# Statistical Considerations for Sequential Analysis of the Restricted Mean Survival Time for Randomized Clinical Trials

Ying Lu and Lu Tian

Department of Biomedical Data Science and Center for Innovative Study Design

Stanford University, Stanford, CA 94305-5464, USA

ylu1@stanford.edu

**Keywords:** Restricted Mean Survival Time (RMST), Group Sequential Design, Sample Size, Interim Analysis

**Abstract**

In this paper, we illustrate the method of designing a group-sequential randomized clinical trial based on the difference in restricted mean survival time (RMST). The procedure is based on theoretical formulations of Murray and Tsiatis (1999). We also present a numerical example in designing a cardiology surgical trial. Various practical considerations are discussed. R codes are provided in the Supplementary Materials. We conclude that the group-sequential design for RMST is a viable option in practice. A simulation study is performed to compare the proposed method to the Max-Combo and conventional log-rank tests. The simulation result shows that when there is a delayed treatment benefit and the proportional hazards assumption is untrue, the sequential design based on the RMST can be more efficient than that based on the log-rank test but less efficient than that based on the Max-Combo test. Compared with Max-Combo test, the RMST-based study design yield coherent estimand, statistical inference and result interpretation.

## Introduction

When the endpoint of a prospective randomized clinical trial is time to an event of interest, such a death and disease progression, the Cox model is the commonly used analytic method to test and estimate the treatment effect in terms of hazard ratio (HR). Oftentimes, the difference or ratio in median survivals or survival probabilities at a fixed time point from two arms is also used to supplement the HR for summarizing the size of the treatment benefit for their transparent interpretations for clinicians. The HR-based method works well in past decades and facilitates numerous advancements in drug development. However, the oncology treatment undergone a rapid revolution more recently: from cytotoxic chemotherapies to cytostatic targeted therapies, to immunotherapies, and to cell-based therapies. Oftentimes, the new treatment has a delayed but sustained survival benefit in comparison with the standard care. In such a case, the proportional hazards (PH) assumption, which is the key to perform statistical analyses using Cox model, becomes problematic. For example, Garon et al (2015) reported the progression-free survival distributions in patients with advance melanoma in three groups categorized by proportion score (PS) and showed that these three Kaplan Meier (KM) curves were approximately the same up to month 2 and then started to separate with the highest PS group with the best survival profile. This is a strong indication that the PH assumption is violated. In the second example, Robert et al. (2015) reported survival distributions in patients with small-cell lung cancer (SCLC) in three treatment groups (Pembrolizumab Q3w, Pembrolizumab Q2w, and Ipilimumab). The KM curves of the survival distributions in Pembrolizumab Q3w and Q2w are similar but crossed several times, again suggesting the violation of the PH assumption.

There are several potential reasons for the violation of the PH assumption. For example, the PH assumption would be violated if the patient population can be divided into several strata with different stratum-specific baseline hazards and the PH assumption is statisfied only within each stratum (Tian, et al., 2019).

Several new approaches have been proposed as alternative to the HR-based statistical inference in analyzing survival data from randomized clinical trials. For example, one may measure the treatment benefit by the difference in median survival as discussed above. However, with the improvement in patient survival, the median survival time may not be estimable due to the limited trial duration (Uno, et al. 2014, 2015[a]). One may also impose a flexible piece-wise exponential model allowing the HR to vary with time and employ the maximum likelihood method to estimate the difference in two survival distributions. When the primary objective is to test the presence of treatment benefit, many flexible tests have been proposed in the literature. For example, Chang and McKeague (2019) proposed to the use empirical likelihood method to test the difference between two or more stochastically ordered survival distributions. Karrison (2016) proposed to use a Max-Combo test, which has a more robust performance than the simple log-rank test, especially when the PH assumption is violated. Several omnibus tests based on the weighted difference between two KM curves has been proposed by Shen and Cai (2001) and Uno et al. (2015)[b], which are also more robust in detecting nonPH alternatives.

In this paper, we focus on an appealing alternative method based on the restricted mean survival time (RMST) (Royston and Parmar, 2013, Trinquart, et al, 2016, 2019, Uno et al. 2014). The $\tau$-RMST is defined as $E(T \wedge \tau)$, the expectation of a truncated survival time, where $T$ is the event time of interest, $\tau$ is a given truncation time point and $a \wedge b = \min(a, b)$. The RMST up to time $\tau$ also equals to

$$\int_0^\tau S(t)dt, \tag{1}$$

where $S(t) = P(T > t)$ is the survival function of $T$. In two group comparisons, the statistical test can be based on a contrast of RMSTs from two arms. For example, we may estimate

$$D(\tau) = E(T_1 \wedge \tau) - E(T_0 \wedge \tau),$$

the difference between two RMSTs up to time $\tau$, by

$$\hat{D}(\tau) = \int_0^\tau \big(\hat{S}_1(t) - \hat{S}_0(t)\big)dt, \quad (2)$$

as a metric of the treatment benefit, where $T_j$ is the survival time in arm $j$, $\hat{S}_j(t)$ is the KM estimator for $T_j$, $j = 0,1$. The 95% confidence interval (CI) can be constructed based on the large sample approximation of the distribution of $\hat{D}(\tau)$. The main advantages of RMST-based analysis include its clinical interpretability and associated robust nonparametric inference. Furthermore, in clinical trials for some disease and treatment combinations, clinicians are interested in time to a clinical event only within a certain time window, because the differences beyond that time interval may not be biologically attributable to the treatment.

Despite those merits, there are many challenges to use RMST as the primary analysis in a randomized clinical trial. Especially, it is not clear how to design and analyze a group-sequential study based on RMST. Murray and Tsiatis (1999) proposed statistical inference procedures for group-sequential studies using the RMST, but did not cover many practical issues related to study design such as sample size estimation, which hinders its penetration to practice (Fleming et al., 1984). Hasegawa et al. (2020) has discussed the statistical information of RMST difference in group-sequential setting by establishing the connection between the RMST differences and weighted logrank test statistics, which shed more light on our understanding of the operational characteristics of RMST-based method in hypothesis testing. In this paper, we will focus on how to design a group-sequential study using the RMST in practice.

**Method**

Designing a group-sequential study needs to consider the following factors:

- $K$: the number of interim analysis;
- $t^{(k)}$: the calendar time for the $k$th interim analysis (without loss of generality, we assume that calendar time of the study initiation is 0);
- $t_F = t^{(K+1)}$: the calendar time of the planned final analysis;
- $\tau^{(k)}$: the truncation time chosen for the RMST in the $k$th interim analysis;
- $\tau_F = \tau^{(K+1)}$: the truncation time chosen for the RMST in the final analysis, and
- $\alpha_k$: the type one error spent at the $k$th interim analysis.

For simplicity, we only consider early stopping for superiority in contrast to futility. In order to describe the inference procedure at the interim and final analyses, we introduce the following notations:

Let $E_{ij}$ be the calendar time, at which the patient $i$ from arm $j$ is enrolled; $T_{ij}$ and $F_{ij}$ be survival time and time to potential loss of follow-up for the same patient, respectively. At the $k$th interim analysis, the censoring time for patient $i$ from arm $j$ is

$$C_{ij}^{(k)} = F_{ij} \wedge (t^{(k)} - E_{ij}), \tag{3}$$

for $E_{ij} < t^{(k)}$, i.e., the patient is enrolled before the interim analysis. Therefore, survival data observed at the interim analysis are

$$Q_j^{(k)} = \left\{ \left( X_{ij}^{(k)}, \delta_{ij}^{(k)} \right) = \left( T_{ij} \wedge C_{ij}^{(k)},\ I\left( T_{ij} < C_{ij}^{(k)} \right) \right) \middle|\ \ i = 1, \cdots, n_j,\ E_{ij} < t^{(k)} \right\}, j = 0, 1,$$

and we may estimate, $D\left(\tau^{(k)}\right)$, the difference in RMST up to time $\tau^{(k)}$ , by

$$\widehat{D}_k\left(\tau^{(k)}\right) = \int_0^{\tau^{(k)}} \left( \hat{S}_1^{(k)}(t) - \hat{S}_0^{(k)}(t) \right) dt, \tag{4}$$

where $\hat{S}_j^{(k)}(\cdot)$ is the KM estimator of $S_j(\cdot)$ based on data $Q_j^{(k)}$. Here we assume that

$$\max_{i=1,\cdots,n_1} \{X_{i1}^{(k)}\} \wedge \max_{i=1,\cdots,n_0} \left\{ X_{i0}^{(k)} \right\} \geq \tau^{(k)}, \tag{5}$$

i.e., the largest follow-up times at the $k$th interim analysis in both arms are greater than $\tau^{(k)}$, to ensure the identifiability of $D\left(\tau^{(k)}\right)$. Under the null hypothesis, it is clear that

$$\sqrt{n}\begin{pmatrix} \widehat{D}_1\left(\tau^{(1)}\right) \\ \cdots \\ \widehat{D}_K(\tau^{(K)}) \\ \widehat{D}_{K+1}(\tau^{(K+1)}) \end{pmatrix} \sim N(0, \quad \Sigma_{K+1}(\pi_1)) \tag{6}$$

where $n = n_1 + n_0$ is the total sample size, $\pi_j = n_j/n$ is the proportion of patients randomized to arm $j$, and the subscript $K+1$ represents the final analysis. Note that at the $k$th interim analysis, $\left(\widehat{D}_1\left(\tau^{(1)}\right), \cdots, \widehat{D}_k\left(\tau^{(k)}\right)\right)'$ also follows a multivariate Gaussian distribution centered at zero with a variance-covariance matrix of $\Sigma_k(\pi_1)$, under the null hypothesis. In order to determine the rejection region at the $k$th interim analysis, we note the expansion

$$\sqrt{n}\begin{pmatrix} \widehat{D}_1\left(\tau^{(1)}\right) \\ \cdots \\ \widehat{D}_k(\tau^{(k)}) \end{pmatrix} = -\sum_{j=0}^{1} \frac{1}{\sqrt{n_j \pi_j}} \sum_{i=1}^{n_j} \begin{pmatrix} \int_0^{\tau^{(1)}} \frac{\int_t^{\tau^{(1)}} S_j(s)ds}{P(X_{ij}^{(1)} \geq t, E_{ij} < t^{(1)})} dM_{ij}^{(1)}(t) \\ \cdots \\ \int_0^{\tau^{(k)}} \frac{\int_t^{\tau^{(k)}} S_j(s)ds}{P(X_{ij}^{(k)} \geq t, E_{ij} < t^{(k)})} dM_{ij}^{(k)}(t) \end{pmatrix} + o_p(1), \tag{7}$$

where $M_{ij}^{(s)}(t) = I\left(E_{ij} < t^{(s)}\right)\left\{I\left(X_{ij}^{(s)} \leq t\right)\delta_{ij}^{(s)} - \int_0^t I\left(X_{ij}^{(s)} \geq s\right)\lambda_j(s)ds\right\}$ and $\lambda_j(s)$ is the hazard function of the survival distribution in arm $j$. Therefore, $\Sigma_k(\pi_1)$ can be consistently estimated as

$$\hat{\Sigma}_k(\pi_1) = \sum_{j=0}^{1} \frac{1}{n_j \pi_j} \sum_{i=1}^{n_j} \begin{pmatrix} \int_0^{\tau^{(1)}} \frac{\int_t^{\tau^{(1)}} \hat{S}_j^{(k)}(s)ds}{\hat{P}(X_{ij}^{(1)} \geq t, E_{ij} < t^{(1)})} d\widehat{M}_{ij}^{(1)}(t) \\ \cdots \\ \int_0^{\tau^{(k)}} \frac{\int_t^{\tau^{(k)}} \hat{S}_j^{(k)}(s)ds}{\hat{P}(X_{ij}^{(k)} \geq t, E_{ij} < t^{(k)})} d\widehat{M}_{ij}^{(k)}(t) \end{pmatrix}^{\otimes 2}, \tag{8}$$

only using data available at the $k$th interim analysis, where $a^{\otimes 2} = aa'$,

$$\hat{P}\left(X_{ij}^{(s)} \geq t, \ E_{ij} < t^{(s)}\right) = n_j^{-1} \sum_{i=1}^{n_j} I(X_{ij}^{(s)} \geq t, \ E_{ij} < t^{(s)}), \tag{9}$$

$$\widehat{M}_{ij}^{(s)}(t) = I\left(E_{ij} < t^{(s)}\right)\left\{I\left(X_{ij}^{(s)} \leq t\right)\delta_{ij}^{(s)} - \int_0^t I\left(X_{ij}^{(s)} \geq s\right) d\widehat{\Lambda}_j^{(k)}(s)ds\right\}, \tag{10}$$

and $\widehat{\Lambda}_j^{(k)}(s) = -\log\left(\hat{S}_j^{(k)}(t)\right)$. Note that this estimator of $\Sigma_k(\pi_1)$ is consistent under both null and alternative hypotheses. Another estimator of $\Sigma_k(\pi_1)$ could be constructed by replacing $\hat{S}_j^{(k)}(\cdot)$ by $\bar{S}^{(k)}(\cdot)$, the KM estimator based on the pooled samples as in Murray and Tsiatis (1999): it is more precise under the null, but not consistent under alternative. The rejection region can therefore be set consecutively. Specifically,

- at the 1st interim analysis, we will reject the null hypothesis if $\sqrt{n}|\widehat{D}_1(\tau^{(1)})| \geq c_1$, where $P(|Z_{11}| \geq c_1) = \alpha_1$, and $Z_{11} \sim N(0, \hat{\Sigma}_1(\pi_1))$;
- at the 2nd interim analysis, we will reject the null hypothesis if $\sqrt{n}|\widehat{D}_2(\tau^{(2)})| \geq c_2$, where $P(|Z_{22}| \geq c_2, |Z_{21}| < c_1) = \alpha_2$, and $(Z_{21}, Z_{22})' \sim N(0, \quad \hat{\Sigma}_2(\pi_1))$;

...

- at the kth interim analysis, we will reject the null hypothesis if $\sqrt{n}|\widehat{D}_k(\tau^{(k)})| \geq c_k$, where $P(|Z_{kk}| \geq c_k, |Z_{k(k-1)}| < c_{k-1}, \cdots |Z_{k1}| < c_1) = \alpha_k$, and $(Z_{k1}, Z_{k2}, \cdots, Z_{kk})' \sim N(0, \quad \hat{\Sigma}_k(\pi_1))$;

...

- at the final analysis, we will reject the null if $\sqrt{n}|\widehat{D}_F(\tau_F)| \geq c_F = c_{K+1}$, where $P(|Z_{(K+1)(K+1)}| \geq c_F, |Z_{(K+1)K}| < c_K, \cdots |Z_{(K+1)1}| < c_1) = \alpha_{K+1}$, and $(Z_{(K+1)1}, Z_{(K+1)2}, \cdots, Z_{(K+1)(K+1)})' \sim N(0, \quad \hat{\Sigma}_{K+1}(\pi_1))$.

Note that although $\hat{\Sigma}_k(\pi_1)$ should approximately be the upper-left $k \times k$ submatrix of $\hat{\Sigma}_{K+1}(\pi_1)$, with latter being a more accurate estimator of the underlying variance-covariance matrix based on more data, only the former is available at the $k$th interim analysis to determine the corresponding rejection region.

The commonly used group-sequential design based on the log-rank test is event driven and the variance-covariance matrix of the test statistics under the null can be derived based on the

number of events. Therefore, the rejection region at each interim analysis can be determined in advance, if the interim analysis is conducted when the planned number of events is reached. On the other hand, the rejection region at each interim analysis based on the RMST needs to be derived based on data available at the corresponding interim analysis. This is due to the more complex structure of the variance-covariance matrix of the test statistics. To estimate the power or plan the sample size for a new study, we need to consider a specific alternative: $S_1(t)$ vs $S_0(t)$, $t \in [0, \tau_F]$. Note that under a general alternative,

$$\sqrt{n}\begin{pmatrix} \widehat{D}_1\left(\tau^{(1)}\right) - D\left(\tau^{(1)}\right) \\ \cdots \\ \widehat{D}_K\left(\tau^{(K)}\right) - D\left(\tau^{(K)}\right) \\ \widehat{D}_{K+1}\left(\tau^{(K+1)}\right) - D\left(\tau^{(K+1)}\right) \end{pmatrix} \sim N(0, \quad \Sigma_{K+1}(\pi_1)) \tag{11}$$

for large $n$. The power of this group-sequential design at the $k$th interim analysis is

$$1 - \beta_k = P\left(\left|Z_k + \sqrt{n}D\left(\tau^{(k)}\right)\right| \geq c_k, \left|Z_s + \sqrt{n}D\left(\tau^{(s)}\right)\right| < c_s, s = 1, \cdots, k-1\right), \tag{12}$$

and the overall power is

$$\sum_{k=1}^{K+1}(1 - \beta_k), \tag{13}$$

where $(Z_1, \cdots, Z_{K+1})' \sim N(0, \Sigma_{K+1}(\pi_1))$.

In order to calculate the power, we need to determine the variance-covariance matrix $\Sigma_{K+1}(\pi_1)$ first. To this end, we first specify the joint distribution of $(E_j, F_j), j = 0, 1$ and values of $\left(t_k, \tau^{(k)}\right)$, $k = 1, \cdots, K+1$. Under those assumptions,

$$\Sigma_{K+1}(\pi_1) = \sum_{j=0}^{1} \frac{1}{\pi_j} E\begin{pmatrix} \int_0^{\tau^{(1)}} \frac{\int_t^{\tau^{(1)}} S_j(s)ds}{P(T_j \wedge F_j \wedge (t^{(1)} - E_j) \geq t)} dM_{ij}^{(1)}(t) \\ \cdots \\ \cdots \\ \int_0^{\tau_F} \frac{\int_t^{\tau_F} S_j(s)ds}{P(T_j \wedge F_j \wedge (\tau^{(K+1)} - E_j) \geq t)} dM_{ij}^{(K+1)}(t) \end{pmatrix}^{\otimes 2}, \tag{14}$$

which can be estimated either by direct integration or more conveniently by Monte-Carlo simulation. To be specific, one may first simulate a data set of a large sample size $M$, $\left\{\left(E_{ij}^*, F_{ij}^*, T_{ij}^*\right), i = 1, \cdots, M\pi_j, \ j = 0, 1\right\}$, and then may calculate the centered RMST estimates,

$$I = \sqrt{M}\begin{pmatrix} D_1^*\left(\tau^{(1)}\right) - D\left(\tau^{(1)}\right) \\ \cdots \\ D_K^*\left(\tau^{(K)}\right) - D\left(\tau^{(K)}\right) \\ D_{K+1}^*\left(\tau^{(K+1)}\right) - D\left(\tau^{(K+1)}\right) \end{pmatrix}. \quad (15)$$

After repeating this process a large number of, say, $B$, times, $\hat{\Sigma}_{K+1}(\pi_1)$ can be approximated by

$$\frac{1}{B}\sum_{b=1}^{B} I_b^{\otimes 2}, \quad (16)$$

Once an estimator of $\Sigma_{K+1}(\pi_1)$ is obtained, we can determine $\{c_1, \cdots, c_{K+1}\}$ based on the prespecified alpha-spending plan, i.e., $\{\alpha_1, \alpha_2, \cdots, \alpha_{K+1}\}$, by solving a system of equations:

$$P(|Z_1| \geq c_1) = \alpha_1,$$

$$P(|Z_2| \geq c_2, |Z_1| < c_1) = \alpha_2,$$

$$\cdots$$

$$P(|Z_{K+1}| \geq c_{K+1}, |Z_K| < c_K, \cdots, |Z_1| < c_1) = \alpha_{K+1}. \quad (17)$$

With an approximation to $\Sigma_{K+1}(\pi_1)$, the variance-covariance matrix of $(Z_1, \cdots, Z_{K+1})'$, the cut-off values $\{c_1, \cdots, c_{K+1}\}$, and the underlying differences in RMST $\left\{D\left(\tau^{(1)}\right), \cdots, D\left(\tau^{(K+1)}\right)\right\}$, one may calculate the power accordingly for any given sample size $n$. Note that $\Sigma_{K+1}(\pi_1)$ under the alternative is different from that under the null and we propose to set the rejection region according to $\Sigma_{K+1}(\pi_1)$ under the general alternative to facilitate the power calculation. Despite the fact that there is no general independent increment structure among test statistics in interim and final analyses, the power can still be estimated under any given alternatives using the proposed method. Lastly, it is important to note that the actual cut-off value at a particular interim analysis only becomes identifiable after the data up to that interim analysis become available. The actual cut-off value may or may not be close to that from the sample size calculation.

**Remarks 1.** If a particular level of power, e.g., 80%, is desired, one may estimate the sample size $n$ by solving the equation

$$\sum_{k=1}^{K+1} P\left(\left|Z_k + \sqrt{n}D\left(\tau^{(k)}\right)\right| \geq c_k, \left|Z_s + \sqrt{n}D\left(\tau^{(s)}\right)\right| < c_s, s = 1, \cdots, k-1\right) = 80\% \quad (18)$$

in terms of $n$. The expected sample size can be calculated based on the stopping probability at the $k$th interim analysis:

$$n\sum_{k=1}^{K+1} P\left(\left|Z_k + \sqrt{n}D\left(\tau^{(k)}\right)\right| \geq c_k, \left|Z_s + \sqrt{n}D\left(\tau^{(s)}\right)\right| < c_s, s = 1, \cdots, k-1\right) P\left(E < t^{(k)}\right), \quad (19)$$

assuming the distributions of enrollment time in two arms are identical.

**Remarks 2.** The futility stopping can be introduced easily. For example, at the $k$th interim analysis, we may decide to

(1) stop the study for futility if $\sqrt{n}\widehat{D}\left(\tau^{(k)}\right) < b_k$,

(2) stop the study for efficacy if $\sqrt{n}\widehat{D}\left(\tau^{(k)}\right) > c_k$ and

(3) continue the study, otherwise.

In such a case, $(b_k, c_k)$ can be selected to satisfy the condition that

$$P\left(Z_{kk} > c_k, Z_{k(k-1)} \in [b_{k-1}, c_{k-1}], \cdots Z_{k1} \in [b_1, c_1]\right) = \alpha_k$$

$$P\left(Z_{kk} + \sqrt{n}D\left(\tau^{(k)}\right) < b_k, Z_{k(k-1)} + \sqrt{n}D\left(\tau^{(k-1)}\right) \in [b_{k-1}, c_{k-1}], \cdots, Z_{k1} + \sqrt{n}D(\tau^{(1)}) \in [b_1, c_1]\right) = \gamma_k$$

where $\{\gamma_1, \cdots, \gamma_K\}$ are the prespecified probabilities controlling early stopping due to futility under the alternative. The introduction of the futility boundary would affect the power negatively. The power at the $k$th interim analysis can be calculated as

$$P\left(Z_{kk} + \sqrt{n}D\left(\tau^{(k)}\right) > c_k, Z_{k(k-1)} + \sqrt{n}D\left(\tau^{(k-1)}\right) \in [b_{k-1}, c_{k-1}], \cdots Z_{k1} + \sqrt{n}D(\tau^{(1)}) \in [b_1, c_1]\right).$$

One sided group sequential study can be designed similarly by dropping the futility boundary.

**Remarks 3.** If the time of interim analysis is relatively close to the final analysis, then it is possible to let $\tau^{(1)} = \tau^{(2)} = \cdots = \tau^{(K+1)}$, i.e., all analyses share the same parameter of interest. In such a case, one may construct a 95% CI for this parameter. The procedure is similar to that for constructing the 95% CI for HR in a group-sequential study (Jennison and Turnbull, 1984). However, the truncation time points are not required to be the same in general.

**A Numerical Example**

In this section, we first present a numerical example of designing a group sequential study. To this end, we consider a study designed to demonstrate the superiority of bariatric surgery approach (treatment arm) in comparison with the standard cardiac ablation procedure (control arm). The endpoint of interest is the duration free from atrial fibrillation (AF) after the treatment. Assuming the maximum duration of the study is 4 years with the first 2.5 years for enrollment and a minimum of follow-up of 1.5 years for all patients. The randomization ratio is one to one. Since the post-surgery clinical treatments will be hard to control, our clinical investigators decided to evaluate the surgical treatment effect to be best evaluated within a time window of 1-2 years after the surgery. The inference based on RMST is an appealing choice. Furthermore, there are two types of patients: those with paroxysmal AF and those with persistent AF. Historical data suggest the AF-free rate at 1 year after standard treatment is 70% for the former and 55% for the latter. Therefore, we assume that the survival distribution in the entire study cohort is a mixture distribution of two exponential distributions with the annual incidence rate of 0.3567 and 0.5978, respectively. The mixing proportion of 40% vs 60% is also based on historical data. Assuming a 20% proportionally increasing in AF-free rate at the end of one year is clinically significant, i.e., the treatment group should achieve an AF-free rate of 84% and 66% by the end of one year, respectively, for those with paroxysmal and persistent AF. The corresponding annual incidence rates for the two prognostic groups should be 0.1744 (paroxysmal AF) and 0.4155 (persistent AF), respectively. Therefore,

$$S_0(t) = (0.40\, e^{-0.3567t} + 0.60\, e^{-0.5978t}) \tag{20}$$

versus

$$S_1(t) = (0.40\, e^{-0.1744t} + 0.60\, e^{-0.4155t}) \tag{21}$$

Because of the mixture population, the HR between treatment and control arms is

$$h(t) = \left[\frac{0.0697e^{-01744t} + 0.2493e^{-0.4155t}}{0.40\ e^{-0.1744t} + 0.60\ e^{-0.4155t}}\right] \Big/ \left[\frac{0.1427\ e^{-0.3567t} + 0.3587e^{-0.5978t}}{0.40\ e^{-0.3567t} + 0.60\ e^{-0.5978t}}\right]$$

The PH assumption is clearly violated. Therefore, we propose to use the RMST up to 1.5 years instead of the HR as the measure of the treatment effect, i.e., the estimand of primary interest.

The study design will be under the following assumptions. First, the enrollment of patients is uniformly over the first 2.5 year. We assume that the annual dropout rate is 15% in both arms following an exponential distribution. To this end, the test statistics is the estimated difference in RMST.

In this sequential design, the first interim analysis is at year 2 after study initiation with $\tau^{(1)} = 1.5$ years, and the final analysis is at year 4 with $\tau_F = \tau^{(2)} = 1.5$ years. The underlying difference in RMST at both interim and final analyses is 0.139 years with the RMST in two arms being 1.059 and 1.198 years, respectively. We plan to have an overall one-sided type I error rate of 2.5% with 0.5% error rate spent at the interim analysis. We will stop for efficacy but not for the futility.

Using the specified distributions in Equations (20) and (21) and formulas in Equations (11-19), we may estimate the variance-covariance matrix $\hat{\Sigma}_2(0.5)$ and the required sample size. However, it is easier to employ the proposed Monte-Carlo method to achieve the same objective. The resulting covariance matrix estimate is

$$\hat{\Sigma}_2(0.5) = \begin{pmatrix} 1.652 & 1.001 \\ 1.001 & 1.024 \end{pmatrix}.$$

We can then determine the critical value for interim and final analyses: $c_1 = 2.5758 \times \sqrt{1.652}$ and $c_2 = 1.9917 \times \sqrt{1.024}$, i.e., we will reject the null, if $\sqrt{n}\hat{D}_1\left(\tau^{(1)}\right) \geq c_1$ at the interim or $\sqrt{n}\hat{D}_2\left(\tau^{(2)}\right) \geq c_2$ at the final analysis. We may calculate the power for a range of sample sizes (Figure 1). The corresponding sample size for 80% statistical power is 212 per arm. Under this

hypothesized alternative, the stopping probability at interim is 36.2% based on equation (19) and the expected total sample size is 197 per arm. To confirm the power estimation, we also conduct a simulation study to estimate the power with Monte-Carlo method. For each given sample size, we repeatedly generate survival data from the hypothesized alternative and conduct the group-sequential analysis with estimated rejection regions. Based on the testing result from 4,000 simulated datasets, we record the empirical power, which is consistent with the analytical result (Figure 1). Specifically, the empirical power corresponding to $n$ =212 per arm is 80.5% (95% CI: 79.3%-81.7%), supporting the estimated sample size. As a reference, the sample size for binary proportion at year 1.5 is 235 per arm after accounting for 22% dropout due to censoring before the end of 1.5 years.

As an illustration, we consider a more complex design with two interim looks at year 2 and year 3. The truncation time point at the interim and final analyses is 1.5, 2.5 and 3.0 years, and the underlying RMST difference is 0.139, 0.303 and 0.390 years, respectively. With Monte-Carlo simulations, the estimated variance-covariance matrix is

$$\hat{\Sigma}_3(0.5) = \begin{pmatrix} 1.651 & 1.821 & 1.959 \\ 1.821 & 4.008 & 4.134 \\ 1.959 & 4.134 & 5.184 \end{pmatrix}.$$

Suppose we plan to spend 0.4%, 0.6%, and 1.5% one-sided type one error at two interim and final analyses, then the critical values of the rejection regions are $2.652 \times \sqrt{1.651}$, $2.445 \times \sqrt{4.008}$, and $2.018 \times \sqrt{5.184}$ at the two interim and final analyses, respectively. Note that the estimated variance of $\sqrt{n}\hat{D}(1.5)$ at the first interim analysis changes slightly (from 1.652 to 1.651) due to Monte Carlo variations. The required sample size for 80% power reduces to 138 patients per arm, reflecting the fact that more information due to prolonged truncation time points for RMST is used in the comparison. In addition, we repeatedly generate survival times for 138 patients per arm from the specified model and conduct the proposed group-sequential

test using RMST. Based on results from 4,000 simulated datasets, the empirical power is 81.1% (95% CI: 79.9%-82.3%). The probability of rejecting the null is 21.9% at the first interim analysis and 34.0% at the second interim analysis.

**A Simulation Study**

We have also conducted a simulation study to examine the empirical property of the proposed methods. Specifically, we have considered the following simple setting that is similar to those in Mehta, et al. (2018). We assume that the sample size for each treatment groups, $n_1 = n_0 = 300$. Patients are uniformly enrolled in the first 12 months, i.e., $E_{ij} \sim U(0, 12)$ months. Censoring time $F_{ij}$ follows exponential distribution with an annual hazard rate of 0.014. We assume that the study duration is 30 months. For control arm, the survival function follows an exponential distribution with $S_0(t) = \exp(-0.1t)$ We choose the following four scenarios of survival distribution for the treatment arm:

- The first scenario is the null case, where $S_1(t) = S_0(t)$;
- The second scenario is for constant PH benefit, where $S_1(t) = \exp(-0.08\, t)$;
- The third scenario is for delayed treatment benefit, where

$$S_1(t) = \exp\{-0.1\, t \wedge 8 - 0.05(t-8)_+\}; \quad (26)$$

- The fourth scenario is for early treatment benefit over a short period, where

$$S_1(t) = \exp\{-0.05\, t \wedge 4 - 0.08\{(t-4)_+ \wedge 4\} - 0.12\,(t-8)_+\}; \quad (27)$$

- The fifth scenario is also for early treatment benefit over a long period, where

$$S_1(t) = \exp\{-0.07t \wedge 8 - 0.08\{(t-8)_+ \wedge 8\} - 0.1286\,(t-16)_+\}; \quad (28)$$

where $a_+ = \max\{0, a\}$. There is only one interim analysis at time $t_1$ and the type one error at the interim analysis is a function of timing of the interim analysis:

$$\alpha_1 = \frac{\alpha\left(1-\exp\left(\frac{5t_1}{30}\right)\right)}{1-\exp(5)}, \quad (29)$$

where $t_1 = 15,\ 18, 21, 24$ and 27 months. Specifically, corresponding alpha spent at the interim analysis is 0.38%, 0.65%, 1.1%, 1.8% and 3.0%, respectively, when the total type one error $\alpha = 0.05$. The RMST for interim and final analyses are up to $\tau_1 = t_1 - 1$ and $\tau_2 = 30 - 1 = 29$ months for interim and final analyses, respectively, to ensure identifiability. The survival curves of four alternatives are given in Figure 2. We have compared the group-sequential design based on the difference in RMST with two alternative methods: the log-rank test and the Max-Combo test for two-group comparisons (Lin et al, 2018). In the Max-Combo test, the test at the kth interim analysis is based on the test statistics

$$M = \max\left(Z_k^{00}, Z_k^{10}, Z_k^{01}\right), \tag{30}$$

where

$$Z_k^{ab} = s_k^{ab} / \hat{\sigma}_k^{ab}\ , \tag{31}$$

$s_k^{ab}$ is the weighted logrank test statistics with the weight $\bar{S}(t)^a \left(1 - \bar{S}^{(k)}(t)\right)^b$ at the failure time $t$ and $\hat{\sigma}_k^{ab}$ is the estimated standard deviation of the weighted logrank test statistics under the null. Here $\bar{S}^{(k)}(t)$ is the KM estimator of the survival function based on data pooled from two arms at the $k$th interim analysis. For each simulated data set, we have conducted group sequential analyses based on logrank test, Max-Combo test, and the test via the difference in RMST. We also have conducted the corresponding analysis without any interim analysis. We repeated the analyses in 40,000 simulated datasets and therefore the Monte-Carlo standard error for estimating 5% type one error and 80% power is 0.1% and 0.2%, respectively. The results are summarized in Table 1.

Based on the simulation results, the type one errors of all tests are well preserved. The statistical power of the RMST is very close to but slightly lower than the power of logrank test under the PH alternative. The power of RMST is slightly higher than that of Max-Combo test in this case. In the third setting with delayed treatment benefit, RMST-based test is less powerful

than other two competitors. It is anticipated, since RMST is not sensitive to the delayed treatment benefit (Tian et al. 2018). In the fourth setting with early treatment benefit, the RMST-based test is more powerful than logrank test but still inferior to the Max-Comb test. In this setting, it interesting to note that the group sequential tests are substantially more powerful than the corresponding single test at the final analysis. This is mainly due to the fact that the power of interim analysis was boosted by early treatment benefit. In the fifth setting with early treatment benefit over a longer period than that in the fourth setting, the RMST-based test is slightly more powerful than either logrank test or the Max-Comb test in the group-sequential analysis. It is also interesting to note that in cases with PH or delayed treatment benefit (settings 2 and 3), the group sequential design doesn't lose much power in comparison with a single final analysis. Overall, the Max-Combo test is the most powerful or nearly most powerful method across different scenarios investigated here. When it is not optimal, its power is not much lower than the better alternatives. Its main limitation is that there is no simple interpretable estimand quantifying the treatment effect associated with the method due to its adaptive nature.

**Discussions**

In this paper, we have illustrated the method of designing a group-sequential randomized clinical trial based on the difference in RMST. Compared with HR-based method, the proposed method is more interpretable and less dependent on stringent model assumptions, such as PH assumption. One concern is that the decision at the interim analysis based on RMST may be premature regardless of its statistical significance level, since the time window of the RMST is too narrow. On the other hand, the HR-based method has the same limitation, which is caused by the fact that the maximum follow-up time at the interim analysis may be too short and all statistical inferences at the interim analysis including those for HR is restricted within this shorter time span. Therefore, it is crucial to choose the timing of the interim analysis, such that the follow-up time is adequately long to allow meaningful clinical decision based on the analysis

result. $\tau^{(k)}$, the truncation time point in the RMST, should be no greater than $t_k$. Otherwise, the difference in RMST is not estimable nonparametrically. In general, we should let $\tau^{(k)}$ as close to $t_k$ as possible to avoid unnecessary loss of information. In Tian et al (2020), we have shown that $\tau^{(k)}$ can be chosen as the minimum of the largest follow-up time in two arms despite the fact that such a choice is data-dependent.

In practice, we often have different milestones in clinical trials. For examples, phase III oncology trial uses overall survival as the primary endpoint but the progress-free survival time as a secondary endpoint. Therefore, the interim analysis $\tau^{(1)}$ can be selected based on the milestone time for progress free survival. Another example is the ESCAPe trial (NCT01283009, www.clinicaltrials.gov), where the primary endpoint is 60-day mortality and secondary endpoints include the mortality at ICU discharge, etc. If the ESCAPe trial uses the RSMT approach, the primary endpoint could be 60-day RMST and the $\tau^{(1)}$ can be 14 days mortality to measure the mortality at ICU discharge. For a superiority trial with the null hypothesis of no difference, a smaller but clinically meaningful $\tau^{(1)}$ ($< \tau^{(F)}$) can help to reject the null hypothesis of no difference sooner for large efficacy or potential detrimental effect to patients (two-sided interim analysis) but will not be proper for futility. Furthermore, for a non-inferiority analysis, we should choose $\tau^{(1)} = \tau_F$, because otherwise, non-inferior at time $\tau^{(1)}$ cannot imply the non-inferiority at time $\tau_F$.

Different test methods are sensitive to different alternatives and the RMST-based method is not the most powerful test to detect crossing survival functions (Tian et al. 2018). In such a case, the Max-Combo test can be more powerful. However, one may argue that detecting crossing in survival functions is merely the first step and a less important step, since in order to claim superiority of one treatment versus another, one needs to consider the difficult trade-off

between short term and long-term survival benefits in such a case. Another important alternative is delayed treatment benefit as in our simulation study. The RMST-based method can still be used. However, in that case, it is even more imperative to make sure that $t_k$ and also $\tau^{(k)}$ are big enough so that the study will not stop prematurely for missing the long-term treatment benefit. A more sensitive measure is the difference in RMST within a prespecified time window $[\tau_1, \tau_2]$:

$$\int_{\tau_1}^{\tau_2} \{S_1(t) - S_0(t)\} dt.$$

One may expect substantial gain in power, if the time window is chosen appropriately (Horiguchi et al. 2018). The group-sequential design based on this generalized RMST is also straightforward. As we discussed in the example, an important potential source of PH assumption violation is the treatment effect heterogeneity. There is a possibility of boosting the power of the overall comparisons and present treatment effect for different subgroups of patients at the same time. For example, Mehrotra et al. (2012) have proposed to first estimate stratum-specific treatment effects and then combine the resulting estimates in an overall comparison. The success of such approaches relies on discovering the patient population structure reflecting the treatment effect heterogeneity.

The Cross-pharma Working Group (Roychoudhury et al., 2019) recommended the use of Max-Combo test based on the weighted log-rank test for statistical inference of survival endpoints that fail the PH assumption. A Max-Combo test can also be constructed based on the weighted K-M tests (Shen and Cai, 2001), in which the simple RSMT test is one of the components of the test. Both these tests can be more powerful, but they do not directly associate with an estimand of the clinical interest, namely the HR, the difference in RMST, or any other metric for the treatment effect. It is a better practice to choose a statistical inference method containing a hypothesis testing procedure consistent with the estimand of interest.

R-code for designing Group-Sequential study based on RMST can be found in the

Supplementary Materials of the paper and at https://web.stanford.edu/~lutian/Software.HTML

**Acknowledgement**

YL is partially supported by grants 4P30CA124435 and 1UL1TR003142 from National Institutes

of Health (USA). LT is partially supported by R01 HL089778 from National Heart, Lung and

Blood Institute and and 1UL1TR003142 from National Institutes of Health (USA).

**References**

Chang, H. and McKeague, I. W. (2019) Nonparametric testing for multiple survival functions with noninferiority margins. The Annals of Statistics 47(1): 205-232.

Fleming, T. R., Harrington, D. P., & O'Brien, P. C. (1984). Designs for group sequential tests. *Controlled clinical trials*, *5*(4): 348-361.

Garon, E. B., Rizvi N. A., Hui, R., et al. (2015) Pembrolizumab for the treatment of non-small-cell lung cancer. The New England Journal of Medicine 372(21):2018-2028.

Hasegawa, Takahiro, Saori Misawa, Shintaro Nakagawa, Shinichi Tanaka, Takanori Tanase, Hiroyuki Ugai, Akira Wakana et al. "Restricted mean survival time as a summary measure of time-to-event outcome." *Pharmaceutical Statistics* (2020).

Horiguchi, M., Tian, L., Uno, H., Cheng, S., Kim, D. H., Schrag, D., & Wei, L. J. (2018). Quantification of long-term survival benefit in a comparative oncology clinical study. *JAMA oncology*, *4*(6): 881-882.

Jennison, C., & Turnbull, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, *5*(1): 33-45.

Karrison TG (2016). Versatile tests for comparing survival curves based on weighted log-rank statistics. Stata Journal, 16(3): 678-690

Lin, Ray S., Ji Lin, Satrajit Roychoudhury, Keaven M. Anderson, Tianle Hu, Bo Huang, Larry F. Leon et al. (2019) "Alternative Analysis Methods for Time to Event Endpoints under Non-proportional Hazards: A Comparative Analysis." *Statistics in Biopharmaceutical Research* (in press).

Mehta, C., & Ghosh, P. (2018) Group sequential designs for non-proportional hazards alternatives. The 2018 FDA & Industry Workshop (https://webcache.googleusercontent.com/search?q=cache:Xjcmb25jg5IJ:https://ww2.amstat.

org/meetings/biopharmworkshop/2018/onlineprogram/ViewPresentation.cfm%3Ffile%3D300802.pdf+&cd=2&hl=en&ct=clnk&gl=us)

Mehrotra, D. V., Su, S. C., & Li, X. (2012). An efficient alternative to the stratified cox model analysis. *Statistics in medicine*, *31*(17), 1849-1856.

Murray, S., & Tsiatis, A. A. (1999). Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics*, *55*(4): 1085-1092.

Robert, C., Schachter, J., Long, G., V., et al. (2015). Pembrolizumab versus Ipilimumab in advanced melanoma. The New England Journal of Medicine 372(26):2521-2532.

Royston, P., & Parmar, M. K. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology*, *13*(1): 152.

Roychoudhury, S., Anderson, K., Ye, J., & Mukhopadhyay, P., (2019). Robust design and analysis of clinical trials with non-proportional hazards: a straw man guidance from a cross-pharma working group.

Tian, L., Fu, H., Ruberg, S. J., Uno, H., & Wei, L. J. (2018). Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, *74*(2): 694-702.

Tian, L., Jin, H., Uno, H., Lu, Y., Huang, B., Anderson, KM. and Wei, L.J. (2020) On the empirical choice of the time window for restricted mean survival time. Biometrics (in press)

Trinquart, L., Jacot, J., Conner, S. C., & Porcher, R. (2016). Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, *34*(15): 1813-1819.

Trinquart, L., Bill-Axelson, A., & Rider, J. R. (2019). Restricted Mean Survival Times to Improve Communication of Evidence from Cancer Randomized Trials and Observational Studies. 76(2):137-139.

Uno, H., Claggett, B., Tian, L. et al. (2014) "Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis." *Journal of Clinical Oncology* 32(22): 2380-2385.

Uno, H., Wittes, J., Fu, H. et al. (2015)[a] "Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies." *Annals of Internal Medicine* 163(2): 127-134.

Uno, H., Tian, L., Claggett, B., and Wei, L. J.. (2015)[b] "A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves." *Statistics in medicine* 34(28): 3680-3695.

Shen, Y. and Cai, J. (2001). Maximum of the Weighted Kaplan-Meier Tests with Application to Cancer Prevention and Screening Trials. Biometrics, Vol. 57(3): 837-843

Table 1. Simulation Results for test based on log-rank test, Max-Combo test and RMST with and without interim analysis (40,000 simulations).

| Scenario | IA (wks) | Single Look | | | 2-Stage | | | Prob Early Stopping | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Logrank | Max-Comb | RMST | Logrank | Max-Comb | RMST | Logrank | Max-Comb | RMST |
| NULL | 15 | 5.0% | 4.8% | 5.0% | 4.8% | 4.9% | 5.1% | 1.1% | 1.1% | 1.1% |
| | 18 | | | | 4.8% | 4.8% | 5.2% | 1.6% | 1.5% | 1.6% |
| | 21 | | | | 4.9% | 4.8% | 5.2% | 2.2% | 2.1% | 2.3% |
| | 24 | | | | 4.9% | 4.9% | 5.2% | 2.9% | 2.8% | 3.0% |
| | 27 | | | | 4.9% | 4.8% | 5.2% | 3.8% | 3.7% | 3.9% |
| PH | 15 | 71.8% | 68.0% | 71.5% | 70.7% | 67.2% | 71.4% | 28.3% | 25.3% | 29.6% |
| | 18 | | | | 70.8% | 67.3% | 71.5% | 39.8% | 36.7% | 40.8% |
| | 21 | | | | 70.8% | 67.3% | 71.6% | 49.9% | 46.5% | 50.9% |
| | 24 | | | | 70.9% | 67.3% | 71.7% | 58.4% | 54.7% | 59.1% |
| | 27 | | | | 71.0% | 67.4% | 71.7% | 65.4% | 61.8% | 66.1% |
| Late Benefit | 15 | 75.7% | 94.5% | 69.1% | 74.0% | 94.2% | 69.3% | 2.6% | 4.8% | 3.1% |
| | 18 | | | | 74.0% | 94.1% | 69.4% | 8.7% | 18.9% | 10.0% |
| | 21 | | | | 73.8% | 94.0% | 69.4% | 24.6% | 48.9% | 22.7% |
| | 24 | | | | 73.7% | 94.1% | 69.4% | 44.7% | 74.1% | 39.3% |
| | 27 | | | | 73.6% | 94.2% | 69.3% | 62.2% | 88.0% | 55.9% |
| Early Benefit (I) | 15 | 48.0% | 88.5% | 65.0% | 78.3% | 90.6% | 81.7% | 76.9% | 86.0% | 78.4% |
| | 18 | | | | 69.9% | 89.1% | 78.0% | 68.3% | 85.2% | 75.4% |
| | 21 | | | | 60.2% | 87.9% | 74.1% | 57.8% | 83.8% | 71.7% |
| | 24 | | | | 53.9% | 87.6% | 70.4% | 51.0% | 84.3% | 68.1% |
| | 27 | | | | 50.2% | 87.6% | 67.6% | 48.0% | 86.1% | 65.9% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Early Benefit (II) | 15 | 72.4% | 81.8% | 80.7% | 78.3% | 81.9% | 82.9% | 65.3% | 61.7% | 66.4% |
| | 18 | | | | 80.8% | 82.3% | 84.1% | 75.2% | 72.6% | 76.7% |
| | 21 | | | | 80.7% | 82.0% | 84.2% | 77.8% | 76.8% | 80.1% |
| | 24 | | | | 78.9% | 81.6% | 83.4% | 77.1% | 78.9% | 80.7% |
| | 27 | | | | 76.2% | 81.3% | 82.3% | 75.0% | 80.4% | 80.6% |

Figure 1. Sample Size and Power Estimation for the Numerical Example (solid thick line: analytic estimates of the power; circle and vertical bar: empirical power estimator and the 95% CI based on the simulations)

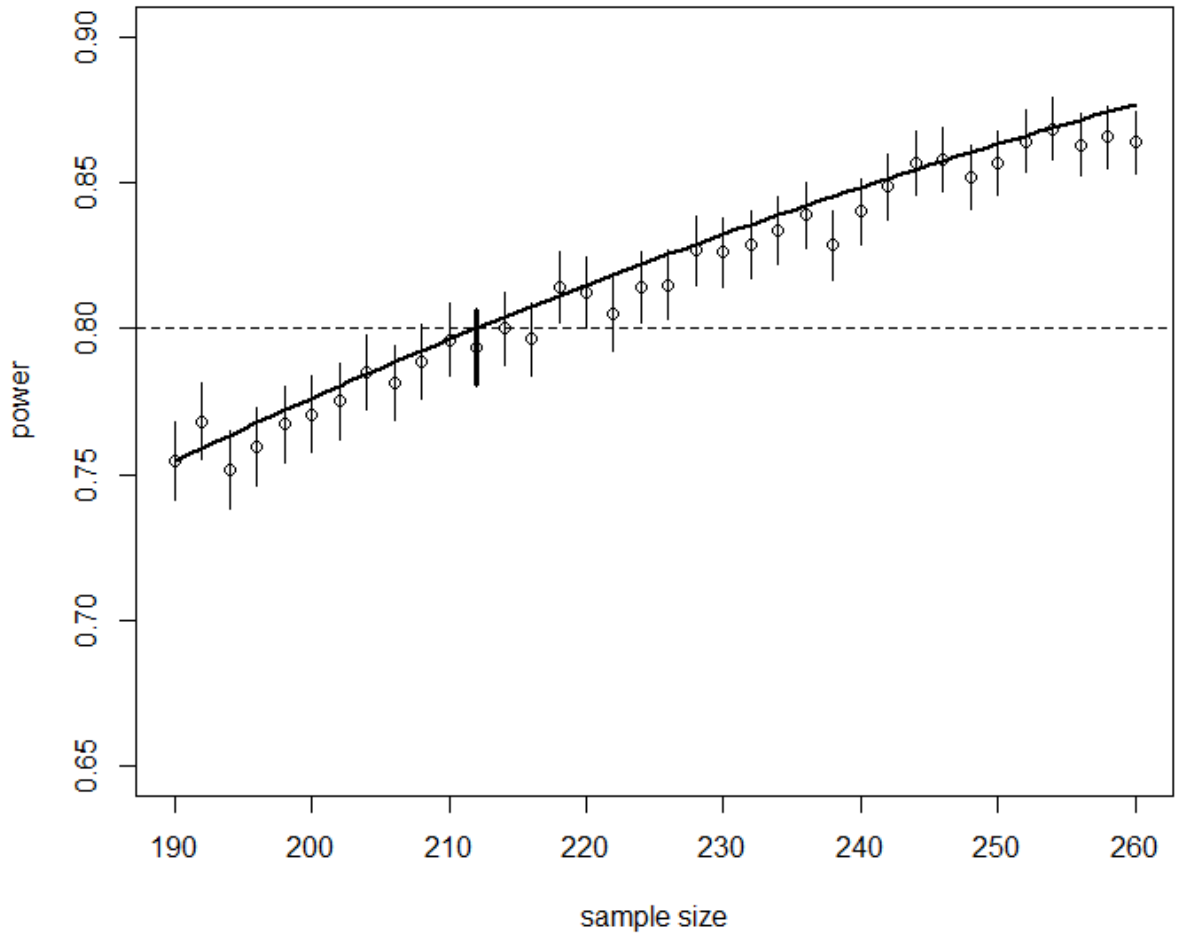

Figure 2. Survival curves of two arms used in the simulation study