{\alpha}{2}}\} = 1 - \beta.$$

This yields

$$P\left\{T_K - \frac{\sqrt{n}[\bar{F}_2(\tau) - \bar{F}_1(\tau)]}{\sqrt{\sigma_{K1}^2(\tau) + \sigma_{K2}^2(\tau)}} > Z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n}|\bar{F}_2(\tau) - \bar{F}_1(\tau)|}{\sqrt{\sigma_{K1}^2(\tau) + \sigma_{K2}^2(\tau)}}\right\} \approx 1 - \beta,$$

and consequently a sample size formula

$$n_K \approx \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2[\sigma_{K1}^2(\tau) + \sigma_{K2}^2(\tau)]}{[\bar{F}_2(\tau) - \bar{F}_1(\tau)]^2}. \quad (6)$$

Next suppose that cWLR is used for sample size calculation. To construct a sample size formula based on the log rank test or its variants, one commonly derives the asymptotic distributions of the test statistics under a proportional hazards assumption and using a local

alternative [10, 15, 16, 17]. This restriction on the alternative hypothesis greatly simplifies the asymptotic means of the test statistics, which facilitates the sample size calculation. We use this approach here as well. Guo [9] studied the asymptotic distribution of the weighted log rank statistic G_n with time dependent weights. As he provided a sketch of the proof of the asymptotic normality of G_n under the local alternative hypothesis, we provide, in the Appendix, a complete proof (also see the Appendix for the exact definition of the local alternative hypothesis), when either time independent or time dependent weights are used. Now based on the asymptotic distribution of G_n given in Theorem 2 in the Appendix, using a similar derivation as above, the corresponding sample size formula using cWLR can be obtained as

$$n_L \approx \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2(\bar{\sigma}_1^2 + \bar{\sigma}_2^2)}{\xi^2 \left[\int_0^\tau \bar{F}_C(t) dF_1(t) \right]^2}, \quad (7)$$

where ξ is the log hazard ratio.

From the above formulae, we need working assumptions on $\sigma_{K_1}^2(\tau)$ and $\sigma_{K_2}^2(\tau)$ or $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$ to calculate the sample sizes. A challenge is that even with the time independent weights these variances depend in a complex manner on the joint distribution of primary time to event and time to response to the initial treatment (note that time to response to the initial treatment is involved in the weight function; furthermore the weight is not predictable thus one cannot simplify the variance formulae by the usual martingale arguments). We will see that the use of potential outcomes along with the use of upper bounds on the variances simplifies the working assumptions. By (6) and Theorem 1 in the Appendix, to use the cWKM for sample size calculation, we need upper bounds for

$$E \left\{ \int_0^\tau \frac{W_j}{\bar{F}_j(t)\bar{F}_C(t)} d[N(t) - Y(t)d\Lambda_j(t)] \right\}^2, \quad j = 1, 2.$$

Consider $j = 1$. To get the upper bound, we use potential outcomes under the two strategies. Moreover, the randomization of second stage treatments ensures that given $R = 1$, A_2 is independent of the potential outcomes (the T_{jk} s) and times to response to the initial treatment (the S_j s). Denote $N_1(u) = I(T_{11} \leq u, T_{11} \leq C)$ and $Y_1(u) = I(T_{11} > u, C > u)$. We obtain the upper bounds by replacing the response indicator R with 1. Noting that W_1 in (1) is nonzero if and only if $N(u) \equiv N_1(u)$ and $Y(u) \equiv Y_1(u)$, the upper bound can be obtained as follows

$$\begin{aligned} & E \left\{ \int_0^\tau \frac{W_1}{\bar{F}_1(t)\bar{F}_C(t)} d[N(t) - Y(t)d\Lambda_1(t)] \right\}^2, \\ &= EE(W_1^2 | T_{11}, C) \left[\int_0^\tau \frac{1}{\bar{F}_1(t)\bar{F}_C(t)} d[N_1(t) - Y_1(t)d\Lambda_1(t)] \right]^2 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{p} EE \left(1 - R + \frac{R}{q} \middle| T_{11}, C \right) \left\{ \int_0^\tau \frac{1}{\bar{F}_1(t)\bar{F}_C(t)} d[N_1(t) - Y_1(t)d\Lambda_1(t)] \right\}^2 \\
&\leq \frac{1}{pq} E \left\{ \int_0^\tau \frac{1}{\bar{F}_1(t)\bar{F}_C(t)} d[N_1(t) - Y_1(t)d\Lambda_1(t)] \right\}^2 \\
&= \frac{1}{pq} \int_0^\tau \frac{d\Lambda_1(t)}{\bar{F}_1(t)\bar{F}_C(t)}. \tag{8}
\end{aligned}$$

The case for $j = 2$ is analogous. For the weighted log rank test, we need upper bounds for $E[\int_0^\tau W_j \{dN(u) - Y(u) d\Lambda_1(u)\}]^2$, $j = 1, 2$. The following inequality gives the upper bound when $j = 1$ and the case for $j = 2$ is similar:

$$\begin{aligned}
&E \left[\int_0^\tau W_1 \{dN(t) - Y(t)d\Lambda_1(t)\} \right]^2 \\
&= EE(W_1^2 | T_{11}, C) \left[\int_0^\tau \{dN_1(t) - Y_1(t)d\Lambda_1(t)\} \right]^2 \\
&= \frac{1}{p} E \left(1 - R + \frac{R}{q} \middle| T_{11}, C \right) \left[\int_0^\tau \{dN_1(t) - Y_1(t)d\Lambda_1(t)\} \right]^2 \\
&\leq \frac{1}{pq} \int_0^\tau \bar{F}_C(t) dF_1(t). \tag{9}
\end{aligned}$$

Since the upper bounds are obtained by replacing the response indicator R with 1, the more subjects randomized at the second stage (e.g. the larger the number of responders), the sharper the upper bounds. In the case that every subject is randomized at the second stage, for example, if all subjects are responders, the upper bounds are precise. Here we choose to use the time independent weights instead of time dependent weights in the weighted Kaplan-Meier estimator and the weighted log rank statistic for sample size calculation, because from the above derivation we see that only using the time independent weights enables us to derive these upper bounds relatively easily by resorting to the martingale theory.

Now replacing the variances in the above sample size formulae by their upper bounds, the sample size calculated by using cWKM and upper bounds of variances is

$$n_K \leq \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \sigma_B^2}{[\bar{F}_2(\tau) - \bar{F}_1(\tau)]^2}, \tag{10}$$

where

$$\sigma_B^2 = \frac{\bar{F}_1^2(\tau)}{pq} \int_0^\tau \frac{d\Lambda_1(t)}{\bar{F}_1(t)\bar{F}_C(t)} + \frac{\bar{F}_2^2(\tau)}{(1-p)(1-q)} \int_0^\tau \frac{d\Lambda_2(t)}{\bar{F}_2(t)\bar{F}_C(t)}.$$

In a similar manner, the sample size calculated from cWLR using upper bounds of the variances is

$$n_L \leq \left[\frac{1}{pq} + \frac{1}{(1-p)(1-q)} \right] \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\xi^2 \int_0^\tau \bar{F}_C(t) dF_1(t)}. \tag{11}$$

We notice that, in order to calculate the sample size using (11), we only need information about the hazard ratio and the integral appearing in it, which is exactly the same information that is needed when one uses the standard log rank test to size a two-arm study in classical survival analysis [14]. Moreover, a nice property of the log rank test is that the integral, $\int_0^\tau \bar{F}_C(t) dF_1(t)$, is equal to the probability of observing an event before time τ when all subjects are offered strategy “11.” This may be easier to elicit from clinicians than the entire distribution function which is needed to use formula (10). In Table 1, we list our two tests to determine sample sizes, the asymptotic variances, as well as the upper bounds of the variances.

“Table 1 about here.”

4. SIMULATION

We conduct a simulation study to assess the performance of the two sample size formulae. In this simulation, the sample sizes are calculated using cWKM or cWLR and upper bounds of variances. For data analyses, we use more powerful tests than those used for sample size formulae derivation. As mentioned above, time dependent weights result in efficiency gain. Therefore, for testing the survival probabilities at a fixed time point, a potentially more powerful test is the one based on the tWKM of the $\bar{F}_j(t)$ s, and for the weighted log rank test, the tWLR is potentially more powerful than the cWLR. Moreover, for testing the survival probabilities at a fixed time point, there are two more tests that are potentially more powerful than the cWKM. One test is based on a slight generalization of the third weighted sample proportion estimator of $\bar{F}_j(t)$ in [1]. In this generalization, we use time dependent weights as defined above instead of time independent weights as in the original estimator. This estimator is the most efficient among the three weighted sample proportion estimators proposed there (see [1] for details), and it is of interest to know if this estimator (or the estimator generalized by using time dependent weights, as we will use) is more efficient than the cWKM. The other test is based on the estimator of $\bar{F}_j(t)$ proposed by Wahed and Tsiatis [7]. They first derived the semiparametric efficient estimator of $\bar{F}_j(t)$, and then, recognizing the difficulty in its implementation, they proposed a less efficient but easier to implement estimator. This proposed estimator is potentially more efficient than all other available estimators of $\bar{F}_j(t)$. However, there is no definite comparison of the relative efficiencies of the modified third estimator in Lunceford et al. [1] (abbreviated “Lunceford 3” hereafter), the tWKM, and the Wahed and Tsiatis estimator (abbreviated “WT estimator” hereafter). Therefore the tests based on these three estimators are all considered and compared in our simulation study.

In the simulation, the duration of the study is supposed to be 16 weeks, i.e., $\tau = 16$. We assume the times to event under strategy “11” and strategy “22” have proportional hazards

in the generative models. We specify the distributions of time to event under strategies “11” and “22” as Weibull distributions, with scale parameters 20 for T_{11} and the scale parameter for T_{22} is determined by the hazard ratio, and a common location parameter 2 to ensure proportional hazards. The hazard ratio takes values 1.25, 1.5 or 2.0. The first and second stage randomization probabilities are $p = q = 0.5$. Denote the time to response when the first stage treatment is $A_1 = j$ as S_j , for $j = 1, 2$. We generate (T_{jj}, S_j) jointly from a Frank copula model [18] with association parameter -5 for $j = 1$ and -6 for $j = 2$, resulting in a negative correlation between time to event and time to response to the first-stage treatment. The censoring time has a point mass at τ and otherwise is uniformly distributed over $(0, \tau)$, and is independent of all other variables. The amount of point mass at τ varies so that about 20% or 40% censorings occur. The times to response, S_1 and S_2 have the same type of marginal distributions as the failure times T_{11} and T_{22} , respectively, but with different parameters, and these parameters take different values such that different percentages of subjects are randomized in the second stage. Specifically, the percentage of subjects randomized at the second stage ranges from about 25% to about 75%. The survival probabilities at the end of study under the two strategies range in $(0.4, 0.6)$ among all scenarios. In one simulation, we calculate the sample sizes using (10), the sample size formulae based on cWKM, and then generate data sets using the calculated sample sizes. For each simulated data set, we calculate the achieved powers of tests based on tWKM, Lunceford 3 and the WT estimator. In another simulation, we calculate the sample sizes using (11), the formula based on the cWLR, and then calculate the simulated power of both cWLR and tWLR under the calculated sample sizes. In these simulations we assume that we know the exact distribution functions (for using (10)) or the hazard ratio and the probability of observing an event before τ (for using (11)) for calculating the sample sizes. In all simulation settings, the sample sizes are calculated based on a significance level 0.05 and a desired power 0.8.

For comparison, in the simulation we also calculate the sample sizes using the method proposed by Feng and Wahed [11]. They considered sample size calculation for two-stage designs by using the second weighted sample proportion estimator proposed in [1] to test the equality of survival probabilities at a fixed time point. In simplifying variance calculation, they made the working assumption that $E(R_j|T_{jj})$ is a constant, where R_j is the response indicator when the first-stage treatment is $A_1 = j$, for $j = 1, 2$, which means that the response status is independent of the potential time to event. They also assumed that the potential times to event follow exponential distributions to make computations easier, although this is not necessary for their method (hence in our simulation we did not assume this in calculating the sample sizes using their method). Then they calculate the sample sizes by using exact variances involved in the test statistics under these working assumptions.

Table 2 lists the powers of three tests of $H_0 : \bar{F}_1(\tau) = \bar{F}_2(\tau)$ based on tWKM, Lunceford 3 and the WT estimator, under sample sizes calculated by using formula (10). Table 3 includes the achieved powers of the cWLR and tWLR under sample sizes calculated using

(11). From all these results, we observe that, under the sample sizes calculated from cWKM, the desired power is achieved if we choose the appropriate tests. Specifically, both tWKM and the test based on WT estimator guarantee the power, while the test based on Lunceford 3 may not. Actually, from results not shown here, we observe that cWKM, as the method for data analysis, also always guarantees the desired power and is conservative, but tWKM is a little more efficient than cWKM. Table 2 also includes sample sizes calculated from the Feng and Wahed method. These sample sizes are a little larger than those calculated from cWKM (and upper bounds of variances), hence the power can also be achieved under these sample sizes. We compared sample sizes calculated from the two methods for a variety of other cases (results not presented here), and found that the sample sizes calculated using formula (10) are also a little smaller than the corresponding ones calculated from Feng and Wahed method. Possible explanations for this include the relative low efficiency of the weighted sample proportion estimator compared with cWKM, as well as the working assumptions that are used in Feng and Wahed’s sample size formula — the response status is independent of the potential failure time. Note that in our simulation settings, this working assumption is not satisfied. We also calculated sample sizes using Feng and Wahed’s method under scenarios in which this working assumption is actually satisfied. Contrary to the above findings, we found that in these cases, the sample sizes calculated using Feng and Wahed’s method are a little smaller than those calculated from (10). And under these sample sizes, the powers of the tests based on tWKM and WT are either over 0.8 or less than but close to 0.8 (in cases where under the sample sizes calculated from (10) the achieved powers are close to 0.8). Under the same sample sizes, the powers of the test based on Lunceford 3 are usually under 0.8.

From Tables 2 and 3, we also observe that, using upper bounds of variances to determine sample sizes results in conservative sample sizes. However, the degree of conservatism depends on the percentage of subjects randomized at the second stage. The higher percentage of subjects randomized at the second stage, the less conservative the sample sizes. Notably, by results in Tables 2 and 3, when the percentage approaches or exceeds 70%, the achieved powers are becoming pretty close to the desired power. Note that the results in Table 2 indicate that the WT estimator is more efficient than all the other estimators, especially when the sample sizes are large. In smaller samples, the efficiency of the WT estimator is reduced due to the variability introduced by estimation of some population quantities, which appear as expectations of some random variables (see [7] for details). Finally, Table 3 shows that tWLR is more efficient than cWLR. Based on these findings, we recommend that the WT (or if a simpler estimator is desired, the tWKM) estimator is used for data analysis to compare the equality of survival probabilities under two strategies at one time point, and the tWLR is used for data analysis to test the equality of the two survival curves.

“Table 2 about here.”

“Table 3 about here.”

Table 4 presents the achieved powers when the quantities needed to calculate the sample sizes are misspecified. We consider the cases where the hazard ratio is 1.25 or 1.5. In the specification of the marginal distributions of T_{jj} for calculating the sample size using (10), we assume that they still follow Weibull distributions but with wrong scale or shape parameters, or they follow exponential distributions with the same survival probabilities at the end of study. The results show that, although misspecification of the marginal distributions of T_{11} and T_{22} or the probability of observing an event before τ may have a large effect on the achieved powers (or sample sizes), the desired powers for both tests are still usually achieved or approximated under reasonable degree of misspecification, which is a consequence of the conservatism of the sample sizes under no misspecification.

“Table 4 about here.”

5. SAMPLE SIZE FORMULAE FOR MORE GENERAL TWO-STAGE DESIGNS

In the simple two-stage design considered in previous sections, only responders to the first-stage treatments are further randomized to one of two second-stage treatments, and the options for the second randomization do not depend on the first-stage treatment. In practice, many variations of the two-stage design exist, which also call for corresponding sample size formulae. In this section we show that these methods for sample size calculation can be generalized to the more general designs as follows. Let A_1 be the coding variable for the options for the first-stage treatment, which takes values $1, 2, \dots, k_1$. Responders to $A_1 = j$ are further randomized to one of the second-stage treatments $A_{2j}^R = 1, 2, \dots$, or k_{2j}^R , and nonresponders to $A_1 = j$ are further randomized to one of the second-stage treatments $A_{2j}^N = 1, 2, \dots$, or k_{2j}^N , for $j = 1, 2, \dots, k_1$. The simple design considered above corresponds to the special case in which $k_1 = 2$, $k_{21}^N = k_{22}^N = 2$ and $k_{21}^R = k_{22}^R = 1$.

The method for deriving sample size formulae in Section 3 also applies to the more general two-stage designs. At first, we can also define test statistics based on the weighted Kaplan-Meier estimator and the weighted log rank test statistic for comparing two treatment strategies in the general designs, just with weights modified (see the following for examples). Further, it is easy to see that in the general designs, the asymptotic variance formulae in Theorems 1 and 2 remain unchanged, except that the formulae for the weight function W_j may change, which results in different upper bounds of the variances and hence different sample size formulae. In the following we illustrate this with two examples.

As the first example, we consider a design in which only responders to the initial treatment $A_1 = 1$ are further randomized to one of two maintenance treatments, while all other subjects are given fixed second-stage treatments. In this design, there are three possible strategies: strategy 1 — offer treatment $A_1 = 1$ first, and if there is response then offer maintenance treatment $A_2 = 1$; strategy 2 — offer treatment $A_1 = 1$ first, and if there is response then offer maintenance treatment $A_2 = 2$; strategy 3 — offer treatment $A_1 = 2$ first, and then offer a (fixed) treatment afterwards, which may or may not depend on the response status. We call the three strategies “11”, “12” and “2”, respectively. Suppose the randomization probabilities are $P(A_1 = 1) = p$ and $P(A_2 = 1|R = 1) = q$. If we use time independent weights, the weight function associated with strategy “11” is the same as the weight function W_1 in (1), while the weight function associated with strategy “2” is $W_2 = I(A_1 = 2)/(1 - p)$. At first, suppose we use the cWKM to test the equivalence of strategies “11” and “2” to size the study. Similarly as in Section 3, we obtain upper bounds for $E \left\{ \int_0^\tau W_1 / [\bar{F}_1(u)\bar{F}_C(u)] d[N(u) - Y(u)d\Lambda_1(u)] \right\}^2$ and $E \left\{ \int_0^\tau W_2 / [\bar{F}_2(u)\bar{F}_C(u)] d[N(u) - Y(u)d\Lambda_2(u)] \right\}^2$, where $\bar{F}_1(u)$, $\bar{F}_2(u)$, $\Lambda_1(t)$ and $\Lambda_2(t)$ are the survival functions and cumulative hazard functions of the potential times to event under strategies “11” and “2”, respectively. The upper bound when W_1 is involved is the same as (9) since the form of W_1 is unchanged. However, when W_2 is involved, we actually do not need an upper bound since the quantity can be calculated precisely as

$$E \left\{ \int_0^\tau \frac{W_2}{\bar{F}_2(u)\bar{F}_C(u)} d[N(u) - Y(u)d\Lambda_2(u)] \right\}^2 = \frac{1}{1-p} \int_0^\tau \frac{d\Lambda_2(u)}{\bar{F}_2(u)\bar{F}_C(u)},$$

because there is no response indicator involved. Consequently, the sample size using a test based on cWKM and upper bounds of variances is

$$n \leq \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \sigma_B^2}{[\bar{F}_2(\tau) - \bar{F}_1(\tau)]^2},$$

where

$$\sigma_B^2 = \frac{\bar{F}_1^2(\tau)}{pq} \int_0^\tau \frac{d\Lambda_{11}(u)}{\bar{F}_1(u)\bar{F}_C(u)} + \frac{\bar{F}_2^2(\tau)}{1-p} \int_0^\tau \frac{d\Lambda_2(u)}{\bar{F}_2(u)\bar{F}_C(u)}.$$

Similarly, the sample size derived from cWLR and upper bounds of variances is

$$n \leq \left(\frac{1}{pq} + \frac{1}{1-p} \right) \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\xi^2 \int_0^\tau \bar{F}_C(t) dF_1(t)}.$$

For a second example, consider a design in which both responders and nonresponders are further randomized. At first, all subjects are randomized to treatment $A_1 = j$ with probability p_{1j} , $j = 1, 2, \dots, k_1$. Those who respond to $A_1 = j$ are further randomized to $A_{2j}^R = i$ with probability p_{2j}^{Ri} , and those who do not respond to $A_1 = j$ are further randomized

to $A_{2j}^N = i$ with probability p_{2j}^{Ni} . Consider the following two strategies: strategy “111” — first offer $A_1 = 1$, and if there is response then offer $A_{21}^R = 1$ but if there is no response, then offer $A_{21}^N = 1$, and strategy “222” — offer $A_1 = 2$ first, and responders to $A_1 = 2$ are then offered $A_{22}^R = 2$ and nonresponders to $A_1 = 2$ are offered $A_{22}^N = 2$. Using time independent weights, the weight function for strategy “111” is

$$W_1 = \frac{I(A_1 = 1)}{p_{11}} \left[\frac{RI(A_{21}^R = 1)}{p_{21}^{R1}} + \frac{(1 - R)I(A_{21}^N = 1)}{p_{21}^{N1}} \right],$$

and the weight function for strategy “222” is

$$W_2 = \frac{I(A_1 = 2)}{p_{12}} \left[\frac{RI(A_{22}^R = 2)}{p_{22}^{R2}} + \frac{(1 - R)I(A_{22}^N = 2)}{p_{22}^{N2}} \right].$$

If we use cWKM to calculate the sample size, we need upper bounds of

$$E \left\{ \int_0^\tau \frac{W_j}{\bar{F}_j(u)\bar{F}_C(u)} d[N(u) - Y(u)d\Lambda_j(u)] \right\}^2, \quad j = 1, 2,$$

where $\bar{F}_j(t)$ and $\Lambda_j(t)$ are the survival function and cumulative hazard function of T_j — the potential time to event under strategy “ jjj ”, $j = 1, 2$. Denote the counting process and the “at risk” process of T_j by $N_j(t)$ and $Y_j(t)$, respectively, for $j = 1, 2$. By repeated expectations, we have, when $j = 1$,

$$\begin{aligned} & E \left\{ \int_0^\tau \frac{W_1}{\bar{F}_1(u)\bar{F}_C(u)} d[N(u) - Y(u)d\Lambda_1(u)] \right\}^2, \\ &= EW_1^2 \left[\int_0^\tau \frac{1}{\bar{F}_1(u)\bar{F}_C(u)} d[N_1(u) - Y_1(u)d\Lambda_1(u)] \right]^2 \\ &= \frac{1}{p_{11}} EE \left[\frac{1}{p_{21}^{N1}} + \left(\frac{1}{p_{21}^{R1}} - \frac{1}{p_{21}^{N1}} \right) R \middle| T_1, C \right] \left[\int_0^\tau \frac{d[N_1(u) - Y_1(u)d\Lambda_1(u)]}{\bar{F}_1(u)\bar{F}_C(u)} \right]^2 \\ &\quad \begin{cases} \leq \frac{1}{p_{11}p_{21}^{R1}} E \left[\int_0^\tau \frac{d[N_1(u) - Y_1(u)d\Lambda_1(u)]}{\bar{F}_1(u)\bar{F}_C(u)} \right]^2, & \text{if } p_{21}^{R1} < p_{21}^{N1}, \\ \leq \frac{1}{p_{11}p_{21}^{N1}} E \left[\int_0^\tau \frac{d[N_1(u) - Y_1(u)d\Lambda_1(u)]}{\bar{F}_1(u)\bar{F}_C(u)} \right]^2, & \text{if } p_{21}^{R1} > p_{21}^{N1}, \\ = \frac{1}{p_{11}p_{21}^{N1}} E \left[\int_0^\tau \frac{d[N_{111}(u) - Y_1(u)d\Lambda_1(u)]}{\bar{F}_1(u)\bar{F}_C(u)} \right]^2, & \text{if } p_{21}^{R1} = p_{21}^{N1}, \end{cases} \\ &= \frac{1}{p_{11} \min(p_{21}^{R1}, p_{21}^{N1})} \int_0^\tau \frac{1}{\bar{F}_1(u)\bar{F}_C(u)} d\Lambda_1(u). \end{aligned}$$

Similarly, for $j = 2$, we have

$$\begin{aligned} & E \left\{ \int_0^\tau \frac{W_2}{\bar{F}_2(u)\bar{F}_C(u)} d[N(u) - Y(u)d\Lambda_1(u)] \right\}^2 \\ &\leq \frac{1}{p_{12} \min(p_{22}^{R2}, p_{22}^{N2})} \int_0^\tau \frac{1}{\bar{F}_2(u)\bar{F}_C(u)} d\Lambda_2(u). \end{aligned}$$

It follows that the sample size for comparing strategies “111” and “222”, calculated from cWKM and upper bounds of variances is

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \sigma_B^2}{[\bar{F}_2(\tau) - \bar{F}_1(\tau)]^2},$$

where

$$\sigma_B^2 = \frac{\bar{F}_1^2(\tau)}{p_{11} \min(p_{21}^{R1}, p_{21}^{N1})} \int_0^\tau \frac{d\Lambda_1(u)}{\bar{F}_1(u)\bar{F}_C(u)} + \frac{\bar{F}_2^2(\tau)}{p_{12} \min(p_{22}^{R2}, p_{22}^{N2})} \int_0^\tau \frac{d\Lambda_2(u)}{\bar{F}_2(u)\bar{F}_C(u)}.$$

Similarly, the sample size based on cWLR and upper bounds of variances is

$$n = \left[\frac{1}{p_{11} \min(p_{21}^{R1}, p_{21}^{N1})} + \frac{1}{p_{12} \min(p_{22}^{R2}, p_{22}^{N2})} \right] \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\xi^2 \int_0^\tau \bar{F}_C(u) dF_1(u)}.$$

Remark. If the randomization probabilities are all equal to each other in the second-stage randomization, then the upper bounds are exact and there is no conservatism in the sample sizes. In this case, the weights are actually unnecessary. This happens when the numbers of options for all second-stage randomizations are the same and all subjects are randomized to the options evenly. Of course, this cannot be true when some subjects are not randomized, for instance, all responders are not rerandomized, implying that the number of options for these subjects is 1.

Finally, sample size formulae for general multi-stage designs when there are more than two stages can be obtained in a similar way as for the designs discussed above. Again, the difference is only in the weight functions and the resulting upper bounds of variances, but the idea for obtaining upper bounds and the form of the upper bounds and sample size formulae is similar.

6. DISCUSSION

For two-stage randomized designs in which the primary outcome is a time to event, we propose using upper bounds of variances involved in test statistics to calculate the conservative sample sizes. These sample size formula ensure a given power for testing if two prespecified adaptive treatment strategies differ either in terms of the probability of survival at a given time point or in terms of the hazard functions. The degree of conservatism depends on the response rate to the first-stage treatment; if for example the proportion of subjects rerandomized at the second stage reaches about 70%, the sample size is exhibits little to no longer conservatism.

The simulations results indicate that the sample sizes obtained using the weighted log rank test are usually considerably smaller than the sample sizes obtained from comparing survival probabilities at a certain time point (see Tables 2 and 3). Moreover, one needs less information in order to calculate the sample size derived from the weighted log rank test. However the log rank is designed to have highest power against proportional hazards [19], thus if the hazards under the alternate hypothesis are nonproportional it maybe that the sample size formula does not provide the desired power. The tests for equality of survival probabilities at one time point have more modest goals (do not attempt to test for differences in survival probabilities at other time points), hence may be better able to ensure power for this limited goal.

As mentioned before, there are several existing sample size formulae for similar settings in the literature. Besides [11], [10] also considered sample size formula for two-stage randomizations with censored data, in which they use the supremum weighted log rank test to calculate the sample size. However, different working assumptions are posited in different methods. For example, [11] assumes that the probability of response does not depend on the potential time to event, and [10] makes more difficult to achieve working assumptions on the data generating model. The working assumptions of the three methods considered in the simulation study are listed in Table 5. The use of the upper bounds enables us to make the same kinds of working assumptions needed for sizing one-stage trials. Moreover, the sample sizes are robust to the incorrectness of the working assumptions.

“Table 5 about here.”

REFERENCES

1. Lunceford, J.K., Davidian M. and Tsiatis, A.A., Estimation of survival distributions of treatment strategies in two-stage randomization designs in clinical trials. *Biometrics* 2002; 58:48-57.
2. Murphy S.A., An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 2005; 24:1455-1481.
3. Lieberman J.A., Stroup T.S., McEvoy J.P., Swartz M.S., Rosenheck R.A., Perkins D.O., Keefe R.S., Davis S.M., Davis, C.E., Lebowitz B.D., Severe J., Hsiao J.K., Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Investigators. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine* 2005; 53(12):1209-1223.

4. Tummarello, D., Mari, D., Graziano, F., Isidori, P., Cetto, G., Pasini, F., Santo, A., and Cellerino, R. A randomized, controlled phase III study of cyclophosphamide, doxorubicin, and vincristine with Etoposide (CAV-E) or Teniposide (CAV-T), followed by recombinant interferon- α maintenance therapy or observation, in small cell lung carcinoma patients with complete responses. *Cancer* 1997; 80:2222-2229.
5. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications* 1986; 14:1393-1512.
6. Lavori PW, Dawson R, Roth AJ. Flexible treatment strategies in chronic disease: clinical and research implications. *Biological Psychiatry* 2000; 48:605-614.
7. Wahed, A.S. and Tsiatis, A.A., Semiparametric efficient estimation of survival distribution in two-stage randomization designs in clinical trials with censored data. *Biometrika* 2006; 93:163-177.
8. Guo X. and Tsiatis A.A., A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics* 2005; 1(1):1-15.
9. Guo X., Statistical analysis in two stage randomization designs in clinical trials. unpublished PhD thesis, Department of Statistics, North Carolina State University, 2005.
10. Feng, W and Wahed, A.S, A supremum log rank test for comparing adaptive treatment strategies and corresponding sample size formula. *Biometrika* 2008; 95, 3, 695-707.
11. Feng, W. and Wahed, S.A., Sample size for two-stage studies with maintenance therapy. *Statistics in Medicine*, 2009; 28:2028-2041.
12. Robins J.M., Rotnitzky A. and Zhao L.P., Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; 89:846-866.
13. Murphy SA, van der Laan MJ, Robins JM, CPPRG. Marginal mean models for dynamic regimes. *Journal of American Statistical Association* 2001; 96:1410-1423.
14. Schonefeld, D.A., The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981; 68:316-319.
15. Lokhnygina Y. and Helterbrand J.D., Cox regression methods for two-stage randomization designs. *Biometrics* 2007; 63:422-428.
16. Chow S.C., Shao J. and Wang H. *Sample Size Calculations in Clinical Research*. Chapman and Hall, 2005.

17. Eng, K.H. and Kosorok, M.R, A sample size formula for the supremum log rank statistic. *Biometrics* 2005; 61:86-91.
18. Roger, B.N. *An Introduction to Copulas*. Springer-Verlag, 1998.
19. Fleming, T.R. and Harrington D.P., *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
20. van der Vaart, A.W., and Wellner, J. A., *Weak Convergence and Empirical Processes*, Springer-Verlag, New York, 1996.

APPENDIX: ASYMPTOTIC RESULTS AND PROOFS

In this appendix, we provide asymptotic properties and their proofs for estimators and test statistics that we used, which are not proved in the literature. These include the asymptotic distributions of the weighted Kaplan-Meier estimators, with time independent weights or time dependent weights, the third weighted sample proportion estimator in [1] (adapted by using time dependent weights), and the weighted log rank statistic under the local alternative hypothesis, with time independent weights or time dependent weights. The results are stated in Theorems 1 and 2, and the proofs follow the theorems.

Theorem 1. Assume that $\bar{F}_1(t) > \delta_0$ and $\bar{F}_C(t) > \delta_0$ for some $\delta_0 > 0$. Then

$$\sqrt{n}\{\hat{F}_{Kj}(t) - \bar{F}_j(t)\} \rightarrow_d N(0, \sigma_{Kj}^2), \quad j = 1, 2,$$

as $n \rightarrow \infty$, where

$$\sigma_{Kj}^2(t) = \bar{F}_j^2(t) E \left\{ \int_0^t \frac{W_j}{\bar{F}_j(u)\bar{F}_C(u)} d[N(u) - Y(u)d\Lambda_j(u)] \right\}^2, \quad (12)$$

when time independent weights are used and

$$\sigma_{Kj}^2(t) = \bar{F}_j^2(t) E \left\{ \int_0^t \frac{W_j(u)}{\bar{F}_j(u)\bar{F}_C(u)} d[N(u) - Y(u)d\Lambda_j(u)] \right\}^2, \quad (13)$$

when time dependent weights are used. Moreover, if we denote as $\hat{F}_{Sj}(t)$ the third weighted sample proportion estimator in [1] (adapted by using time dependent weights), then

$$\sqrt{n}\{\hat{F}_{Sj}(t) - \bar{F}_j(t)\} \rightarrow_d N(0, \sigma_{Sj}^2), \quad j = 1, 2,$$

where

$$\begin{aligned} \sigma_{Sj}^2(t) &= E\{W_j(T)I(T \leq t) - F_{jj}(t) - \alpha_j(W_j(T) - 1)\}^2 \\ &\quad + \int_0^t \frac{E\{L_j(t, u)\}^2}{\bar{F}_C(u)} \lambda^c(u) du, \end{aligned} \quad (14)$$

with

$$L_j(t, u) = [W_j(T)I(T \leq t) - G_j(t, u) - \alpha_j\{W_j(T) - 1 - G_{W_j}(u)\}]I(T \geq u),$$

$\lambda^c(t)$ being the hazard function of the censoring time C , and $G_{W_j}(u) = E\{(W_j(U) - 1)I(T \geq u)\}/P(T \geq u)$, $j = 1, 2$. \square

The asymptotic properties of the weighted log rank test statistic are those under the proportional hazards assumption and the local alternative hypothesis. The particular local alternative hypothesis we use is represented by

$$H_n : \lambda_2^n(t) = \lambda_1(t) \exp(\gamma/\sqrt{n}), \quad n \geq 1, \quad (15)$$

where γ is a constant. Under the local alternative hypothesis, we denote the distribution of the observed data under H_n as P_n , and denote the distribution under the null hypothesis as P_0 . Theorem 2 below gives the asymptotic distribution of the weighted log rank test statistic.

Theorem 2: Assume that $\bar{F}_1(\tau) > \delta_0$ and $\bar{F}_C(\tau) > \delta_0$ for some $\delta_0 > 0$. Then

$$\sqrt{n}G_n \rightarrow_d N(\mu_L, (\bar{\sigma}_1^2 + \bar{\sigma}_2^2)/4)$$

under P_n , as $n \rightarrow \infty$, where $\mu_L = \gamma \int_0^\tau \bar{F}_C(t) dF_1(t)/2$ and

$$\bar{\sigma}_j^2 = P_0 \left\{ \int_0^\tau W_j [dN(t) - Y(t)d\Lambda_1(t)] \right\}^2, \quad j = 1, 2,$$

when time independent weights are used, and

$$\bar{\sigma}_j^2 = P_0 \left\{ \int_0^\tau W_j(t) [dN(t) - Y(t)d\Lambda_1(t)] \right\}^2, \quad j = 1, 2,$$

when time dependent weights are used.

Proof of Theorem 1:

A. Weighted Kaplan-Meier Estimator

We only prove the result for $\hat{F}_{K1}(t)$ with time dependent weights. The proof for $\hat{F}_{K2}(t)$ and proofs when time independent weights are used are parallel.

Let X be the observed data. Let P be the probability measure of X . Denote $\mathbb{P}_n f(X) = \sum_{i=1}^n f(X_i)/n$, $Pf = \int f dP$, and $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$ for any function f of X . Let $dN_{W_j}(u) = W_j(u)dN(u)$ and $Y_{W_j}(u) = W_j(u)Y(u)$, $j = 1, 2$. At first, by Proposition 2 in [8], for any function $\theta(u)$ on the real line,

$$P \int_0^t \frac{\theta(u)}{\bar{F}_1(u)} [dN_{W_1}(u) - Y_{W_1}(u)d\Lambda_1(u)] = 0.$$

It follows from this equality, along with a similar argument as in the proof of Theorem 3.2.3 in [19], that

$$\begin{aligned} & \sqrt{n}[\hat{F}_{K1}(t) - \bar{F}_1(t)] \\ &= -\bar{F}_1(t)\mathbb{G}_n \int_0^t \frac{\hat{F}_{K1}(u-)}{\bar{F}_1(u)} \frac{1}{\bar{Y}_{W1}(u)} [dN_{W1}(u) - Y_{W1}(u)d\Lambda_1(u)]. \end{aligned} \quad (16)$$

We first show that the estimator $\hat{F}_{K1}(u)$ is uniformly consistent in $[0, t]$, for any $0 < t \leq \tau$. In order to do this, we need to show that some classes of functions are Donsker [20]. Define the classes of functions

$$\begin{aligned} \Phi &= \{\phi(u) : \phi(u) \text{ is a monotone function from } [0, t] \text{ to } [\delta_0, 1]\}, \\ \Theta &= \left\{ \theta(u) = \frac{\phi_1(u)}{\phi_2(u)} : \phi_1(u) \in \phi, \phi_2(u) \in \Phi \right\}, \end{aligned}$$

and

$$\mathcal{F} = \left\{ f_\theta(X) = \int_0^t \frac{\theta(u)}{\bar{F}_1(u)} [dN_{W1}(u) - Y_{W1}(u)d\Lambda_1(u)] : \theta(u) \in \Theta \right\}.$$

In the following, denote C to be a generic constant. For any real function f define on $[0, t]$, denote $\|f\|_1^2 = \int_0^t f^2(u)dF_1(u)$. Let $\phi_1^L(u), \phi_1^U(u), \dots, \phi_K^L(u), \phi_K^U(u)$ be the set of ε -brackets covering ϕ , where $K = \exp(-C/\varepsilon)$. For any function $\theta(u)$ in Θ , there exist functions $\phi_1(u)$ and $\phi_2(u)$ in Φ such that $\theta(u) = \phi_1(u)/\phi_2(u)$. Let $\phi_{i1}^L(u), \phi_{i1}^U(u)$ and $\phi_{i2}^L(u), \phi_{i2}^U(u)$ be ε -brackets for $\phi_1(u)$ and $\phi_2(u)$, respectively. Then $\phi_{i1}^L(u)/\phi_{i2}^U(u) \leq \theta(u) \leq \phi_{i2}^U(u)/\phi_{i1}^L(u)$ and

$$\left\| \frac{\phi_{i1}^L(u)}{\phi_{i2}^U(u)} - \frac{\phi_{i1}^U(u)}{\phi_{i2}^L(u)} \right\|_1^2 \leq C\{\|\phi_{i1}^L(u) - \phi_{i1}^U(u)\|_1^2 + \|\phi_{i2}^L(u) - \phi_{i2}^U(u)\|_1^2\}.$$

This implies that the ε -entropy of Θ is of the same order as that of Φ , which, by Theorem 2.7.5 in [21], is C/ε . Therefore, there exist functions $\theta_j^L(u) \in \Theta, \theta_j^U(u) \in \Theta, 1 \leq j \leq K$ such that $\|\theta_j^L - \theta_j^U\|_1 \leq \varepsilon, 1 \leq j \leq K$, and for any $\theta(u) \in \Theta$, $\theta_j^L(u) \leq \theta(u) \leq \theta_j^U(u)$ for some $1 \leq j \leq K$. Consequently, the function $f_\theta(X)$ in \mathcal{F} satisfies

$$f_\theta(X) \geq \int_0^t \frac{\theta_j^L(u)}{\bar{F}_1(u)} dN_{W1}(u) - \int_0^t \frac{\theta_j^U(u)}{\bar{F}_1(u)} Y_{W1}(u) d\Lambda_1(u) \equiv f_j^L(X)$$

and

$$f_\theta(X) \leq \int_0^t \frac{\theta_j^U(u)}{\bar{F}_1(u)} dN_{W1}(u) - \int_0^t \frac{\theta_j^L(u)}{\bar{F}_1(u)} Y_{W1}(u) d\Lambda_1(u) \equiv f_j^U(X),$$

and moreover, if we define $\|f\|_2^2 = Pf^2$, then

$$\|f_j^L - f_j^U\|_2 = \left\| \int_0^t \frac{\theta_j^L(u) - \theta_j^U(u)}{\bar{F}_1(u)} dN_{W1}(u) - \int_0^t \frac{\theta_j^U(u) - \theta_j^L(u)}{\bar{F}_1(u)} Y_{W1}(u) d\Lambda_1(u) \right\|_2$$

$$\begin{aligned}
&\leq \left\| \int_0^t \frac{\theta_i^L(u) - \theta_i^U(u)}{\bar{F}_1(u)} dN_{W_1}(u) \right\|_2 + \left\| \int_0^t \frac{\theta_i^U(u) - \theta_i^L(u)}{\bar{F}_1(u)} Y_{W_1}(u) d\Lambda_1(u) \right\|_2 \\
&\leq C \left\{ \left[\int_0^t (\theta_i^L(u) - \theta_i^U(u))^2 dF_1(u) \right]^{\frac{1}{2}} + \left[\int_0^t (\theta_i^U(u) - \theta_i^L(u))^2 dF_1(u) \right]^{\frac{1}{2}} \right\} \\
&\leq C\varepsilon.
\end{aligned}$$

It follows that the functions $f_1^l, f_1^u, \dots, f_K^l, f_K^u$ are $C\varepsilon$ brackets and they cover \mathcal{F} . Hence the bracket entropy of \mathcal{F} is also of order C/ε , and therefore \mathcal{F} is a P -Donsker class. It follows that, by the continuous mapping theorem,

$$\begin{aligned}
&\left| \mathbb{G}_n \int_0^t \frac{\hat{F}_{K_1}(u-)}{\bar{F}_1(u)} \frac{1}{\bar{Y}_{W_1}(u)} [dN_{W_1}(u) - Y_{W_1}(u)d\Lambda_1(u)] \right| \\
&\leq \sup_{\theta \in \Theta} \left| \mathbb{G}_n \int_0^t \frac{\theta(u)}{\bar{F}_1(u)} [dN_{W_1}(u) - Y_{W_1}(u)d\Lambda_1(u)] \right| \\
&\rightarrow \sup_{\theta \in \Theta} |\mathbb{G}f_\theta|, \quad \text{as } n \rightarrow \infty.
\end{aligned}$$

In light of (16), this implies that $\hat{F}_{K_1}(t) \rightarrow \bar{F}_1(t)$, with rate $1/\sqrt{n}$. Since both $\hat{F}_{K_1}(\cdot)$ and $\bar{F}_1(\cdot)$ are increasing and bounded functions, and $\bar{F}_1(\cdot)$ is continuous, it follows that

$$\sup_{u \in [0, t]} |\hat{F}_{K_1}(u) - \bar{F}_1(u)| \rightarrow 0, \quad \text{and} \quad \sup_{u \in [0, t]} |\hat{F}_{K_1}(u-) - \bar{F}_1(u)| \rightarrow 0. \quad (17)$$

This holds for every $t \in [0, \tau]$.

To show the asymptotic normality of $\hat{F}_{K_1}(t)$, we write

$$\begin{aligned}
\hat{F}_{K_1}(t) - \bar{F}_1(t) &= -\bar{F}_1(t) \left\{ (\mathbb{P}_n - P) \int_0^t \frac{1}{\bar{Y}_{W_1}(u)} [dN_{W_1}(u) - Y_{W_1}(u)d\Lambda_1(u)] \right. \\
&\quad \left. + (\mathbb{P}_n - P)D_n(X) \right\}, \quad (18)
\end{aligned}$$

where $y_{W_1}(u) = E[W_1 Y(u)] = \bar{F}_1(u)\bar{F}_C(u)$, and

$$D_n(X) = \int_0^t \left\{ \frac{1}{\bar{Y}_{W_1}(u)} - \frac{\hat{F}_{K_1}(u-)}{\bar{F}_1(u)} \frac{1}{\bar{Y}_{W_1}(u)} \right\} [dN_{W_1}(u) - Y_{W_1}(u)d\Lambda_1(u)].$$

Let $\theta_n(u) = \hat{F}_{K_1}(u-)/\bar{Y}_{W_1}(u)$ and $\theta_0 = \bar{F}_1(u)/y_{W_1}(u)$. Then by (17) and the law of large numbers, $\|\theta_n - \theta_0\|_\infty \equiv \sup_{0 \leq u \leq \tau} |\theta_n(u) - \theta_0(u)| \rightarrow 0$, as $n \rightarrow \infty$. Since the class of functions \mathcal{F} is Donsker, by equicontinuity (see [20]), for large n we have, for some sequence $\delta_n \rightarrow 0$,

$$\begin{aligned}
|\sqrt{n}(\mathbb{P}_n - P)D_n(X)| &= |\mathbb{G}_n[f_{\theta_n}(X) - f_{\theta_0}(X)]| \\
&\leq \sup_{\|\theta - \theta_0\|_\infty \leq \delta_n} |\mathbb{G}_n[f_\theta(X) - f_{\theta_0}(X)]| \\
&\leq \sup_{\|f_\theta - f_{\theta_0}\|_2 \leq C\delta_n} |\mathbb{G}_n[f_\theta(X) - f_{\theta_0}(X)]| \\
&\rightarrow_p 0.
\end{aligned}$$

Now the asymptotic normality of $\hat{F}_{K1}(t)$ follows from this and (18), with an asymptotic variance that is stated in the Theorem.

B. The Third Weighted Sample Proportion Estimator with Time Dependent Weights

We only prove the result for $\hat{F}_{S1}(t)$.

By the consistency of $\hat{\alpha}_1$ and $\hat{F}_C(u)$, we can write

$$\begin{aligned}\hat{F}_{S1}(t) - \bar{F}(t) &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\Delta_i W_{1i}(U_i)}{\bar{F}_C(T_i)} I(T_i > t) - \bar{F}_1(t) \right\} \\ &\quad - \alpha_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i}{\bar{F}_C(U_i)} (W_{1i}(U_i) - 1) \right\} + o_P(1).\end{aligned}$$

The first term and the second term on the right hand side of the above equality both converge to 0 in probability by the law of large numbers. So we conclude that $\sup_{u \in [0, t]} |\hat{F}_{S1}(u) - \bar{F}_1(u)| \rightarrow_P 0$, as $n \rightarrow \infty$.

We now prove the asymptotic normality. At first, we can write

$$\begin{aligned}\sqrt{n}\{\hat{F}_{S1}(t) - \bar{F}_1(t)\} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\Delta_i W_{1i}(U_i)}{\hat{F}_C(T_i)} I(T_i > t) - \bar{F}_1(t) \right\} \\ &\quad - \alpha_1 \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\Delta_i}{\hat{F}_C(U_i)} [W_{1i}(U_i) - 1] \right\} + o_P(1).\end{aligned}\quad (19)$$

Denote

$$\begin{aligned}\hat{G}_1(t, u) &= \frac{1}{n\hat{F}(u-)} \sum_{i=1}^n \frac{\Delta_i W_{1i}(U_i) I(T_i \leq t) I(T_i \geq u)}{\hat{F}_C(T_i)}, \\ G_1^*(t, u) &= \frac{1}{\hat{F}(u-)} \left\{ \bar{F}_1(t) - \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i W_{1i}(U_i) I(T_i < u)}{\hat{F}_C(T_i)} \right\},\end{aligned}$$

and $G_1(t, u) = P(u \leq T_{11} \leq t) / \bar{F}(u)$.

Following Zhao and Tsiatis (1997), the first term on the right hand side of (19) can be written as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \{W_{1i}(U_i) I(T_i \leq t) - \bar{F}_1(t)\}$$

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t \frac{dM_i^c(u)}{\bar{F}_C(u)} \{W_{1i}(U_i)I(T_i \leq t) - G_1(t, u)\} + D_n, \quad (20)$$

where

$$\begin{aligned} D_n &= -\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t \frac{dM_i^c(u)}{\bar{F}_C(u)} \{G_1(t, u) - G_1^*(t, u)\} \\ &\quad + \frac{1}{\sqrt{n}} \{\hat{F}_1(t) - F_1(t)\} \sum_{i=1}^n \int_0^t \frac{dM_i^c(u)}{\bar{F}_C(u) \hat{F}(u-)}. \end{aligned} \quad (21)$$

Note that $M_i^c(u)$ is the martingale for the counting process of the censoring time C . Now we can prove that $D_n = o_P(1)$ by showing both terms in D_n is $o_{P_n}(1)$. To show that the first term is of order $o_{P_n}(1)$, we use the following inequality, which is a corollary of Lengart's inequality [19]:

$$P \left\{ \sup_{t \leq T} \left[\int_0^t H(s) dM(s) \right]^2 \geq \varepsilon \right\} \leq \frac{\eta}{\varepsilon} + P \left\{ \int_0^T H^2(s) d\langle M, M \rangle(s) \geq \eta \right\},$$

for any $\varepsilon > 0$ and $\eta > 0$, where M is the martingale corresponding to a counting process, $H(s)$ is a predictable and locally bounded process and T is a stopping time such that $P(T < \infty) = 1$. By this inequality, for the first term in D_n , we have, for any $\varepsilon > 0$ and $\eta > 0$,

$$\begin{aligned} &P \left\{ \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t \frac{dM_i^c(u)}{\bar{F}_C(u)} \{G_1(t, u) - G_1^*(t, u)\} \right]^2 \geq \varepsilon \right\} \\ &\leq \frac{\varepsilon}{\eta} + P \left\{ \int_0^t \frac{[G_1(t, u) - G_1^*(t, u)]^2}{\bar{F}_C^2(u)} \frac{1}{n} \sum_{i=1}^n Y_i^c(u) \lambda^c(u) du \geq \eta \right\} \\ &\leq \frac{\varepsilon}{\eta} + P \left\{ \int_0^t \frac{[G_1(t, u) - G_1^*(t, u)]^2}{\bar{F}_C^2(u)} \lambda^c(u) du \geq \eta \right\}. \end{aligned}$$

By the above inequality, it suffices to verify that $\sup_{0 \leq u \leq \tau} |G_1(t, u) - G_1^*(t, u)| \rightarrow_P 0$. By the uniform consistency of $\hat{F}(u-)$ and $\hat{F}_C(u)$, this claim reduces to

$$\sup_{0 \leq u \leq t} \left| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i W_{1i}(U_i) I(T_i < u)}{\bar{F}_C(T_i)} - P(T_{11} < u) \right| \rightarrow_P 0,$$

which follows from law of large numbers. On the other hand, we can prove that $1/\sqrt{n} \sum_{i=1}^n \int_0^t dM_i^c(u)/[\bar{F}_C(u) \hat{F}(u-)]$ converges in distribution. First note that, by defining $\mathcal{F} = \left\{ \int_0^t dM_i^c(u) / \phi(u) : \phi \in \phi \right\}$, we can prove that \mathcal{F} is a Donsker class, the details of which are omitted. Then by asymptotic equi-continuity of the process $\sqrt{n}(\mathbb{G}_n - P)$, $1/\sqrt{n} \sum_{i=1}^n \int_0^\tau dM_i^c(u)/[\bar{F}_C(u) \hat{F}(u-)]$

differs from $1/\sqrt{n} \sum_{i=1}^n \int_0^t dM_i^c(u)/[\bar{F}_C(u)\bar{F}(u-)]$ by a quantity of order $o_P(1)$. Now the claim that $D_n \rightarrow 0$ follows from this and the central limit theorem.

In an analogous manner, we can show that for the second term (in the braces) on the right hand side of (19) is equal to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ W_{1i}(U_i) - 1 - \int_0^\tau \frac{W_{1i}(u) - 1 - G_{W1}(u)}{\bar{F}_C(u)} dM_i^c(u) \right\} + o_P(1), \quad (22)$$

where $G_{W1}(u) = P\{(W_1(U) - 1)I(T \geq u)\}/P(T \geq u)$. Now the theorem follows from (19), (20), (21) and (22) and the central limit theorem. The asymptotic variance of the $\hat{F}_{S1}^\tau(t)$ is the variance of the major terms, which is equal to σ_{S1}^2 in (14). \square

Proof of Theorem 2:

In the following we also assume that the weights are time dependent. The proof is similar for time independent weights.

Define

$$\mathcal{F} = \left\{ \int_0^\tau \theta(u)[dN_{W2}(u) - Y_{W2}(u)d\Lambda_1(u)] : \theta(u) \in \Theta \right\},$$

where Θ is defined in the proof of Theorem 1. We first prove that $\sqrt{n}(\mathbb{P}_n - P_n)$ converges to \mathbb{G}_{P_0} in $\ell^\infty(\mathcal{F})$ under P_n , as $n \rightarrow \infty$. The proof of the asymptotic normality of G_n rely on the asymptotic equi-continuity of the process $\sqrt{n}(\mathbb{P}_n - P_n)$ implied by this result. We use Theorem 2.8.9 in [20] for the proof. In order to use that theorem, we need to verify the following two conditions:

$$\sup_{f,g \in \mathcal{F}} |\rho_{P_n}(f,g) - \rho_{P_0}(f,g)| \rightarrow 0, \quad (23)$$

where $\rho_P(f,g) = \text{var}_P(f-g)$; and there exists an envelope function F of \mathcal{F} such that

$$\limsup_{n \rightarrow \infty} P_n F^2 I(F \geq \varepsilon\sqrt{n}) = 0, \quad (24)$$

and $P_n F^2 = O(1)$. Moreover, we need to ascertain that both $\mathcal{F}_{\delta, P_n} = \{f-g : f, g \in \mathcal{F}, \|f-g\|_{P_n,2} < \delta\}$ and $\mathcal{F}_\infty = \{(f-g)^2 : f, g \in \mathcal{F}, \|f-g\|_{P_n,2} < \delta\}$ are P_n -measurable for every $\delta > 0$ and n .

We first check condition (23). Let

$$f = \int_0^\tau \theta_1(u)[dN_{W2}(u) - Y_{W2}(u)d\Lambda_1(u)],$$

and

$$g = \int_0^\tau \theta_2(u)[dN_{W_2}(u) - Y_{W_2}(u)d\Lambda_1(u)],$$

for some functions $\theta_1(u), \theta_2(u) \in \Theta$. Denote $f_P(t_{22}, s)$ and $f_P(t_{22})$ to be the probability density functions of (T_{22}, S) and T_{22} under probability measure P of the observed data, and denote $f_C(c)$ to be the probability density function of the censoring time C . Denoting $Y_{22}(u) = I(T_{22} > u, C > u)$, we can write

$$\begin{aligned} \rho_P(f, g) &= P \left\{ [\theta_1(T_{22}) - \theta_2(T_{22})] I(T_{22} \leq C) I(T_{22} \leq \tau) W_2(T_{22}) \right. \\ &\quad \left. - \int_0^\tau [\theta_1(u) - \theta_2(u)] W_2(u) Y_{22}(u) d\Lambda_1(u) \right\}^2 \\ &= \int_0^\infty \int_0^\infty \int_0^\infty K(t_{22}, s, c) f_P(t_{22}, s) f_C(c) dt_{22} ds dc, \end{aligned} \quad (25)$$

where the function $K(t_{22}, s, c)$ can be bounded by a constant A that does not depend on n , by our assumptions on Θ . Under H_n , we have $\Lambda_2^{(n)}(u) = \Lambda_1(u) \exp(c/\sqrt{n})$, which implies that

$$f_{P_n}(t_{22}) - f_{P_0}(t_{22}) = f_{P_0}(t_{22}) \left\{ \exp\left(\frac{\gamma}{\sqrt{n}}\right) [\bar{F}_1(t_{22})]^{\exp(\frac{\gamma}{\sqrt{n}})-1} - 1 \right\}.$$

This, combined with the fact that $f_{P_n}(s|t_{22}) = f_{P_0}(s|t_{22})$, yields

$$f_{P_n}(t_{22}, s) - f_{P_0}(t_{22}, s) = f_{P_0}(t_{22}, s) \left\{ \exp\left(\frac{\gamma}{\sqrt{n}}\right) [\bar{F}_1(t_{22})]^{\exp(\frac{\gamma}{\sqrt{n}})-1} - 1 \right\}. \quad (26)$$

Now by (25) and (26), it follows that, for any $f, g \in \mathcal{F}$,

$$\begin{aligned} &|\rho_{P_n}(f, g) - \rho_{P_0}(f, g)| \\ &\leq \int_0^\infty \int_0^\infty \int_0^\infty K(t_{22}, s, c) |f_{P_n}(t_{22}, s) - f_{P_0}(t_{22}, s)| f_C(c) dt_{22} ds dc \\ &\leq A \int_0^\infty \int_0^\infty \int_0^\infty f_C(c) f_{P_0}(t_{22}, s) \left| \exp\left(\frac{\gamma}{\sqrt{n}}\right) [\bar{F}_1(t_{22})]^{\exp(\frac{\gamma}{\sqrt{n}})-1} - 1 \right| dt_{22} ds dc. \end{aligned} \quad (27)$$

Since the absolute value in the integrand converges to 0 as $n \rightarrow \infty$, the right hand side of (27) converges to 0 as $n \rightarrow \infty$. This follows from the dominated convergence theorem as follows. When $\gamma > 0$, the absolute value in the integrand is bounded by $\exp(\gamma) + 1$. When $\gamma < 0$, it is bounded by $[\bar{F}_1(t_{22})]^{\exp(\gamma)-1} + 1$. Plugging the absolute value by this bound, the integral is bounded by $P_0[\bar{F}_1(T_{11})]^\alpha + 1$, where $\alpha = \exp(\gamma) - 1 < 0$. Since $\bar{F}_1(T_{11})$ is uniformly distributed in $[0, 1]$ and $\alpha > -1$, $P_0[\bar{F}_1(T_{11})]^\alpha < \infty$.

To check condition (24), note that the functions $\int_0^\tau \theta(u)[dN_{W_2}(u) - Y_{W_2}(u)d\Lambda_1(u)]$ are bounded by a constant under our assumptions. So we can choose the envelope function F

to be the upper bound. For such an envelope function, condition (24) is obviously satisfied. And we also have that $P_n F^2 = O(1)$.

Finally, the P_n -measurability of $\mathcal{F}_{\delta, P_n}$ and \mathcal{F}_{∞}^2 follows since θ is a monotone function divided by another monotone function. For any monotone function, it is the (pointwise) limit of a series of step functions of the form $\sum_{i=1}^n c_i I(t_{i-1} < t \leq t_i)$, where all the t_i s are rational numbers. Since the set of all such functions is countable, by Example 2.3.4 in van der Vaart and Wellner (1996), both $\mathcal{F}_{\delta, P_n}$ and \mathcal{F}_{∞}^2 are P_n -measurable.

Now we conclude from Theorem 2.8.9 in [20] that $\sqrt{n}(\mathbb{P}_n - P_n)$ converges to \mathbb{G}_{P_0} in $\ell^\infty(\mathcal{F})$ under P_n , as $n \rightarrow \infty$.

Similarly, if we define

$$\mathcal{F}' = \left\{ \int_0^t \theta(u) [dN_{W_1}(u) - Y_{W_1}(u) d\Lambda_1(u)] : \theta(u) \in \Theta \right\},$$

then we can also show that $\sqrt{n}(\mathbb{P}_n - P_n)$ converges to \mathbb{G}_{P_0} in $\ell^\infty(\mathcal{F}')$ under P_n .

Before we can show the asymptotic normality of G_n , we need to show that

$$\sup_{t \in [0, \tau]} \left| \frac{\bar{Y}_{W_j}(t)}{\bar{Y}_{W_1}(t) + \bar{Y}_{W_2}(t)} - \frac{1}{2} \right| \rightarrow_{P_n} 0, j = 1, 2, \text{ as } n \rightarrow \infty.$$

This follows from the fact that $\bar{Y}_{W_j}(t) - y_{W_j}(t) \rightarrow_{P_n} 0$, $j = 1, 2$, which is a consequence of the asymptotic normality of $\sqrt{n}[\bar{Y}_{W_j}(t) - y_{W_j}(t)]$ under P_n . The latter can be proved by the Lindeberg-Feller central limit theorem, the details of which are omitted here. From these results, it follows that, under P_n ,

$$\begin{aligned} \sqrt{n}G_n &= \sqrt{n}(\mathbb{P}_n - P_n) \int_0^\tau \frac{\bar{Y}_{W_2}(t)}{\bar{Y}_{W_1}(t) + \bar{Y}_{W_2}(t)} [dN_{W_1}(t) - Y_{W_1}(t) d\Lambda_1(t)] \\ &\quad - \sqrt{n}(\mathbb{P}_n - P_n) \int_0^\tau \frac{\bar{Y}_{W_1}(t)}{\bar{Y}_{W_1}(t) + \bar{Y}_{W_2}(t)} [dN_{W_2}(t) - Y_{W_2}(t) d\Lambda_2(t)] \\ &\quad + \sqrt{n} \int_0^\tau \frac{\bar{Y}_{W_1}(t) \bar{Y}_{W_2}(t)}{\bar{Y}_{W_1}(t) + \bar{Y}_{W_2}(t)} [\lambda_1(t) - \lambda_2(t)] dt \\ &= \frac{\sqrt{n}}{2} (\mathbb{P}_n - P_n) \int_0^\tau [dN_{W_1}(t) - Y_{W_1}(t) d\Lambda_1(t)] \\ &\quad - \frac{\sqrt{n}}{2} (\mathbb{P}_n - P_n) \int_0^\tau [dN_{W_2}(t) - Y_{W_2}(t) d\Lambda_1(t)] \\ &\quad + \frac{\gamma}{2} \int_0^\tau \bar{F}_1(t) \bar{F}_C(t) d\Lambda_1(t) + o_{P_n}(1). \end{aligned}$$

Again by the Lindeberg-Feller central limit theorem, the first term on the right hand side of the above equality converges in distribution to $N(0, \bar{\sigma}_1^2/4)$, and the second term converges to $N(0, \bar{\sigma}_2^2/4)$, both under P_n , where

$$\bar{\sigma}_j^2 = P_0 \left(\int_0^\tau W_j [dN(t) - Y(t)d\Lambda_1(t)] \right)^2, \quad j = 1, 2.$$

Finally, $G_n/\sqrt{n} \rightarrow_d N(\mu, (\bar{\sigma}_1^2 + \bar{\sigma}_2^2)/4)$ under P_n , where $\mu = \gamma \int_0^\tau \bar{F}_C(t)d\Lambda_1(t)/2 = \gamma \int_0^\tau \bar{F}_C(t)dF_1(t)/2$. \square

Table 1: Variances and their upper bounds involved in the test statistics based on cWKM or cWLR

Method	weighted Kaplan-Meier	weighted log rank
variance of test statistic	$\sum_{j=1}^2 \bar{F}_j^2(\tau) E \left\{ \int_0^\tau \frac{W_j d[N(u) - Y(u)d\Lambda_j(u)]}{\bar{F}_j(u)\bar{F}_C(u)} \right\}^2$	$\frac{1}{4} \sum_{j=1}^2 P_0^* \left\{ \int_0^\tau W_j d[N(u) - Y(u)d\Lambda_1(u)] \right\}^2$
upper bound of variance	$\left\{ \frac{1}{pq} + \frac{1}{(1-p)(1-q)} \right\} \sum_{j=1}^2 \bar{F}_j^2(\tau) \int_0^\tau \frac{d\Lambda_j(u)}{\bar{F}_j(u)\bar{F}_C(u)}$	$\frac{1}{4} \left\{ \frac{1}{pq} + \frac{1}{(1-p)(1-q)} \right\} \int_0^\tau \bar{F}_C(u) dF_1(u)$

P_0 : expectation under the assumption that $F_1(t) \equiv F_2(t)$

Table 2: Achieved powers of different tests under sample sizes obtained by using (10). (T_{11}, S_1) and (T_{22}, S_2) both follow Frank copula models with association parameters -5 and -6, respectively. The marginal distribution of T_{11} is Weibull with scale parameter 20 and shape parameter 2 and the marginal distribution of T_{22} is also Weibull, with the same shape parameter as T_{11} and the scale parameter is determined by the hazard ratio. The significance level of the test is 0.05 and the desired power is 0.8.

Hazard ratio	$n_K(n^*)$	% rerandomized	tWKM	Lunceford 3	WT
1.25	2824(2938)	25	0.92	0.87	0.92
		50	0.87	0.84	0.90
		75	0.84	0.77	0.89
1.5	805(894)	25	0.91	0.83	0.92
		50	0.85	0.76	0.91
		75	0.83	0.75	0.89
2	251(314)	25	0.92	0.76	0.89
		50	0.85	0.68	0.86
		75	0.80	0.67	0.85

n^* : sample size calculated from the Feng and Wahed method.

Table 3: Achieved powers of the weighted log rank tests under sample sizes obtained by (11). The true model is the same as that described in Table 2. The significance level of the test is 0.05 and the desired power is 0.8.

Hazard ratio	n_L	25% rerandomized		50% rerandomized		75% rerandomized	
		cWLR	tWLR	cWLR	tWLR	cWLR	tWLR
1.25	2651	0.94	0.95	0.86	0.87	0.81	0.83
1.5	753	0.95	0.96	0.87	0.88	0.81	0.83
2	234	0.91	0.92	0.85	0.86	0.82	0.83

Table 4: Achieved powers of tests under misspecifications of the true model. In the true model, (T_{11}, S_1) and (T_{22}, S_2) both follow Frank copula models with association parameters -1 and -2, respectively. The marginal distribution of T_{11} is Weibull with scale parameter 20 and shape parameter 2 and the marginal distribution of T_{22} is also Weibull, with the same shape parameter as T_{11} and the scale parameter is determined by the hazard ratio. The first column indicates the misspecification of the true model. The significance level of the test is 0.05 and the desired power is 0.8.

cWKM			cWLR		
Specification of the model	sample size	Power [‡]	Specification of true quantities	sample size	Power [‡]
True model 1: scaleT11=20, shapeT11=shapeT12=2, HR=1.25 P(observe an event before time τ)=0.39					
no misspecification	3345	0.89	no misspecification	3210	0.93
scaleT11=17	2738	0.83	P(event)=0.45 ^b	2214	0.75
scaleT11=23	4098	0.95	P(event)=0.50	2522	0.81
shapeT11=shapeT22=1.75	4154	0.96	P(event)=0.60	2101	0.70
shapeT11=shapeT22=2.25	2784	0.85	P(event)=0.30	4203	0.97
exponential*	2825	0.81	P(event)=0.25	5044	0.97
True model 2: scaleT11=20, shapeT11=shapeT12=2, HR=1.5 P(observe an event before time τ)=0.37					
no misspecification	1072	0.87	no misspecification	1035	0.90
scaleT11=17	866	0.86	P(event)=0.45	849	0.82
scaleT11=23	1325	0.94	P(event)=0.50	764	0.77
shapeT11=shapeT22=1.75	1319	0.95	P(event)=0.60	637	0.71
shapeT11=shapeT22=2.25	901	0.83	P(event)=0.30	1273	0.95
exponential*	806	0.80	P(event)=0.25	1528	0.97

[‡]: All powers are under the same tests as used to calculate the sample size.

^b: P(event)=P(observe an event before τ).

*: The marginal distributions of T_{11} and T_{22} are supposed to be exponential distributions with the same survival probabilities with the true model at τ .

Table 5: Methods for sample size calculation and working assumptions needed to calculate the sample size

Method	Working assumptions
cWKM	knowledge of $\bar{F}_j(t)$ and $\bar{F}_C(t)$ for all t
Feng & Wahed	the response status is independent of the potential time to event, knowledge of $\bar{F}_j(t)$ and $\bar{F}_C(t)$ for all t
cWLR	proportional hazards, knowledge of hazard ratio and $P(\text{observe an event before } \tau)$