Sample Size Formulae for Two-Stage Randomized Trials with Survival Outcomes

Zhiguo Li
Department of Biostatistics and Bioinformatics
Duke University Medical Center, Durham, North Carolina 27710, U.S.A.
and Susan A. Murphy
Department of Statistics, University of Michigan
Ann Arbor, Michigan 48109, U.S.A.

Abstract

Two-stage randomized trials are growing in importance in developing adaptive treatment strategies, i.e., treatment policies or dynamic treatment regimes. Usually the first stage involves randomization to one of several initial treatments. The second stage of treatment begins when an early nonresponse criterion or response criterion is met. In the second stage nonresponding subjects are rerandomized among second-stage treatments. Sample size calculations for planning these two-stage randomized trials with failure time outcomes are challenging because the variances of common test statistics depend in a complex manner on the joint distribution of time to the early nonresponse criterion or response criterion and the primary failure time outcome. We produce simple, albeit conservative, sample size formulae by using upper bounds on the variances. The resulting formulae only require the same working assumptions needed to size a standard single stage randomized trial and, in common settings, are only mildly conservative. These sample size formulae are based on either a weighted Kaplan–Meier estimator of survival probabilities at a fixed time point or a weighted version of the log rank test.

KEY WORDS: Dynamic treatment regime; Sample size calculation; Sequential multiple assignment randomized trial; Weighted Kaplan–Meier estimator; Weighted log rank test.

1 Introduction

Adaptive treatment strategies, also called treatment policies or dynamic treatment regimes (Robins, 1986; Lavori et al., 2000; Murphy et al., 2001), are growing in importance in the management of chronic disorders and in the treatment of disorders for which there are no universally effective treatments. In both cases initial treatments may only be acutely effective for about 40-60% of the patients; thus multiple stages of treatment are frequently required to obtain good outcomes. Formally, an adaptive treatment strategy is a sequence of decision rules one per stage of the treatment. Each decision rule inputs patient information and outputs a recommended treatment. Examples of adaptive treatment strategies abound. See McKay et al. (2004), Murphy et al.(2007), and Marlowe et al. (2008), for an example in the treatment of alcohol dependence, in continuing care for drug abuse, and in criminology, respectively. The second author is currently collaborating with scientists on improving the following simple adaptive treatment strategy for attention deficit hyperactivity disorder in children: provide behavioral modification therapy initially, and beginning at month 2 and every month thereafter, assess the child's classroom behavior; if the behavior problems in the classroom exceed a prespecified criterion, then augment the behavioral therapy with methylphenidate.

With the increase in the use of adaptive treatment strategies there has been an increase in calls for clinical trial designs for use in developing these treatment strategies. Sequential, multiple assignment, randomized trials (Lavori & Dawson, 2003; Murphy, 2005; Murphy et al., 2007) have been proposed. In these trials each subject can proceed through multiple stages of treatment and may be randomized at each stage among treatments. Thus each subject may be randomized multiple times through the course of the trial. Precursors of this design have been used in a variety of medical fields (Stone et al., 1995; Tummarello et al., 1997; Rush et al., 2004; Lieberman et al., 2005). In this paper we focus on a particularly common version of the sequential multiple assignment randomized design, a two-stage randomized trial: in the first stage subjects are randomized to one of two initial treatments. The second stage of treatment begins when and if early signs of nonresponse occur. In the second stage non-responding subjects are rerandomized among two subsequent treatments. Responding subjects stay on the initial treatment or are assigned another fixed second-stage treatment. Variants of this design are discussed in the Supplementary Material.

Because sequential multiple assignment randomized trials are intended to provide data that can be used to assist in the development of an adaptive treatment strategy, they tend to be sized for comparing two subgroups involved in the trial (Murphy, 2005; Murphy et al., 2007; Oetting et al., 2007). In the two-stage randomized trial, the two subgroups might be the two treatments for the nonresponders or alternatively may be two of the adaptive treatment strategies implemented in the trial (Murphy et al. 2007). Here we consider sizing the study to compare two adaptive treatment strategies beginning with different initial treatments. In practice these two adaptive treatment strategies might be the most intensive and the least intensive strategy or might represent opposing clinical approaches. In general the sequential multiple assignment randomized design should be followed by a more standard randomized clinical trial, in which the developed strategy is compared to an appropriate alternative. See the above references for discussions of these issues.

We focus on sizing a two-stage randomized trial for a failure time outcome, for example, time until a first school disciplinary event in the case of the following trial for attention deficit hyperactivity disorder or time until treatment dropout in the case of some schizophrenia and substance abuse trials. This paper is motivated by our experience in designing several sequential multiple assignment randomized studies, one of which is the following 36 week trial for attention deficit hyperactivity disorder. The primary purpose of this trial is to assist clinical scientists in constructing an adaptive treatment strategy composed of behavioral and/or medication components. In this study children with this disorder are first randomized to either a low intensity behavioral modification therapy or a low dose of medication. Beginning at 2 months and every month thereafter each child's classroom behavior is assessed and compared to a prespecified criterion. Exceeding the criterion is interpreted as an early sign of nonresponse; the nonresponding children are then rerandomized to either intensification of current treatment or a combined treatment. Children who do not show signs of nonresponse continue on their first-stage treatment. The trial design is provided in Figure 1. Although this trial was originally sized using the outcome of end of school year child behavior problems, one of the interesting outcomes was the time until a first school disciplinary event.

As is clear from the study in attention deficit hyperactivity disorder there are two time-to-event outcomes, that is, the time to early nonresponse and the time to the primary outcome. To improve clarity we use the term, failure time, to refer to the time to the primary outcome. The failure time can occur before or after the time to early nonresponse, e.g., the time until a first school disciplinary event can occur before or after the child's classroom behavior meets the criterion for early nonresponse.

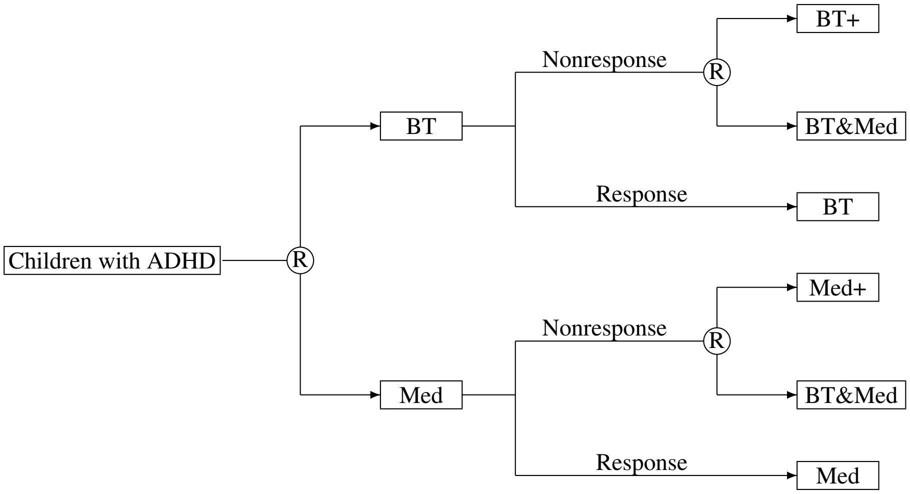

Fig. 1. Design of the trial in attention deficit hyperactivity disorder, abbreviated ADHD. R: randomization; BT: behavioral therapy; Med: medication; BT+: intensified behavioral therapy; Med+: higher dose medication; BT&Med: combined behavioral therapy and medication.

We develop relatively simple, easy to use, sample size formulae for comparing two adaptive treatment strategies that begin with different initial treatments. We provide sample size formulae based on two different tests: a test of the equality of survival probabilities at one time point using a weighted Kaplan–Meier estimator and a test of the equality of hazard functions based on a weighted version of the log rank test. The challenge to developing easy to use sample size formulae is that the variances involved in the test statistics are complex functionals of the joint distribution of time to early nonresponse and failure time; these two times are likely to be dependent. For example, in the study in attention deficit hyperactivity disorder the time to child behavior exceeding the non-response criterion and the time to a first school disciplinary event are likely dependent.

To achieve the goal of easy to use sample size formulae, we simplify the formulae by replacing the variance terms in them by appropriate upper bounds. The resulting sample size formulae require similar information to that needed to size a two group randomized trial for a failure time outcome; in particular we need not make assumptions concerning the dependence between the time to early nonresponse and the failure time. To construct the upper bounds we first express the variances in terms of potential outcomes. Second we use time independent weights instead of time dependent weights to construct the test statistics underlying the sample size formulae. The resulting sample size formulae will be conservative both due to the use of upper bounds on the variances and due to the fact that tests which are potentially more efficient than those used to derive the sample size formulae are used in data analysis. See the next sections for details about this. However as will be seen, in common settings in which about 40–60% of subjects experience early signs of nonresponse and are thus rerandomized, these sample sizes are only minimally conservative.

In addition to providing simple, easy to use, sample size formulae, we provide the asymptotic theory for the weighted Kaplan–Meier estimator. Moreover, we provide here-to-fore omitted theory justifying the use of the weighted version of log rank test. Details are provided in the Supplementary Material.

Sample size formulae for sizing two-stage randomized trials for failure time outcomes have been proposed and studied in Feng & Wahed (2008, 2009). Feng & Wahed (2009) developed a sample size formula based on a weighted sample proportion estimator of survival function whereas Feng & Wahed (2008) developed a sample size formula based on the supremum of a weighted version of the log rank test. These formulae require working assumptions on the relationship between the time to early nonresponse and the failure time. As will be seen these assumptions are unnecessary in the approach proposed here.

2 Test Statistics

Suppose that n subjects are to be randomized to one of two first-stage treatments, denoted by $A_1 = 1,\ 2$. For example, in the study in attention deficit hyperactivity disorder these correspond to low intensity behavioral modification therapy or low dose methylphenidate, respectively. Nonresponders are further randomized to one of two second-stage treatments, denoted by $A_2 = 1,\ 2$. These could be intensification of current treatment or augmentation of current treatment, respectively. Recall that responding subjects are assigned a fixed second-stage treatment which could be the same as the initial treatment. Let T denote the failure time, S denote the time to early nonresponse and C denote the censoring time. On each subject we observe $A_1,\ R,\ RA_2,\ S',\ \Delta$ and U, where $R = I\{S \leq \min(T, C)\}$ is the nonresponse indicator or equivalently in this case, the rerandomization indicator, $S' = \min(T, S, C)$, $\Delta = I(T \leq C)$ and $U = \min(T, C)$. It is worthwhile to mention that, in different trials, the nonresponse, or alternatively the response, is assessed differently. For example, in some trials it is assessed at a fixed time point after the initial treatment. Then the definition of R should be changed accordingly. However, this will not affect our development in the sequel, as long as R is the indicator for rerandomization. As is customary, we assume throughout that the censoring time C is independent of all other variables including A_1 and A_2. Note that S can be censored by either T or C; indeed the failure time T may occur prior to the time to nonresponse, S. Denote the randomization probabilities by $p = \mathrm{pr}(A_1 = 1)$ and $q = \mathrm{pr}(A_2 = 1 \mid R = 1)$. Denote jk to be the treatment strategy for which $A_1 = j$ and then $A_2 = k$ for nonresponders, for $j = 1, 2$ and $k = 1, 2$. Finally, the duration of the study is τ. Data from this trial design provides information on four adaptive treatment strategies, i.e., strategies 11, 12, 21 and 22. The trial design described here is just a special case of general two-stage randomized trial designs. Sample size formulae for other types of trial designs are discussed in the Supplementary Material.

A variety of statistics have been proposed for comparing two adaptive treatment strategies with data from two-stage randomized trials and can thus be used to construct test statistics and related sample size formulae. A first approach is to compare survival probabilities under different strategies at a certain time point, for example, survival probabilities at the end of the study period. Estimators of survival functions can be found in Lunceford et al. (2002) who proposed several versions of a weighted sample proportion estimator, in Wahed & Tsiatis (2006) who derived both a semiparametric efficient estimator and a less efficient, but easier to implement estimator, in Guo & Tsiatis (2005) who proposed a weighted Nelson–Aalen estimator of the cumulative hazard function, and recently, in Miyahara & Wahed (2010) who proposed a weighted Kaplan–Meier estimator. A second approach is to compare survival functions of the two adaptive treatment strategies using a weighted version of the log rank test as in a 2005 PhD thesis by Xiang Guo at the Department of Statistics, North Carolina State University, which can be found at: http://www.lib.ncsu.edu/resolver/1840.16/5768.

We utilize test statistics based on the two approaches discussed above to derive sample size formulae. In the following we consider powering the two-stage randomized trial to detect a difference, if any, between strategies 11 and 21. Comparisons between strategies 12 and 21, 11 and 22, and 12 and 22 are similar. In the context of the study in attention deficit hyperactivity disorder this means that the trial is sized to ensure power to compare two competing approaches to clinical management of the disease: using behavioral modification therapy with intensification if needed, which is favored by clinical psychologists, versus using medication with intensification if needed, which is favored by physicians. Since only strategies 11 and 21 are considered in the following, to simplify notation, from here on we denote the survival function, cumulative hazard function and hazard function of the failure time under strategy 11 as $\bar{F}_1(t)$, $\Lambda_1(t)$ and $\lambda_1(t)$ and those under strategy 21 as $\bar{F}_2(t)$, $\Lambda_2(t)$ and $\lambda_2(t)$.

All of the statistics discussed above involve weights, which come in two forms: time independent weights or time dependent weights. Weights are required because in this design, for any given adaptive treatment strategy, responding subjects who have an initial treatment consistent with this strategy are over represented in this strategy and nonresponding subjects who have an initial treatment consistent with this strategy are under represented. For example, all subjects starting with $A_1 = 1$ who responded have a treatment sequence that is automatically consistent with strategy 11, while only a proportion of subjects started with $A_1 = 1$ who failed to respond have a treatment sequence that is consistent with strategy 11, which depends on the second randomization probability q. To adjust for this design characteristic we use inverse probability weights (Robins et al., 1994; Murphy et al., 2001):

$$W_j = \frac{I(A_1 = j)}{p}\left\{1 - R + \frac{I(A_2 = 1)}{q}R\right\},\ j = 1, 2,$$

for strategies 11 and 21, respectively, where $I(\cdot)$ denotes the indicator function. These weights are similar to the weights used in Lunceford et al. (2002) and Miyahara & Wahed (2010). The difference is that they consider strategies with the same first-stage treatment and hence the factors $I(A_1 = 1)/p$ and $I(A_1 = 2)/(1-p)$ are omitted.

Alternatively one could use time dependent weights:

$$W_j(t) = \frac{I(A_1 = j)}{p}\left\{1 - R(t) + \frac{I(A_2 = 1)}{q}R(t)\right\},\ j = 1, 2,$$

where $R(t) = I\{S \leq \min(t, T, C)\}$. These time dependent weights are used in Guo's dissertation and by Feng & Wahed (2008). Time dependent weights permit more subject information to be used in the estimation at each time t. Consider strategy 11 and subjects who have $A_1 = 1$ as the initial treatment. If time dependent weights are used then all of these subjects have a nonzero time dependent weight at time points t less than $\min(S, T, C)$, i.e., before they meet the nonresponse criterion, regardless of the value of A_2. However, if time independent weights are used then those subjects who have $A_2 = 2$ will never have a nonzero weight for strategy 11. Hence, intuitively the use of time dependent weights results in efficiency gains compared with the time independent weights.

We first consider the test statistic based on the weighted Kaplan–Meier estimator. This test statistic is used for testing for $H_0 : \bar{F}_1(t) = \bar{F}_2(t)$ 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 & Wahed (2010), which is defined as

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

where the subscript i indicates the ith subject, $i = 1, \ldots, n$ and $N(u) = \Delta I(U \leq u)$ is the counting process for the failure time, and $Y(u) = I(U \geq u)$ is the at risk process. The estimator in (1) uses time dependent weights; however we can also use time independent weights by replacing $W_{ji}(u)$ with W_{ji} in (1). 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 in data analysis for evaluating the sample size formula the test statistic with time dependent weights will also be considered. It is worth mentioning that another estimator of $\bar{F}_j(t)$ can be defined as $\exp\{-\hat{\Lambda}_j(t)\}$, where $\hat{\Lambda}_j(t)$ is the weighted Nelson–Aalen estimator of $\Lambda_j(t)$ proposed in Guo & Tsiatis (2005). Theorem 1 in the Supplementary Material implies that the asymptotic distribution of the weighted Kaplan–Meier estimator is the same as that of this estimator.

By Theorem 1 in the Supplementary Material, the asymptotic distribution of $n^{1/2}\{\hat{\bar{F}}_{Kj}(t) - \bar{F}_j(t)\}$ is $N\{0, \sigma^2_{Kj}(t)\}$, where

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

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

$$T_K(t) = \frac{n^{1/2}\{\hat{\bar{F}}_{K1}(t) - \hat{\bar{F}}_{K2}(t)\}}{\{\hat{\sigma}^2_{K1}(t) + \hat{\sigma}^2_{K2}(t)\}^{1/2}},$$

where

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

is a consistent estimator of $\sigma^2_{Kj}(t)$, for $j = 1, 2$. The variance of the numerator of the test statistic is the sum of variances since $\hat{\bar{F}}_{K1}(t)$ and $\hat{\bar{F}}_{K2}(t)$ are independent. This independence occurs because the two strategies begin with different treatments; when the strategies begin with the same treatment, the Kaplan–Meier estimators are dependent. See the Supplementary Material, in the paragraph following the proof of Theorem 1, for details regarding this situation. If time independent weights are used in the weighted Kaplan–Meier estimator then $W_j(u)$ and $W_{ji}(u)$ in the above expressions for $\sigma^2_{Kj}(t)$ and $\hat{\sigma}^2_{Kj}(t)$ are replaced by W_j and W_{ji}, respectively, and the resulting estimators of variances remain consistent. Under the null hypothesis, the test statistic $T_K(t)$ has an asymptotic $N(0, 1)$ distribution.

Now we consider the weighted log rank statistic. This test statistic is used for testing $H_0 : \bar{F}_1 \equiv \bar{F}_2$ versus $H_1 : \bar{F}_1(t) \neq \bar{F}_2(t)$ for some $t \leq \tau$. The log rank test is the most commonly used test for comparing the distributions of two failure times. It is also commonly used to calculate sample sizes in classical survival analysis. See, for example, Schoenfeld (1981). Weighting each subject as above, the following statistic is an analogue of the usual log rank statistic:

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

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 Guo's dissertation, uses time dependent weights. One can use time independent weights as well, just by replacing $W_{ji}(t)$ in the definition with W_{ji}. By Theorem 3 in the Supplementary Material, the asymptotic distribution of $\sqrt{n} L_n$ under the null hypothesis is $N\{0, (\sigma^2_{L1} + \sigma^2_{L2})/4\}$, where

$$\sigma^2_{Lj} = E\left[\int_0^\tau W_j(t)\{dN(t) - Y(t) d\Lambda_1(t)\} \right]^2, \; j = 1, \; 2.$$

Again, see the Supplementary Material, the paragraph following the proof of Theorem 3, for testing two strategies starting with the same initial treatment. Let

$$\hat{\sigma}^2_{Lj} = \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^\tau W_{ji}(t)\{dN_i(t) - Y_i(t) d\hat{\Lambda}_1(t)\} \right]^2, \; j = 1, 2,$$

where

$$d\hat{\Lambda}_1(t) = \frac{\sum_{i=1}^{n} W_{1i} dN_i(t) + \sum_{i=1}^{n} W_{2i} dN_i(t)}{\sum_{i=1}^{n} W_{1i} Y_i(t) + \sum_{i=1}^{n} W_{2i} Y_i(t)}$$

is obtained by pooling the two groups. We can use the following test statistic to test for $H_0 : \bar{F}_1 \equiv \bar{F}_2$:

$$T_L = \frac{2n^{1/2} L_n}{\left(\hat{\sigma}_{L1}^2 + \hat{\sigma}_{L2}^2\right)^{1/2}}.$$

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 will use the test with time independent weights for sample size calculation but for data analysis we also consider the test with time dependent weights. 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 estimator and with the literature in this area. Lastly, Lokhnygina & Helterbrand (2007) proposed a pseudo score test of the log hazard ratio in a Cox proportional hazards model for comparing two adaptive treatment strategies. In their pseudo score function, each subject is weighted by a time independent weight. The statistic L_n defined above with time independent weights is equivalent to the pseudo score function defined there, but their expression of its asymptotic variance is in a different form.

3 Sample Size Calculation

As mentioned above, 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 comparatively simple sample size formulae, which will be seen below. On the other hand we suggest that test statistics using time dependent weights be used in the data analyses as these tests are potentially more powerful. We first derive sample size formulae using exact variances, and then replace the variances in the sample size formulae with their upper bounds to get our final sample size formulae.

First suppose that we wish to test $H_0 : \bar{F}_1(t) = \bar{F}_2(t)$ versus $H_1 : \bar{F}_1(t) \neq \bar{F}_2(t)$ for some t satisfying $0 < t \leq \tau$ using the test based on the weighted Kaplan–Meier estimator with time independent weights. For definiteness, we suppose that the survival probabilities at the end of the study are to be compared, i.e, $t = \tau$, and write T_K instead of $T_K(\tau)$. Under 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 in the Supplementary Material, the distribution of T_K is approximately normal with mean $n^{1/2}\{\bar{F}_1(\tau) - \bar{F}_2(\tau)\}/\left\{\sigma_{K1}^2(\tau) + \sigma_{K2}^2(\tau)\right\}^{1/2}$ and variance 1 under the alternative hypothesis. To detect a difference in survival probabilities at time τ of size $\delta_K = \bar{F}_1(\tau) - \bar{F}_2(\tau)$ with power $1 - \beta$, we set

$$\text{pr}\left(T_K > Z_{1-\frac{\alpha}{2}} \text{ or } T_K < -Z_{1-\frac{\alpha}{2}}\right) = 1 - \beta,$$

which yields the sample size formula

$$n_K \approx \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \left\{\sigma^2_{K1}(\tau) + \sigma^2_{K2}(\tau)\right\}}{\left\{\bar{F}_1(\tau) - \bar{F}_2(\tau)\right\}^2}.$$

Next suppose that we wish to test $H_0 : \bar{F}_1 \equiv \bar{F}_2$ versus $H_0 : \bar{F}_1(t) \neq \bar{F}_2(t)$ for some $t \leq \tau$ and the weighted log rank test with time independent weights is used. To construct a sample size formula based on the log rank test or its variants, the asymptotic distributions of the test statistics are usually derived under a proportional hazards assumption with a local alternative for the log hazards ratio (Schoenfeld, 1981; Chow et al., 2005; Eng & Kosorok, 2005; Feng & Wahed, 2008). The proportional hazards assumption is $\lambda_2(t) = \lambda_1(t)e^{\xi}$, where ξ is the log hazard ratio. And the local alternative assumption we use is represented by $\xi = \gamma/n^{1/2}$ for some constant γ. The use of a local alternative greatly simplifies the asymptotic means of the test statistics, which facilitates the sample size calculation. We use this approach here as well. Guo's dissertation studied the asymptotic distribution of the weighted log rank statistic L_n with time dependent weights. While he provided an outline, we provide a complete proof of the asymptotic normality of L_n under the local proportional hazards alternative in the Supplementary Material. Based on the asymptotic distribution of L_n given in Theorem 3 in the Supplementary Material and using a similar derivation as above, the corresponding sample size formula can be obtained as

$$n_L \approx \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \left(\sigma^2_{L1} + \sigma^2_{L2}\right)}{\xi^2 \left\{\int_0^\tau \bar{F}_C(t) dF_1(t)\right\}^2}.$$

To use the above formulae to calculate sample sizes we need values of $\sigma^2_{K1}(\tau)$ and $\sigma^2_{K2}(\tau)$ or σ^2_{L1} and σ^2_{L2}. A challenge is that, even with time independent weights, these variances depend in a complex manner on the joint distribution of the failure, T, and time to early nonresponse, S. In particular, R in the weight functions W_1 and W_2 depends on S; furthermore, as discussed in the Supplementary Material, the weights are not predictable, thus one cannot simplify the variance formulae by the usual martingale arguments. Thus a direct approximation of the values for these variances requires working assumptions on the joint distribution of T and S. We avoid making working assumptions on the joint distribution by the use of upper bounds on the variances; this then permits the use of working assumptions similar to those used to size standard one-stage trials.

To obtain interpretable upper bounds on the variances, we use potential outcomes (Holland, 1986) notation. The use of potential outcomes permits us to form upper bounds that are easily interpretable to scientists and thus enables the scientist to more easily provide the necessary information needed in the sample size calculation. Furthermore the potential outcome notation allows us to make our arguments precise. Let S_j and $\tilde{T}_j$ be the time to early nonresponse and the failure time, respectively, if a subject is assigned first-stage treatment j. The $\tilde{T}_j$s correspond to failure times in a one-stage study; that is, a study in which there is no change in treatment assignment. If $S_j < \tilde{T}_j$ then define D_{jk} to be the interval of time between early nonresponse and the failure time if the subject is assigned second-stage treatment k. Intuitively the failure time in response to the adaptive treatment strategy, jk, is the mixture of two times, one failure time in which there would have been no further treatment assignment, and the other failure time in which there would be a further treatment assignment once early nonresponse occurs. That is, the potential failure time under the assignment of strategy jk is $T_{jk} = \tilde{T}_j I(S_j > \tilde{T}_j) + (S_j + D_{jk}) I(S_j \le \tilde{T}_j)$. On the event $\{S_j > \tilde{T}_j\}$, $T_{j1} = T_{j2}$. We make the following consistency assumptions (Robins, 1997). If a subject is assigned first-stage treatment j then the time to early nonresponse S is equal to S_j; similarly if the assigned first-stage treatment is j, then $T = T_{jk}$ for all nonresponding subjects assigned second-stage treatment k. The failure time for all responding subjects satisfies $T = T_{jk}$ for $k = 1, 2$.

Now consider the variance term $\sigma^2_{K1}(\tau)$. First, when $W_1 \neq 0$, $N(t) = \Delta I(T \le t)$ and $Y(t) = I(U \ge t)$ can be replaced by $N_1(t) = I(T_{11} \le t, T_{11} \le C)$ and $Y_1(t) = I\{\min(T_{11}, C) \ge t\}$, respectively. Thus

$$\sigma^2_{K1}(\tau) = \bar{F}_1^2(\tau) E\left[\int_0^\tau \frac{W_1}{\bar{F}_1(t)\bar{F}_C(t)} \{dN_1(t) - Y_1(t) d\Lambda_1(t)\}\right]^2.$$

Next since $N_1(t)$ and $Y_1(t)$ are functions of T_{11} and C only, and the weight is time independent, the above is

$$\sigma^2_{K1}(\tau) \quad = \quad \bar{F}_1^2(\tau) E\left(E(W_1^2 \mid T_{11}, C) \left[\int_0^\tau \frac{1}{\bar{F}_1(t)\bar{F}_C(t)} \{dN_1(t) - Y_1(t) d\Lambda_1(t)\}\right]^2 \right).$$

Because the randomization of first-stage and second-stage treatments ensures that A_1 and A_2 are independent of the potential outcomes for the S_js and the T_{jk}s, we can replace $E(W_1^2 | T_{11}, C)$ in the above display by $E(1 - R + R/q \mid T_{11}, C)/p$. Next replace R by 1 to produce the upper bound,

$$\begin{aligned}
\sigma^2_{K1}(\tau) &\le \frac{\bar{F}_1^2(\tau)}{pq} E\left[\int_0^\tau \frac{1}{\bar{F}_1(t)\bar{F}_C(t)} \{dN_1(t) - Y_1(t) d\Lambda_1(t)\}\right]^2 \\
&= \frac{\bar{F}_1^2(\tau)}{pq} \int_0^\tau \frac{d\Lambda_1(t)}{\bar{F}_1(t)\bar{F}_C(t)}, \qquad (2)
\end{aligned}$$

where martingale theory is used to obtain the integral in the last equality. An upper bound for $\sigma^2_{K2}(\tau)$ is obtained similarly, and it is the bound where we replace p with $(1 - p)$ and the subscript 1 by subscript 2 in (2). A similar argument can be used to derive the upper bounds for σ^2_{L1} and σ^2_{L2} as well; in the case of σ^2_{L1} the upper bound is

$$\sigma^2_{L1} \le \frac{1}{pq} \int_0^\tau \bar{F}_C(t) dF_1(t).$$

The upper bound on σ^2_{L2} is similar, but p is replaced by $(1 - p)$ and the subscript 1 is replaced by subscript 2.

Since the upper bounds are obtained by replacing R with 1, the larger proportion of subjects randomized at the second stage, or equivalently, the larger proportion of nonresponders, the sharper the upper bounds. In the case that every subject is randomized at the second stage, for example, if all subjects are nonresponders, the upper bounds are precise.

Now replacing the variances in the above sample size formulae by their upper bounds, the sample size based on the weighted Kaplan-Meier estimator with time independent weights and upper bounds on the variances is

$$n_K \leq \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \sigma_B^2}{\left\{\bar{F}_1(\tau) - \bar{F}_2(\tau)\right\}^2}, \tag{3}$$

where

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

In a similar manner, the sample size based on the weighted log rank test with time independent weights and upper bounds on the variances is

$$n_L \leq \left\{\frac{1}{pq} + \frac{1}{(1-p)q}\right\} \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{\xi^2 \int_0^\tau \bar{F}_C(t) dF_1(t)}. \tag{4}$$

In order to calculate the sample size using formula (4), we only need information about the hazard ratio and the integral $\int_0^\tau \bar{F}_C(t)dF_1(t)$, which is exactly equal to the probability of observing an event before time τ when all subjects are assigned strategy 11. This is the same information used to size a two-arm one-stage trial using the log rank test (Schoenfeld, 1981). In contrast, if we use formula (3) to calculate the sample size, we need working assumptions on the distribution function of the potential failure time under each strategy, i.e., $F_1(t)$ and $F_2(t)$, as well as $F_C(t)$, the distribution function of the censoring time. In practice, one could assume these functions have a parametric form, for example, exponential or Weibull distributions. For the censoring distribution $\bar{F}_C(t)$, a uniform distribution over $(0, \tau)$ with a point mass at τ is often reasonable. Then one only needs to guess at the parameters in these distributions to calculate the integrals in the formula. Alternatively, one could make guesses at these distribution functions at some fixed time points before τ, and then approximate the integrals using numerical approximation.

Although formula (3) needs more inputs than (4), it is made much simpler by using upper bounds on variances. We illustrate this by comparing our working assumptions needed for (3) with those in a sample size formula developed by Feng & Wahed (2009). Their sample size formula is also based on testing the equality of survival probabilities at a fixed time point but using the second weighted sample proportion estimator with time independent weights proposed in Lunceford et al. (2002). Instead of forming upper bounds on the variances, they simplify the variance terms by making working assumptions on the joint distribution of the potential outcomes. They use slightly different potential outcomes than used here. In their setting R_j denotes the nonresponse indicator under first-stage treatment j; T_{j0} is the failure time if the subject responds and T_{jk}^* is the failure time if the subject does not respond and is assigned treatment k in the second stage. Both times are durations beginning at the start of first-stage treatment. Then the failure time in response to strategy jk is $T_{jk} = (1 - R_j)T_{j0} + R_j T_{jk}^*$. Using these potential outcomes they assume that $E(R_j \mid T_{jk}) = E(R_j)$. See, for example, their derivation of display (12) and their simulation models. One might interpret this working assumption as: the chance that a subject exhibits early nonresponse is independent of the subject's failure time. Since they do not use upper bounds on variances, they need to make more working assumptions than our formula (3). Besides the assumption just mentioned, they also need assumptions about the distribution functions of T_{j0} and T^*_{jk}, and the nonresponse rates $\mathrm{pr}(R_j = 1)$, $j, k = 1, 2$, to calculate the sample size.

We further compare working assumptions used by the different sample size formulae in the context of the trial for attention deficit hyperactivity disorder. First consider sizing the trial to test if the chance of a school disciplinary event occurring by 36 weeks differs between strategies 11 and 21. In this example, the working assumption required to use the formula in Feng and Wahed (2009) is that, for both initial treatments, early nonresponse is independent of the time to a school disciplinary event. Moreover, to apply their formula, we need to know the distribution functions of the times to a school disciplinary event for both those who respond to behavioral therapy and those who respond to medication. We also need to know the distribution functions of time to a school disciplinary event for those who do not respond to behavioral therapy and are then assigned more intensified behavioral therapy, and for those who do not respond to medication and are then assigned higher dose medicine, respectively. Lastly, we need to know the proportion of subjects who respond to behavioral therapy and proportion of subjects who respond to medication. In contrast, the sample size formula (3) does not require the independence assumption stated above. Moreover, we only need to know the distribution functions of time to a school disciplinary event for subjects who are assigned strategy 11 and 21, respectively. If instead we wish to size the trial to test the difference between the distributions of the times to a school disciplinary event under strategies 11 and 21 using the weighted log rank test, then the working assumptions are even simpler. To use the sample size formula (4), our first assumption is proportional hazards between strategies 11 and 21. In addition, we need to specify the hazard ratio and the probability of observing the first school disciplinary event before the end of study in children who are assigned strategy 11. The working assumptions in general settings for the three sample size formulae are provided in Table 1.

"Table 1 about here."

4 Simulation

We conducted a simulation study to assess the performance of the sample size formulae. In evaluating the formulae we include tests based on time independent weights and tests based on time dependent weights. Moreover, for comparing the survival probabilities at a given time point, we include two additional tests. The first test is based on a slight generalization of the third weighted sample proportion estimator of $\bar{F}_j(t)$ in Lunceford et al. (2002), which is the most efficient among the three estimators proposed there. Our generalization uses time dependent weights $W_1(t)$ and $W_2(t)$ instead of the time independent weights. The second test is based on the pseudo semiparametric efficient estimator of $\bar{F}_j(t)$ proposed by Wahed & Tsiatis (2006).

In the simulations, we suppose $\tau = 36$. We generate (T_{j1}, S_j) jointly from a Frank copula model (Roger, 1998) with association parameter 5 when $j = 1$ and 6 when $j = 2$, resulting in a positive correlation between the failure time and the time to early nonresponse. The marginal distributions of T_{11} and T_{21} are Weibull distributions, with a common location parameter equal to 2 to ensure proportional hazards. The scale parameter for T_{11} is 50 and the scale parameter for T_{21} is determined by the desired hazard ratio. The hazard ratio takes values 1.25, 1.5 or 2.0. The marginal distributions of the times to early nonresponse, S_1 and S_2, also have Weibull distributions with both scale and location parameters varied so as to achieve varying percentages of subjects who are randomized in the second stage. Specifically, the percentage of subjects randomized at the second stage ranges from about 25% to about 75% throughout. The censoring time, C, has a point mass at τ and is otherwise uniformly distributed over $(0, \tau)$, and is independent of all other variables. The size of the point mass at τ is varied so that approximately 20% or 40% of the failure times are censored. The survival probabilities at the end of study under the two strategies 11 and 21 range in $(0.4, 0.6)$ among all scenarios. The first- and second-stage randomization probabilities are $p = q = 0.5$. All simulations are based on 1000 simulated data sets. The significance level is $\alpha = 0.05$ and the desired power is 80%.

Table 2 and Table 3 provide the results of the simulation. From both tables, we observe that the desired power is achieved by all the tests except the one based on the third estimator in Lunceford et al. (2002) when comparing survival probabilities at the end of study. The tests based on time dependent weights are more efficient than the corresponding tests based on time independent weights. By results in Table 2, the pseudo efficient estimator of Wahed and Tsiatis is slightly more efficient than the weighted Kaplan–Meier estimator with time dependent weights when the sample sizes are large. In smaller samples, the efficiency of the pseudo efficient estimator is reduced due to the variability introduced by the estimation of probabilities of potentially rare events. See Wahed & Tsiatis (2006) for details. Table 2 also includes sample sizes calculated from the formula in Feng & Wahed (2009). This sample size formula assumes that the potential outcomes R_j and T_{jk} are independent; this assumption does not hold here. Their sample sizes are a little larger than those calculated from our formula (3). This is possibly due to the relatively low efficiency of the sample proportion estimator compared with the weighted Kaplan-Meier estimator.

“Table 2 about here.”

“Table 3 about here.”

From Tables 2 and 3 we also observe that the degree of conservatism of the sample size formulae 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, when the percentage of subjects who are randomized at the second stage approaches or exceeds 70%, the achieved powers are close to the expected power. We can also use simulations to search for the nonconservative sample sizes and compare them with our conservative ones. For example, when the hazard ratio is 1.5, and when 25%, 50% and 75% subjects are rerandomized, the sample sizes needed to guarantee 80% power using the test based on weighted Kaplan–Meier estimator with time independent weights are about 680, 710 and 770, respectively, compared with 880 subjects required by our conservative sample size formula. When the weighted log rank test with time independent weights are used, the nonconservative sample sizes are about 620, 690 and 720, when 25%, 50% and 75% subjects are rerandomized, respectively, while our conservative sample size is 753.

Table 4 presents the results of another simulation which illustrate the achieved powers when working assumptions are incorrect, that is, the quantities needed to calculate the sample sizes are misspecified. We consider cases where the hazard ratio is 1.25 or 1.5. In the case of the sample size formula (3), the wrong scale or shape parameter is used in the specification of the Weibull distributions of T_{11} and T_{21}. In the case of the sample size formula (4), the probability of observing an event before the end of study is misspecified. The results show that, the desired powers for both tests are usually achieved or approximated under reasonable degree of misspecification. This robustness is a consequence of conservatism of the sample size formulae.

"Table 4 about here."

In all these simulations we generated data in which the failure times and the times to early nonresponse are positively associated. We also conducted simulations in which they are negatively associated; the results are very similar to those shown here.

5 Discussion

Although the focus of this work is on trials intended for use in refining or developing adaptive treatment strategies, in some situations two fully developed adaptive treatment strategies exist and scientific interest rest primarily in contrasting the strategies. In this case a traditional two-arm trial of the strategies is more appropriate than the sequential multiple assignment randomized trial discussed here and has the advantage of requiring a smaller sample size.

In this work we assume that the censoring time distribution does not depend on the initial treatment. In the case of the weighted log rank test, removing this assumption will result in a more complicated sample size formula. However, removing this assumption for the sample size formula based on the weighted Kaplan–Meier estimator requires minimal change as follows. Denote $\bar{F}_{Cj}(t)$ to be the survival function of the censoring time for subjects starting with initial treatment $A_1 = j$, $j = 1, 2$. Then it is easy to see that the asymptotic distribution of $\hat{\bar{F}}_j(t)$ is normal with the same mean but $\bar{F}_C(t)$ in the variance formula is replaced by $\bar{F}_{Cj}(t)$. Therefore, the corresponding sample size formula is the same as formula (3) except that σ_B^2 now becomes

$$\sigma_B^2 \quad = \quad \frac{\bar{F}_1^2(\tau)}{pq}\int_0^\tau \frac{d\Lambda_1(t)}{\bar{F}_1(t)\bar{F}_{C1}(t)} + \frac{\bar{F}_2^2(\tau)}{(1-p)q}\int_0^\tau \frac{d\Lambda_2(t)}{\bar{F}_2(t)\bar{F}_{C2}(t)}.$$

The simulation results in the previous section indicate that the sample sizes obtained using the weighted log rank test are usually considerably smaller than those obtained from comparing survival probabilities at a given time point. Moreover, less information is needed to calculate the sample size derived from the weighted log rank test. Hence we have developed a Web applet which can be used to size studies based on the weighted log rank test. This Web applet, along with the simulation code for our simulations, can be found at http://methodologymedia.psu.edu/logranktest/samplesize. However, the log rank test is designed to have highest power against proportional hazards (Peto & Peto, 1972), thus the use of the sample size formula based on the log rank test may not provide the desired power if the hazards under the alternate hypothesis are likely nonproportional. The test for equality between survival probabilities at one time point has a more modest goal, hence may be better able to ensure power.

Feng & Wahed (2008) derived a sample size formula for the same problem using a supremum weighted log rank test comparing two strategies. Their sample size formula requires more detailed working assumptions than that required by the formula based on the weighted log rank test given here. It would be desirable to develop a sample size formula based on the supremum weighted log rank test that requires less detailed working assumptions.

SUPPLEMENTARY MATERIAL

Supplementary Material available at Biometrika online includes a discussion of sample size formulae for two variants of two-stage randomized trials and proofs of theoretical results.

ACKNOWLEDGEMENTS

The authors thank the editor, an associate editor and the referee for valuable comments. This research was supported by the National Institute of Health, U.S.A.

References

CHOW S.C., SHAO J. & WANG H. (2005). SAMPLE SIZE CALCULATIONS IN CLINICAL RESEARCH. Chapman and Hall.

COLLINS L.M., MURPHY S.A., NAIR V. & STRECHER V. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine* **30**, 65-73.

ENG, K.H. & KOSOROK, M.R. (2005). A sample size formula for the supremum log rank statistic. *Biometrics* **61**, 86-91.

FENG, W. & WAHED, A.S. (2008). A supremum log rank test for comparing adaptive treatment strategies and corresponding sample size formula. *Biometrika* **95(3)**, 695-707.

FENG, W. & WAHED, S.A. (2009). Sample size for two-stage studies with maintenance therapy. *Statistics in Medicine*, **28**, 2028-2041.

GUO X. & TSIATIS A.A. (2005). A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics* **1(1)**, 1-15.

HOLLAND P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945-960.

LAVORI P.W., & DAWSON R. (2003). Dynamic treatment regimes: practical design considerations. *Clinical Trials* **1**, 9-20.

LAVORI P.W., DAWSON R. & ROTH A.J. (2000). Flexible treatment strategies in chronic disease: clinical and research implications. *Biological Psychiatry* **48**, 605 614.

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. (2005). Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Investigators. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine* **53(12)**, 1209-1223.

LOKHNYGINA Y. & HELTERBRAND J.D. (2007). Cox regression methods for two-stage randomization designs. *Biometrics* **63**, 422-428.

LUNCEFORD J.K., DAVIDIAN M. & TSIATIS, A.A. (2002). Estimation of survival distributions of treatment strategies in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48-57.

MAKAY J.R., LYNCH K.G., SHEPARD D.S., RATICHEK S., MORRISON R., KOPPENHAVER J. & PETTINATI H.M. (2004). The effectiveness of telephone-based continuing care in the clinical management of alcohol and cocaine use disorders: 12-month outcomes. *Journal of Consulting and Clinical Psychology* **72(6)**, 967-979.

MARLOWE D.B., FESTINGER D.S., ARABIA P.L., DUGOSH K.L., BENASUTTI K.M., CROFT J.R. & MCKAY J.R. (2008). Adaptive interventions in drug court: a pilot experiment. *Criminal Justice Review* **33(3)**, 343-360.

MIYAHARA S. & WAHED A.S. (2010). Weighted Kaplan–Meier estimators for two-stage treatment regimes. *Statistics in Medicine* **29**, 2581-2591.

MURPHY S.A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24**, 1455-1481.

MURPHY S.A., LYNCH K.G., OSLIN D., MCKAY J.R. & TENHAVE T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence* **88(2)**, S24-S30.

MURPHY S.A., VAN DER LAAN M.J., ROBINS J.M., CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of American Statistical Association* **96**, 1410-1423.

OETTING A.I., LEVY J.A., WEISS R.D. & MURPHY S.A. (2007). Statistical methodology for a SMART design in the development of adaptive treatment strategies. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (Shrout P.E., Ed.) Arlington VA: American Psychiatric Publishing, Inc.*

PETO R.& PETO J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society. Series A* **96**, 1410-1423.

ROBINS J.M. (1986). 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* **14**, 1393-1512.

ROBINS J.M. (1997). Causal inference from complex longitudinal data. *Lecture Notes in Statistics, Berkane M. Ed.* Springer Verlag, **120**, 69-117.

ROBINS J.M., ROTNITZKY A. & ZHAO L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846-866.

ROGER, B.N. (1998). AN INTRODUCTION TO COPULAS. Springer-Verlag, 1998.

RUSH A.J., FAVA M., WISNIEWSKI S.R., LAVORI P.W., TRIVEDI M.H., SACKEIM H.A., THASE M.E., NIERENBERG A.A., QUITKIN F.M. & KASHNER T.M. (2004) Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials* **25**, 119-142.

SCHOENFELD, D.A. (1981). The asymptotic properties of nonprametric tests for comparing survival distributions. *Biometrika* **68**, 316-319.

STONE, R.M., BERG, D.T., GEORGE, S.L., DODGE, R.K., PACIUCCI, P.A., SCHULMAN, P., LEE, E.J., MOORE, J.O., POWELL, B.L. & SCHIFFER, C.A. (1995). Granulocyte- macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myeloge- nous leukemia. *The New England Journal of Medicine* **332**, 1671-1677.

TUMMARELLO, D., MARI, D., GRAZIANO, F., ISIDORI, P., CETTO, G., PASINI, F., SANTO, A., & CELLERINO, R. (1997). 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* **80**, 2222-2229.

WAHED, A.S. & TSIATIS, A.A. (2006). Semiparametric efficient estimation of survival distribution in two-stage randomization designs in clinical trials with censored data. *Biometrika* **93**, 163-177.

Table 1: Methods for sample size calculation, working assumptions made, and quantities needing to be guessed to calculate the sample size

Method	**Working assumptions**	**Quantities needing to be guessed**
cWKM	none	$\bar{F}_j(t),\ j=1,2$, and $\bar{F}_C(t)$
cWSP	the response status is independent of the potential failure time	pr$(R_j=1)$,$\bar{F}_{j0}(t)$, $\bar{F}^*_{j1}(t)$, $j=1,\ 2$, and $\bar{F}_C(t)$
cWLR	proportional hazards	hazard ratio, pr(observe an event before τ under strategy 11)

cWKM, cWLR, cWSP: sample size formulae based on weighted Kaplan–Meier estimator and weighted log rank test with time independent weights, and sample size formula based on the weighted sample proportion estimator in Feng and Wahed (2009), respectively; $\bar{F}_{j0}(t)$, $\bar{F}^*_{j1}(t)$: survival functions of T_{j0} and T^*_{j1}, respectively; R_j: indicator for rerandomization if assigned initial treatment j.

Table 2: Achieved powers (%) for the sample sizes based on the weighted Kaplan–Meier estimator with time independent weights. The significance level of the test is 5% and the desired power is 80%.

Hazard ratio	$n_K(n^*)$	% rerandomized	cWKM	tWKM	Lunceford 3	WT
1.25	2862(2995)	25	88	91	85	93
		50	83	85	82	87
		75	81	83	76	85
1.5	880(934)	25	87	92	85	91
		50	82	84	76	88
		75	81	82	77	83
2	278(307)	25	87	90	78	87
		50	83	85	71	83
		75	80	82	68	81

n_K, n^*: sample sizes calculated from (3) and from the formula in Feng and Wahed (2009), respectively; cWKM, tWKM: the tests based on the weighted Kaplan–Meier estimator with time independent and time dependent weights, respectively; Lunceford 3: the test based on the third estimator in Lunceford et al. (2002); WT: the test based on the estimator in Wahed and Tsiatis (2006).

Table 3: Achieved powers (%) of the weighted log rank tests under sample sizes obtained by cWLR. The significance level of the test is 5% and the desired power is 80%.

		25% rerandomized		50% rerandomized		75% rerandomized	
Hazard ratio	n_L	cWLR	tWLR	cWLR	tWLR	cWLR	tWLR
1.25	2651	92	94	87	89	84	85
1.5	753	93	94	85	87	81	83
2	234	92	93	83	83	80	83

n_L: sample size based on the weighted log rank test with time independent weights;
cWLR, tWLR: weighted log rank test with time independent weights and time dependent weights, respectively.

Table 4: Achieved powers (%) of tests under misspecifications of the true model. In the true model, (T_{11}, S_1) and (T_{21}, S_2) both follow Frank copula models with association parameters 1 and 2, respectively. The marginal distributions of T_{11} and T_{21} are Weibull with scale parameter and shape parameter specified below. The first and fourth columns indicate the misspecification of the true model. All simulated powers are those of the same tests as used to calculate the sample size. The significance level of the tests is 5% and the desired power is 80%.

cWKM			cWLR		
Specification of the model	n_K	Power	Specification of the model	n_L	Power
True model 1: scaleT11=50, shapeT11=shapeT21=2, hazard ratio=1.25 pr(observe an event before τ)=0.61					
no misspecification	2862	91	no misspecification	2651	92
scaleT11=45	2405	84	pr(event)=0.40	1726	78
scaleT11=55	3729	94	pr(event)=0.45	1953	84
shapeT11=shapeT21=1.75	3812	96	pr(event)=0.38	1658	73
shapeT11=shapeT21=2.25	2385	84	pr(event)=0.65	2825	96
exponential	2478	81	pr(event)=0.70	3045	98
True model 2: scaleT11=50, shapeT11=shapeT12=2, hazard ratio=1.5 pr(observe an event before τ)=0.37					
no misspecification	880	91	no misspecification	753	90
scaleT11=45	647	86	pr(event)=0.28	578	82
scaleT11=55	1096	94	pr(event)=0.25	510	77
shapeT11=shapeT21=1.75	978	95	pr(event)=0.36	737	71
shapeT11=shapeT21=2.25	794	83	pr(event)=0.48	973	95
exponential	601	80	pr(event)=0.50	1015	97

scaleTjk and shapeTjk: the scale and shape parameters of T_{jk}, respectively; pr(event): pr(observe an event before τ); exponential: the marginal distributions of T_{11} and T_{21} are supposed to be exponential distributions with the same survival probabilities as the true model at τ; n_K, n_L: sample sizes calculated from (3) and (4), respectively.