Received: 2 April 2018 | Accepted: 2 November 2018
DOI: 10.1111/biom.13007

ORIGINAL ARTICLE

WILEY Biometrics

Nonparametric group sequential methods for evaluating survival benefit from multiple short-term follow-up windows

Meng Xia[1] | **Susan Murray**[1] | **Nabihah Tayob**[2]

[1]University of Michigan, Department of Biostatistics, Ann Arbor, Michigan 48109

[2]The University of Texas MD Anderson Cancer Center, Department of Biostatistics, Houston, Texas 77030

Correspondence
Meng Xia, University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109
Email: summerx@umich.edu

Abstract

This article takes a fresh look at group sequential methods applied to two-sample tests of censored survival data and proposes an alternative method of defining and evaluating treatment benefit. Our method re-purposes traditional censored event time data into a sequence of short-term outcomes taken from (potentially overlapping) follow-up windows. A new two-sample restricted means test based on this restructured follow-up data is proposed along with group sequential methods for its use in the clinical trial setting. This method compares favorably with existing methods for group sequential monitoring of time-to-event outcomes, including methods for monitoring the restricted means test and the logrank test. Our method performs particularly well in cases where there is a delayed treatment effect and/or a subset of cured patients. As part of developing group sequential methods for these analyses, we consider asymmetric error spending approaches that differentially limit the chances of stopping incorrectly for perceived efficacy versus perceived harm attributed to the investigational arm of the trial. Recommendations for how to choose proper group sequential stopping boundaries are given, with supporting simulations and an example from the AIDS Clinical Trial Group.

KEYWORDS

asymmetric error spending, group sequential methods, nonparametric test, survival analysis

1 | INTRODUCTION

Traditionally in the censored time-to-event setting, with or without group sequential monitoring, two-sample treatment comparisons are based on restricted mean event times or integrated weighted hazard differences estimated over many follow-up years (Mantel, 1963; Gehan, 1965; Mantel, 1966; Breslow, 1970; Peto and Peto, 1972; Prentice, 1978; Harrington and Fleming, 1982; Tsiatis, 1982; Pepe and Fleming, 1989; Li, 1999; Murray and Tsiatis, 1999). Investigators and biostatisticians alike hope that treatment differences will emerge throughout the trial, anticipating Kaplan–Meier curves that snake farther and farther apart as the end of follow-up draws near.

In this article we embrace the philosophy that for each patient in a clinical trial, short-term survival over repeated, overlapping intervals are observed, and that each of these has value in assessing treatment benefit. In short, time-to-event data can be reformulated as repeated short-term longitudinal outcomes subject to censoring, and then analyzed using methodology that takes into account both the censored nature of the data as well as the correlation between short-term events measured from the same individual.

Tayob and Murray (2016) followed this train of thought when they evaluated the behavior of an overall τ-restricted mean estimated from multiple, overlapping τ-length follow-up windows. Their overall estimated τ-restricted mean integrates area under an estimated survival curve, but instead of using time-to-event data in its original form, Tayob and

Murray estimate the curve from a massive censored longitudinal repeated measures dataset with multiple overlapping short-term outcomes taken from each individual's observed follow-up. Corresponding confidence intervals nonparametrically take into account the correlation between outcomes taken from the same individual. The choice of τ is typically taken from the context in which the method is applied. For instance, in pulmonary literature a 1-year restricted mean is common, and fairly stable over time as seen in Tayob and Murray. In scenarios where τ-restricted means are not stable over time, the overall τ-restricted mean is an estimate from a mixture distribution that results from combining information from overlapping follow-up windows.

In this article we propose a new two-sample test comparing τ-year restricted means estimated in the manner proposed by Tayob and Murray. As with existing two-sample tests, this test is valid under the null hypothesis of no treatment difference regardless of the distributions under study. We also develop group sequential methods for monitoring a clinical trial via the proposed statistic, along with graphics displaying the estimated overall years of life gained per τ time units when assigned the superior treatment.

Group sequential monitoring via nonparametric two-sample tests has a long and respected history in clinical trial design. Classic group sequential analysis literature gives stopping rules for statistically significant treatment benefit or harm (Pocock, 1977; O'Brien and Fleming, 1979). The most common approach for controlling type I error throughout a trial is to use error spending functions proposed by Lan and DeMets (1983), which allow for both symmetric and asymmetric stopping rules. Symmetric stopping rules imply that stopping early for statistically significant treatment differences have the same cost, whether benefit or harm is attributed to the experimental therapy. Asymmetric bounds are useful when consequences of stopping early are different according to the treatment difference that is emerging (Tsiatis, 1981; DeMets and Ware, 1982). Futility bounds have become increasingly popular as a mechanism for stopping a trial that is unlikely to end in a new treatment recommendation (Harrington, 2012; Friedman et al., 2015). These types of bounds also avoid the ethically uncomfortable scenario of trial termination only after statistical proof of increased mortality from the new treatment.

Our article proceeds with a description of notation in Section 2. In Section 3 we describe the proposed test statistic in the case where a single analysis is performed, with an extension to the group sequential setting given in Section 4; Derivations behind methods in Sections 3 and 4 are relegated to Supplementary Materials. In Section 4, we also review symmetric versus asymmetric stopping boundaries, with a modified recommendation for safety monitoring. Section 5 summarizes finite sampling behavior of our group sequential monitoring procedure in a variety of clinical trial settings. An example from the AIDS Clinical Trial Group is given in Section 6 and followed by discussion in Section 7.

2 | NOTATION

Our ultimate goal is to group sequentially monitor two-sample tests that compare estimates of τ-restricted mean lifetimes, $\mu_g(s,\tau)$, with group subscript, $g = 1, 2$, incorporating information from multiple, potentially overlapping, short-term follow-up windows of length τ. For simplicity, we first describe notation for the one-sample case, submerging the g subscript.

2.1 | Description of random variables

Suppose $i = 1, \ldots, N$ patients participate in a clinical trial. Patient-specific random variables are measured against two different time scales in the group sequential setting: calendar time, s, and study time, t. Study time, t, indexes time from a patient's clinical trial entry; length of life, length of follow-up and other clinical trial endpoints are described on this time-scale. Calendar time, s, indexes time from the initiation of the overall study; patient entry times and interim analysis times are described on this time scale.

In particular, study time indexed random variables include failure times, T_i and potential loss-to-follow-up times $V_i, i = 1, \ldots, N$. On the calendar time scale, we define random study entry times, E_i, for participant $i = 1, \ldots, N$, as well as interim analysis times, $s = s_1, s_2, \ldots$, which are (non-random) study design parameters. At interim analysis time s, $n(s) = \sum_{i=1}^{N} I(E_i \leq s)$ individuals have entered the trial with $n(s) = N$ for $s \geq max(E_1, \ldots, E_N)$. An individual's maximum follow-up time at analysis time s is administratively capped at $s - E_i$. Hence, the censoring random variable, $C_i(s) = min(V_i, s - E_i)$, for individual i can potentially change at each analysis time s, depending on the censoring mechanism. We assume that T_i is independent of $C_i(s), i = 1, \ldots, N$. For patients who have entered the trial, observed event times at analysis time s are $X_i(s) = min\left\{T_i, C_i(s)\right\}$, with corresponding failure indicator variables $\delta_i(s) = I\{T_i \leq C_i(s)\}$, $i = 1, \ldots, n(s)$.

Notation for residual lifetime random variables are needed to define short-term outcomes during several, potentially overlapping, τ-length follow-up windows of interest. The starting times of these follow-up windows, $t \in \{t_1, t_2, \ldots, t_b\}$, are non-random design parameters measured on the study time scale with $t_1 = 0$, and b indicating the total number of windows. We define the residual lifetime from study time t observed at analysis time s as $X_i(s,t) = (X_i(s) - t)I\{X_i(s) \geq t\}$ with corresponding failure indicator $\delta_i(s,t) = \delta_i(s)I\{X_i(s) \geq t\}$. A third time-scale metric, window time u, indexes time from the beginning of each follow-up window. We use the window time metric as a

common time-scale for residual lifetime random variables, $X_i(s,t_1), X_i(s,t_2), \ldots, X_i(s,t_b)$.

Figure 1 displays data for 3 example individuals, with random variables specific to subject A given in detail. Patient entry times E_A, E_B, E_C and interim analysis times s_1, s_2 are given on the calendar time scale. Death, loss to follow-up, administrative censoring and window start times are given on the study time scale. At the second interim analysis conducted on January 1, 2016, $n(s_2) = 3$ individuals have entered the study. Subject A contributes information from three windows starting at $t_1 = 0, t_2 = 6$ months and $t_3 = 12$ months. Observed residual lifetime and censoring indicator data pairs contributed by Subject A at the second analysis time are $(17,1), (11,1)$ and $(5,1)$. In terms of short-term follow-up windows of length $\tau = 12$ months, Subject A contributes uncensored information from three windows: in the first window, Subject A lives 12 of 12 months, in the second Subject A lives 11 of 12 months, and in the third window Subject A lives 5 of 12 months. Any test statistic incorporating multiple short-term outcomes taken from an individual as laid out in Figure 1 will need to account for potential correlation between these outcomes.

2.2 | Counting process notation and estimation

For an individual i who has entered the trial by interim analysis time s, $N_i(s,t,u) = I\{X_i(s,t) \le u, \delta_i(s,t) = 1\}$ and $Y_i(s,t,u) = I\{X_i(s,t) \ge u\}$ are the counting and at risk processes for the number of events occurring no later than window time u within the follow-up window starting at study time t. From Figure 1, consider Subject A's data at the 2^{nd} interim analysis time, s_2, from the follow-up window starting at $t_2 = 6$ months. Subject A's corresponding counting process data at window times $u = 11^-, 11$, and 11^+ months are $\{N_A(s_2,t_2,11^-) = 0, Y_A(s_2,t_2,11^-) = 1\}$, $\{N_A(s_2,t_2,11) = 1, Y_A(s_2,t_2,11) = 1\}$ and $\{N_A(s_2,t_2,11^+) = 1, Y_A(s_2,t_2,11^+) = 0\}$.

Let $N(s,t,u) = \sum_{i=1}^{n(s)} N_i(s,t,u)$ and $Y(s,t,u) = \sum_{i=1}^{n(s)} Y_i(s,t,u)$ represent processes summed across individuals entered by interim analysis time s. For individual i at interim analysis s, let $N_i(s,u) = \sum_{j=1}^{b} N_i(s,t_j,u)$ count the observed residual lifetime events across the b follow-up windows attributed to individual i that are seen prior to window time u; the corresponding at risk process is $Y_i(s,u) = \sum_{j=1}^{b} Y_i(s,t_j,u)$. $N_i(s,u)$ has the potential to count

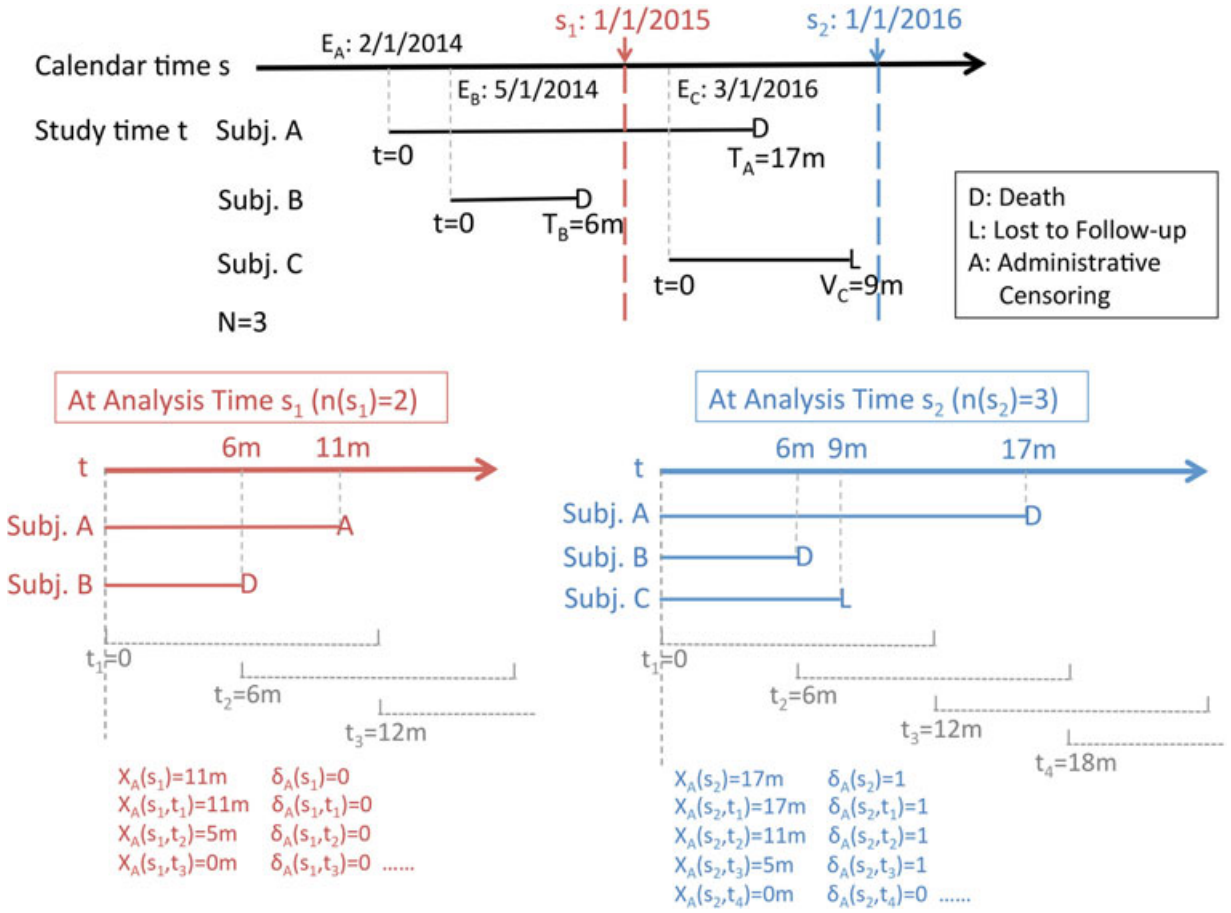


FIGURE 1 Notation for 3 example individuals, with random variables specific to Subject A given in detail. (This figure appears in color in the electronic version of this article.)

the same event more than once, since this event may be contained in more than one follow-up window. Likewise, $Y_i(s,u)$, includes at-risk processes from the same individual more than once from follow-up windows that overlap. Combining all information available at interim analysis time s regarding event and at-risk information for window time u we define $N(s,u) = \sum_{i=1}^{n(s)} N_i(s,u)$ and $Y(s,u) = \sum_{i=1}^{n(s)} Y_i(s,u)$.

At analysis time s, let hazard function $\lambda(s,t,u) = \lim_{\Delta u \to 0}[Pr\{u \leq X_i(s,t) \leq u + \Delta u, \delta_i(s,t) = 1 | X_i(s,t) \geq u\}/\Delta u]$ and

$$\lambda^W(s,u) = \frac{\sum_{j=1}^{b} \lambda(s,t_j,u) Pr\{X_i(s,t_j) \geq u\}}{\sum_{l=1}^{b} Pr\{X_i(s,t_l) \geq u\}}.$$

As in standard group sequential methods, we assume that analysis time does not affect the true event-time hazard, so that the superfluous s notation in $\lambda(s,t,u)$ can be dropped to become $\lambda(t,u)$. However, because $\lambda^W(s,u)$ corresponds to a mixture distribution of residual lifetimes contributed from individuals at time s, and is a function of $Pr\{X_i(s,t) \geq u\}$ that depends on follow-up, analysis time s can influence this term.

3 | TWO-SAMPLE TEST AT A SINGLE ANALYSIS TIME, s

In this section, we propose a two-sample test that compares average lifetime per τ follow-up years. The test is inspired by overall τ-restricted mean estimates developed by Tayob and Murray (2016) that incorporate information from repeated, overlapping follow-up windows of length τ, subject to censoring. Additional subscripts $g, g = 1,2$, indicate treatment group when used with notation from the last section; random variables from different treatment groups are assumed independent. We assume a single analysis at calendar time s.

For treatment g at analysis time s, following results from Tayob and Murray,

$$\hat{\mu}_g(s,\tau) = \int_0^{\tau} exp\left\{ - \int_0^{u_2} \frac{dN_g(s,u_1)}{Y_g(s,u_1)} \right\} du_2$$

consistently estimates $\mu_g(s,\tau) = \int_0^{\tau} exp\left\{ - \int_0^{u_2} \lambda_g^W(s,u_1) du_1 \right\} du_2$, the average lifetime per τ time units as measured from the mixture distribution of short-term, overlapping τ-length follow-up windows starting at times $t_1, \ldots, t_b$. Our proposed two-sample test becomes

$$\mathcal{T}(s) = \sqrt{\frac{n_1(s) n_2(s)}{n_1(s) + n_2(s)}} \{\hat{\mu}_1(s,\tau) - \hat{\mu}_2(s,\tau)\}.$$

Let $\hat{\pi}_g(s) = n_g(s)/\{n_1(s) + n_2(s)\}$, $g = 1,2$. As shown in Web Appendix A, under the null hypothesis of $\mu_1(s,\tau) = \mu_2(s,\tau)$, the asymptotic limiting distribution of $\mathcal{T}(s)$ has a mean 0 Normal distribution with variance that can be estimated by $\hat{\pi}_2(s)\hat{\sigma}_1^2(s) + \hat{\pi}_1(s)\hat{\sigma}_2^2(s)$, where $\hat{\sigma}_g^2(s) = \sum_{i=1}^{n_g(s)} [z_i\{\hat{\mu}_g(s,\tau)\} - \bar{z}\{\hat{\mu}_g(s,\tau)\}]^2/[n_g(s) - 1]$, with $z_i\{\hat{\mu}_g(s,\tau)\} = \sum_{j=1}^{b} z_{ij}\{\hat{\mu}_g(s,\tau)\}$; $\bar{z}\{\hat{\mu}_g(s,\tau)\} = \sum_{i=1}^{n_g(s)} z_i\{\hat{\mu}_g(s,\tau)\}/n_g(s)$ and

$$z_{ij}\{\hat{\mu}_g(s,\tau)\} = \int_0^{\tau} exp\left\{ - \int_0^{u_2} \frac{dN_g(s,u_1)}{Y_g(s,u_1)} \right\} \times \left\{ \int_0^{u_2} \frac{dN_{gi}(s,t_j,u_1) - Y_{gi}(s,t_j,u_1)\frac{dN_g(s,u_1)}{Y_g(s,u_1)}}{Y_g(s,u_1)/n_g(s)} \right\} du_2.$$

An approximate $1 - \alpha$ level confidence interval for the average difference in lifetime per τ time units, $\mu_1(s,\tau) - \mu_2(s,\tau)$, becomes $\{\hat{\mu}_1(s,\tau) - \hat{\mu}_2(s,\tau)\} \pm \mathcal{Z}_{1-\alpha/2}\sqrt{\hat{\sigma}_1^2(s)/n_1(s) + \hat{\sigma}_2^2(s)/n_2(s)}$, where $\mathcal{Z}_{1-\alpha/2}$ is the $100 \times (1 - \alpha/2)\%$ quantile of the standard Normal distribution. A standard Normal(0,1) version of the test statistic can be calculated using

$$\tilde{\mathcal{T}}(s) = \frac{\mathcal{T}(s)}{\sqrt{\hat{\pi}_2(s)\hat{\sigma}_1^2(s) + \hat{\pi}_1(s)\hat{\sigma}_2^2(s)}} = \sqrt{\frac{n_1(s) n_2(s)}{n_2(s)\hat{\sigma}_1^2(s) + n_1(s)\hat{\sigma}_2^2(s)}} \{\hat{\mu}_1(s,\tau) - \hat{\mu}_2(s,\tau)\}.$$

In describing efficiency of their estimation procedure, Tayob and Murray (2016) give guidance on selection of follow-up window start times $t_1, t_2, \ldots, t_b$ based on the special case where event-times follow an exponential distribution. In this case, an analysis of their closed form asymptotic variance showed that, for a fixed number b of incorporated windows, equal spacing of $t_1, t_2, \ldots, t_b$ gave the smallest possible variability. For any fixed duration follow-up period, simulations also indicated increased efficiency in estimation with increasing b, even though increases in b create increasing amounts of overlap between a patient's incorporated short-term follow-up windows. However, Tayob and Murray (2016) found that increasing b beyond approximately $(2s - \tau)/\tau$ gave diminishing returns in efficiency; they ultimately recommended incorporating outcomes from follow-up windows starting after every $\frac{\tau}{2}$ units of follow-up time, i.e., $t = \{0, \frac{\tau}{2}, \tau, \ldots, s - \tau\}$. For instance, with $\tau = 1$ year and an interim analysis 3 years into the trial, we would incorporate information from 1-year duration follow-up windows starting at $t_1 = 0$, $t_2 = 0.5$ years, $t_3 = 1$ years, $t_4 = 1.5$ years, and $t_5 = 2$ years.

4 | MORE THAN ONE ANALYSIS AT CALENDAR TIMES, $s_1, \ldots, s_K$

At analysis time s, a decision to continue or end the clinical trial is based on the standardized test statistic, $\tilde{\mathcal{T}}(s)$, exceeding predetermined lower or upper critical values (CVs), $c_L(s)$ and $c_U(s)$, respectively. When $K > 1$ analyses are planned, group sequential methodology tells us that CVs, $\{c_L(s_1), c_U(s_1)\}, \ldots, \{c_L(s_K), c_U(s_K)\}$, corresponding to test statistics, $\tilde{\mathcal{T}}_K = \{\tilde{\mathcal{T}}(s_1), \ldots, \tilde{\mathcal{T}}(s_K)\}$, must be carefully chosen to preserve an overall type I error of α (Pocock, 1977; O'Brien and Fleming, 1979; DeMets and Lan, 1994).

CVs, $\{c_L(s_k), c_U(s_k)\}$, for the k^{th} analysis ($k = 1, \ldots, K$) can be calculated from the multivariate distribution of $\tilde{\mathcal{T}}_k$. As shown in Web Appendices A and B, $\tilde{\mathcal{T}}_k$ has a mean zero multivariate Normal distribution with $k \times k$ covariance matrix Σ, where the diagonal elements are equal to one and the off-diagonal elements, $\sigma_{k_1k_2} = \sigma_{k_2k_1}, k_1 < k_2$, can be estimated by

$$\begin{aligned}\hat{\sigma}_{k_1k_2} &= \{\hat{\pi}_2(s_{k_1})\tilde{\sigma}_1^2(s_{k_1}) + \hat{\pi}_1(s_{k_1})\tilde{\sigma}_2^2(s_{k_1})\}^{-\frac{1}{2}} \\ &\times \{\hat{\pi}_2(s_{k_2})\hat{\sigma}_1^2(s_{k_2}) + \hat{\pi}_1(s_{k_2})\hat{\sigma}_2^2(s_{k_2})\}^{-\frac{1}{2}} \\ &\times \sum_{g=1}^{2} \sqrt{\hat{\pi}_{3-g}(s_{k_1})\hat{\pi}_{3-g}(s_{k_2})\hat{\psi}_g(s_{k_1}, s_{k_2})} \\ &\times \Bigg(\sum_{i=1}^{n_g(s_{k_1})} \{n_g(s_{k_1}) - 1\}^{-1} \\ &\times \left[\tilde{z}_i\{\hat{\mu}_g(s_{k_1}, \tau)\} - \bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1}, \tau)\}\right] \\ &\times \left[z_i\{\hat{\mu}_g(s_{k_2}, \tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2}, \tau)\}\right] \Bigg)\end{aligned}$$

where $\hat{\pi}_g$, $\hat{\sigma}_g^2(s_{k_2})$, $z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}$ and $\bar{z}\{\hat{\mu}_g(s_{k_2}, \tau)\}$ have been defined in Section 3 with $s = s_{k_2}$ and $\hat{\psi}_g(s_{k_1}, s_{k_2}) = n_g(s_{k_1})/n_g(s_{k_2})$. The $z_{ij}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ terms in $\hat{\sigma}_g^2(s_{k_1})$, $z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}$ and $\bar{z}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ are replaced with

$$\begin{aligned}\tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1}, \tau)\} &= \int_0^{\tau} exp\{-\int_0^{u_2} \frac{dN_g(s_{k_1}, u_1)}{Y_g(s_{k_1}, u_1)}\} \Bigg[\int_0^{u_2} \\ &\times \Bigg\{ \sum_{l=1}^{b} \Bigg(\sum_{i=1}^{n_g(s_{k_2})} I\{T_{gi} \geq u_1 + t_l\} \sum_{i'=1}^{n_g(s_{k_1})} I\{C_{gi'}(s_{k_1}) \geq u_1 + t_l\} \Bigg) \Bigg\}^{-1} \\ &\times n_g(s_{k_1}) n_g(s_{k_2}) Y_{gi}(s_{k_1}, t_j, u_1) \\ &\times \left\{ \frac{dN_{gi}(s_{k_2}, t_j, u_1)}{Y_{gi}(s_{k_2}, t_j, u_1)} - \frac{dN_g(s_{k_1}, u_1)}{Y_g(s_{k_1}, u_1)} \right\} \Bigg] du_2\end{aligned}$$

when calculating $\tilde{\sigma}_g^2(s_{k_1})$, $\tilde{z}_i\{\hat{\mu}_g(s_{k_1}, \tau)\}$ and $\bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1}, \tau)\}$.

Examples of calculating CVs based on the joint distribution of $\tilde{\mathcal{T}}_k$ are described further in Sections 4.1 and 4.2. Section 4.1 reviews how to calculate CVs based on symmetric type I error spending functions that are in common use. In Section 4.2 we describe calculation of CVs based on asymmetric error spending approaches that differentially limit the chances of stopping incorrectly for perceived efficacy versus harm attributed to the investigational arm.

4.1 | Symmetric spending functions

Interim analysis CVs are often based on a monotonically increasing spending function, $\alpha(\gamma), 0 \leq \gamma \leq 1$, with $\alpha(0) = 0$ and $\alpha(1) = \alpha$, the desired overall type I error. A valuable advantage of spending functions is increased flexibility in scheduling interim analyses, for instance as prespecified accrual and follow-up targets are met. Spending functions that approximate the Pocock (P) and the O'Brien-Fleming (OF) approaches to type I error control are $\alpha_{OF}(\gamma) = 2 - 2\Phi(\mathcal{Z}_{1-\alpha/2}/\sqrt{\gamma})$ and $\alpha_P(\gamma) = \alpha \ln\{1 + (e-1)\gamma\}$, respectively. At interim analysis time s, γ is often taken to be the proportion of available statistical information relative to the information anticipated at the final analysis. Another common choice for γ is the proportion of expired calendar time relative to the planned trial duration.

As a simple example of the OF spending function with $\alpha = 0.05$, suppose $K = 2$ analyses are planned at s_1 and s_2. We choose to use symmetric bounds so that $c_L(s_1) = -c_U(s_1)$ and $c_L(s_2) = -c_U(s_2)$. Further suppose that at s_1, $\gamma = \frac{2}{3}$, giving $\alpha_{OF}\left(\frac{2}{3}\right) = 0.016$; at the final analysis time $\gamma = 1$ and $\alpha_{OF}(1) = 0.05$ by design. Since under the null hypothesis $\tilde{\mathcal{T}}(s_1)$ has an approximate Normal(0,1) distribution, and no type I error has been spent prior to s_1, $\{c_L(s_1), c_U(s_1)\} = \{\mathcal{Z}_{0.016/2}, \mathcal{Z}_{1-0.016/2}\}$. Calculation of $\{c_L(s_2), c_U(s_2)\}$ is not as straightforward due to stochastic dependence between $\tilde{\mathcal{T}}(s_1)$ and $\tilde{\mathcal{T}}(s_2)$. The symmetric OF spending function allows $0.05 - 0.016 = 0.034$ type I error to be spent at the 2^{nd} analysis, with 0.017 error allocated towards incorrectly claiming a statistically significant treatment benefit and 0.017 error towards incorrectly claiming statistically significant treatment harm.

Calculations for $\{c_L(s_2), c_U(s_2)\}$ are only relevant when the trial continues beyond the first interim analysis ($\mathcal{Z}_{0.016/2} < \tilde{\mathcal{T}}(s_1) < \mathcal{Z}_{1-0.016/2}$) and need to satisfy:

$$\begin{aligned}&Pr\left\{\tilde{\mathcal{T}}(s_2) \notin (c_L(s_2), c_U(s_2)) | \mathcal{Z}_{0.016/2} < \tilde{\mathcal{T}}(s_1) < \mathcal{Z}_{1-0.016/2}, H_0\right\} \\ &= \frac{\left\{\mathcal{Z}_{0.016/2} < \tilde{\mathcal{T}}(s_1) < \mathcal{Z}_{1-0.016/2}, \tilde{\mathcal{T}}(s_2) \notin (c_L(s_2), c_U(s_2)) | H_0\right\}}{Pr\left\{\mathcal{Z}_{0.016/2} < \tilde{\mathcal{T}}(s_1) < \mathcal{Z}_{1-0.016/2} | H_0\right\}} = \frac{0.034}{1-0.016} \approx 0.035.\end{aligned}$$

Suppose the estimated correlation between $\tilde{\mathcal{T}}(s_1)$ and $\tilde{\mathcal{T}}(s_2)$, i.e. σ_{12}, is 0.5. Modern software packages can easily generate a large number of mean zero bivariate normal iterates with correlation 0.5, $\{Z_m(s_1), Z_m(s_2)\}, m = 1, \ldots, M$; in simulation we used

$M = 10$ million. The desired CVs, $c_L(s_2) = -c_U(s_2)$, satisfying $Pr\{\tilde{\mathcal{T}}(s_2) \notin (c_L(s_2), c_U(s_2)) \mid \mathcal{Z}_{0.016/2} < \tilde{\mathcal{T}}(s_1) < \mathcal{Z}_{1-0.016/2}, H_0\} = 0.035$ are calculated by first subsetting the iterates who failed to reject at s_1, i.e., the set $\mathcal{S}(s_1) = \left\{m \in 1, \ldots, M : \mathcal{Z}_{0.016/2} < Z_m(s_1) < \mathcal{Z}_{1-0.016/2}\right\}$. Then $c_U(s_2) = -c_L(s_2)$ is the $1 - 0.035 = 0.965$ percentile of $|Z_m(s_2)|$ iterates taken from $\mathcal{S}(s_1)$.

The calculation of CVs in the general case with an arbitrary spending function $\alpha(\gamma)$ is similar. At analysis time s_k with γ_k, estimate Σ_k and generate M mean zero multivariate normal iterates, $\left\{Z_m(s_1), \ldots, Z_m(s_k)\right\}$, with correlation (covariance) matrix $\Sigma_k, m = 1, \ldots, M$. Calculate the subset of iterates $\mathcal{S}(s_{k-1})$ that fail to reject the null hypothesis at all previous interim analyses $1, \ldots, k-1$. Then $c_U(s_k)$ is the $1 - \frac{\alpha(\gamma_k) - \alpha(\gamma_{k-1})}{1 - \alpha(\gamma_{k-1})}$ percentile of $|Z_m(s_k)|$ iterates taken from the set $\mathcal{S}(s_{k-1})$, and $c_L(s_k) = -c_U(s_k)$.

4.2 | Asymmetric type I error control and patient protection

Symmetric stopping boundaries make it equally difficult to reject the null hypothesis due to treatment benefit or harm. These bounds are appropriate when trial monitors are blinded to the identity of the superior treatment arm at each analysis. Modern Data and Safety Monitoring Committees are rarely blinded, however, and in cases where the control is a viable therapeutic choice, there is additional motivation to end a trial where the investigational arm is trending towards harm. For the remainder of this section we consider asymmetric stopping boundaries, classified as efficacy, safety or futility bounds.

The priority of the efficacy stopping bound is to limit the clinical trial false positive rate to $\alpha/2$, where a false positive clinical trial is defined as a trial that incorrectly stops in favor of the investigational arm. Typically we choose $\alpha/2 = 2.5\%$ and use a traditional spending function approach for this bound. This bound is tightly linked to overall study power. When triggered, futility and safety bounds stop the trial without favoring the investigational arm, but are motivated by different desired operating characteristics of the trial.

The goal of a futility boundary is to terminate the trial once it seems unlikely to end with statistical evidence favoring the investigational arm (Ware et al., 1985). Criteria for defining a futility boundary are variable, chosen to have simulated operating characteristics attractive to the trial sponsor and investigative team in the trial's design phase. Such boundaries are much more aggressive at ending an unpromising trial than when compared to a symmetric stopping rule; trial sponsors using a futility boundary avoid spending resources that prove their latest offering is significantly worse than the current standard of care. Although this logic suggests a cost-benefit motivation, such boundaries have the added attraction of stopping a trial before even weak statistical evidence of harm attributed to the clinical trial has been obtained. Further discussion of futility stopping boundaries with examples can be found in Friedman et al. (2015) and Harrington (2012). If the only goal of a clinical trial is to move forward with a new therapeutic, the financial and ethical protection afforded by futility boundaries are quite attractive.

The distinction we place between a safety boundary and a futility boundary is that safety boundaries never recommend ending a trial early if the investigational arm is performing at the level of or superior to the control arm. Symmetric OF and Pocock stopping rules include a boundary that can be classified as a safety boundary, the boundary that ends the trial in favor of the control when crossed. Hereafter, we refer to these as OF or Pocock safety boundaries. It is possible to mix and match efficacy and safety boundaries using commercial software, for instance an OF efficacy boundary may be paired with a Pocock safety boundary (Proschan et al., 2006). Traditional type I error is maintained at level α, with $\alpha/2$ type I error generated from efficacy and safety boundaries, respectively. The OF efficacy bound encourages additional follow-up time for collecting data on secondary endpoints when the investigative arm is favored, while the Pocock safety boundary allows for an earlier average stopping time when the treatment arm reflecting current medical practice is favored.

In updating our own thoughts on safety boundaries, we note that (1) in the era of big data (proteomics, genetics. microbiome, etc.), clinical trial auxiliary data is tremendously valuable. Clean prospective longitudinal follow-up can generate preliminary data on disease mechanism, therapeutics and personalized medicine, for a start. For this reason, futility boundaries with very early termination of unpromising therapies seem less appealing. However, (2) we feel uncomfortable with current OF and Pocock safety boundaries that require statistically significant harm attributed to the investigational therapy before stopping a trial.

For our own clinical trials, we have sought solutions via asymmetric boundaries inspired by Jennison and Turnbull with ideas incorporated from Proschan et al. as well as DeMets and Lan (DeMets and Lan, 1994; Jennison and Turnbull, 2000; Proschan et al., 2006). In particular, we recommend a safety bound modified from a Jennison and Turnbull (JT) spending function, $\alpha_{JT}(\gamma) = \gamma^{\omega}\alpha_{\text{safety}}$, where γ is the proportion of information at the interim analysis, $\omega > 0$ is a user-defined shape parameter and $\alpha_{\text{safety}} > 0$ is a user-specified overall error rate for exceeding the safety boundary and stopping the trial under the null hypothesis. Our recommendation for ω is $\log\left\{\alpha_{\text{safety}}^{-1}\alpha/2\right\}/\log(\gamma_1)$ with γ_1 being the proportion of information at the first analysis, which allows the trial to terminate at the first interim analysis time if the test statistic indicates harm from the investigational therapy at the $\alpha/2$ significance level; hereafter we call this the JT safety boundary.

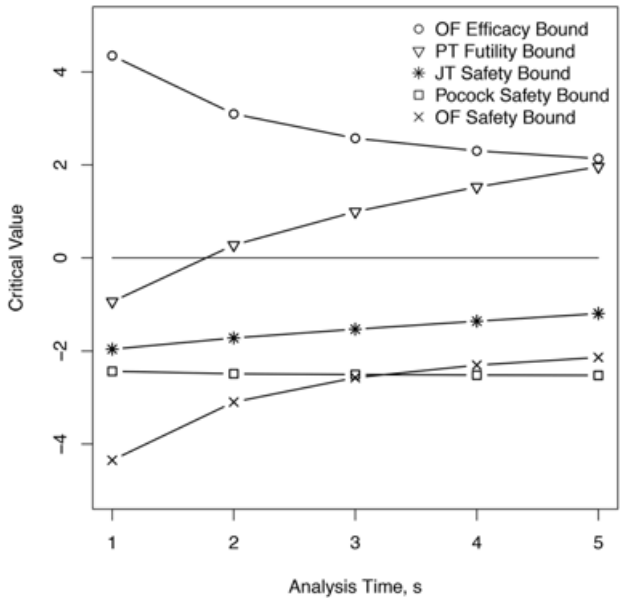


FIGURE 2 Example of efficacy, futility, and safety boundaries (OF: O'Brien and Fleming; PT: Pampallona and Tsiatis; JT: Jennison and Turnbull).

Figure 2 displays symmetric, futility and safety boundaries for a trial planning 5 interim analyses using a standardized test statistic; an OF efficacy bound with a 2.5% false positive clinical trial rate is also shown. OF and Pocock safety boundaries are also shown, where the overall probability of ending the trial incorrectly due to safety is taken to be 2.5% for each of these boundaries. The displayed JT safety boundary assumes $\alpha_{\text{safety}} = 0.20$ and $\alpha/2 = 0.025$, so that $\omega \approx 1.29$. The displayed Pampallona and Tsiatis (PT) futility bound (Pampallona and Tsiatis, 1994) is the only bound with potential to stop the trial while the investigational arm is performing at or above the level of the control.

5 | SIMULATION STUDY

In this section we summarize finite sample operating characteristics of our test statistic, with $\tau = 1$ year, against the most popular group sequentially monitored tests: the logrank test and the restricted mean survival test (RMS). In Web Appendix D of supplementary materials we summarize results for our test statistic using alternative choices of $\tau = 0.25, 0.50$ and 0.75 years as well as results for weighted logrank tests that use Peto & Peto's weight favoring early treatment differences and Fleming and Harrington's $(0.5, 0.5)$ weight favoring late differences.

In each setting we use an OF efficacy bound with a 2.5% false positive clinical trial rate. For safety, we consider (1) an OF safety boundary and (2) a Pocock safety boundary, where each of these assume an overall 2.5% chance of ending the trial incorrectly due to safety. And finally, we consider (3) a JT safety boundary assuming $\alpha_{\text{safety}} = 0.20$ and $\alpha/2 = 0.025$.

Each scenario assumes a 5 year study with 100 participants per treatment arm; 50 participants per group are accrued at baseline with the remainder accrued uniformly over 4 years. Interim analysis are conducted annually ($K = 5$). In addition to administrative censoring at each analysis time, we assume a loss-to-follow-up mechanism, $V_i = 5B_i + \tilde{E}_i \times (1 - B_i)$, where B_i and $\tilde{E}_i$ are distributed as Bernoulli(0.3) and Exponential with hazard 0.3, respectively.

Event times are generated from exponential or piecewise exponential distributions. In Scenario 1, both intervention and control arms have hazards of 0.5 throughout follow-up (null hypothesis scenario). Scenarios 2–9, shown in Figure 3 with piecewise hazards superimposed over the various survival curves, consider proportional hazard alternatives (Scenarios 2 and 3), delayed treatment effect alternatives (Scenarios 4 and 5), early treatment differences that attenuate over time (Scenarios 6 and 7) and alternatives subject to a cure pattern (Scenarios 8 and 9). Left and right panels of Figure 3 show scenarios where the investigational arm is beneficial or harmful, respectively; asymmetric stopping rules have different operating characteristics depending on the benefit/harm profile of the investigational arm.

Tables 1 and 2 summarize group sequential operating characterics in Scenarios 1 though 9. Table 1 shows rates of stopping for perceived efficacy (column 3) or a perceived safety signal (columns 4–6). Table 2 shows the average study time (AST), the average sample number (ASN) and the average number of events (ANE) for each scenario. For improved precision, scenario 1 includes 10,000 iterations; scenarios 2–9 include 1000 iterations.

Table 1, Scenario 1, shows that under the null hypothesis, all of the estimated efficacy and safety stopping rates meet their corresponding design targets within our tolerance for simulation error, where these targets were 0.025 for the OF Efficacy boundary, 0.20 for the JT safety boundary and 0.025 for the OF and Pocock safety boundaries. The JT safety boundary ends the trial more frequently (Table 1) and earlier (AST in Table 2) than either the Pocock or OF safety boundaries in Scenario 1. Regardless of the test statistic used, the JT safety boundary tends to end the trial 0.3–0.4 years earlier with 5 fewer patients enrolled and 7–9 fewer events observed (See AST, ASN and ANE, respectively in Table 2).

Regardless of test statistic used, in scenarios where the investigational drug is harmful (Scenarios 3, 5, 7, and 9), the JT safety boundary reaches a safety signal at a much higher rate than its competitor safety bounds (Table 1) and with a much smaller AST, ASN and ANE (Table 2). In scenarios where the investigational drug is beneficial (scenarios 2, 4, 6, and 8), the additional safety conferred by use of the JT bound does not reduce study power except very modestly in scenario 4, where the treatment benefit does not emerge until after the first interim analysis. In this one case, less than a

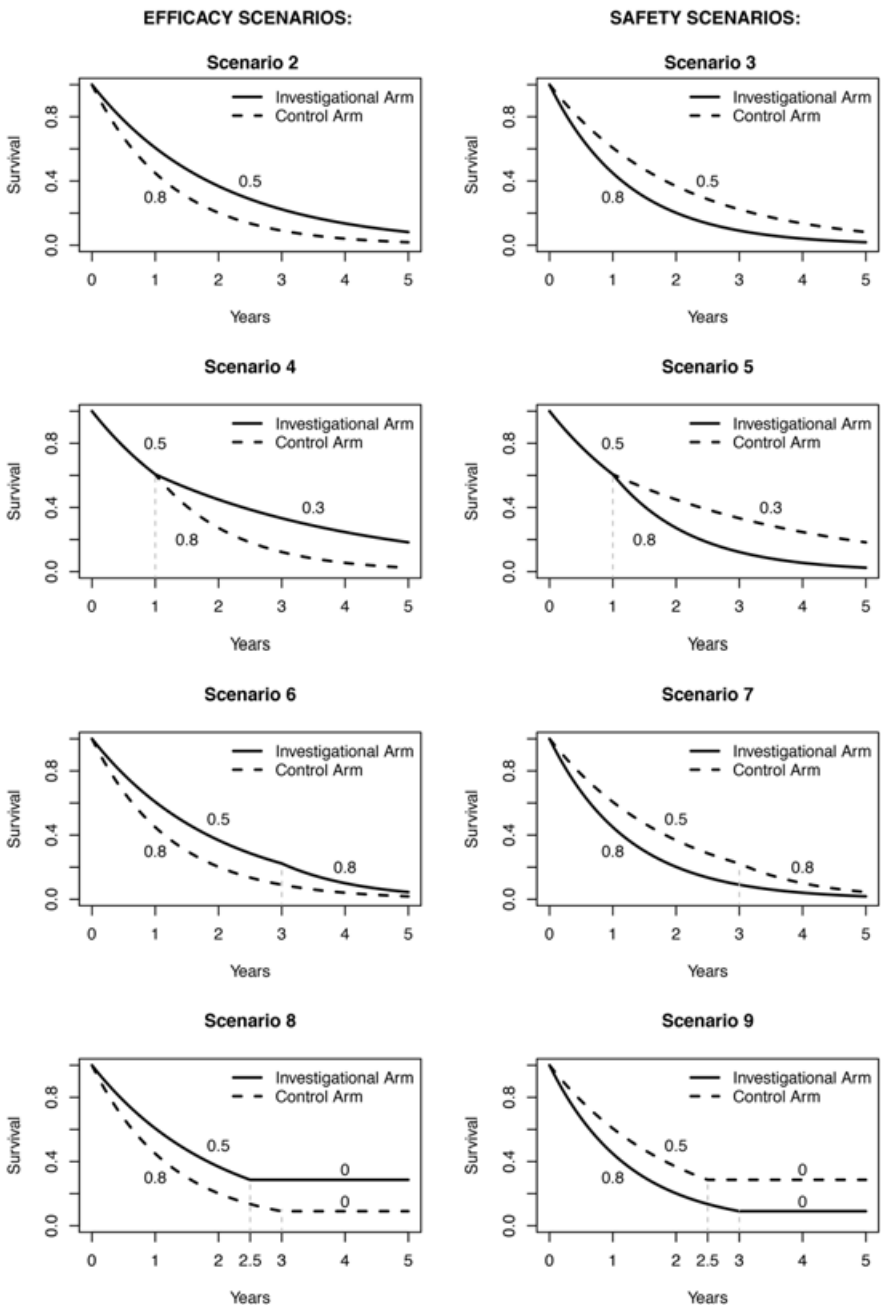


FIGURE 3 Survival probabilities of the efficacy and safety scenarios.

percentage point of simulated power is lost when using the JT safety boundary compared to the other safety boundaries.

For proportional hazards scenarios (Table 1, Scenarios 2 and 3), all three test statistics have comparable probabilities of stopping for efficacy (Scenario 2) or safety (Scenario 3), with the logrank test edging out its competitors very slightly. Table 2, likewise, gives very similar AST, ASN and ANE results for the three test statistics.

In Scenarios 4 and 5, where there is a delayed treatment effect, the proposed statistic has at least 10% higher power (Scenario 4, Table 1) with a better safety profile (Scenario 5, Table 1) compared with both the logrank and RMS tests. Modest improvements in AST, ASN and ANE are also attributed to use of the proposed test statistic (Scenarios 4 and 5, Table 2).

In Scenarios 6 and 7, where an early treatment difference emerges but becomes attenuated over time, power increases by approximately 2 percentages points when moving from the proposed to the logrank test, and from the logrank to the RMS test (Scenario 6, Table 1). Safety profiles, AST, ASN and ANE likewise slightly favor the RMS procedure over the logrank and proposed test, respectively (Scenario 7, Tables 1 and 2).

TABLE 1 Rates of stopping for efficacy (OF Efficacy) or for safety (JT Safety, P Safety, OF Safety).

Scenario	Test statistic	OF Efficacy	JT Safety	P Safety	OF Safety
	Proposed	0.022	0.198	0.026	0.025
1	RMS	0.023	0.198	0.027	0.027
	Logrank	0.022	0.193	0.022	0.024
	Proposed	0.807	0.002	0	0
2	RMS	0.816	0.001	0	0
	Logrank	0.820	0.001	0	0
	Proposed	0	0.982	0.790	0.838
3	RMS	0	0.980	0.786	0.852
	Logrank	0	0.987	0.799	0.849
	Proposed	0.855–0.863 †	0.021	0.007	0.001
4	RMS	0.715–0.722 †	0.034	0.010	0
	Logrank	0.745–0.749 †	0.026	0.007	0
	Proposed	0	0.979	0.787	0.860
5	RMS	0	0.939	0.619	0.731
	Logrank	0	0.959	0.642	0.765
	Proposed	0.761	0	0	0
6	RMS	0.802	0	0	0
	Logrank	0.781	0	0	0
	Proposed	0	0.960	0.709	0.738
7	RMS	0	0.965	0.736	0.786
	Logrank	0	0.971	0.730	0.764
	Proposed	0.884	0	0	0
8	RMS	0.771	0.001	0	0
	Logrank	0.863	0	0	0
	Proposed	0	0.989	0.847	0.885
9	RMS	0	0.955	0.727	0.770
	Logrank	0	0.983	0.826	0.871

† There is potential for OF efficacy rates to be affected by the safety boundary used, for instance when an efficacy boundary would have been crossed if not for an earlier safety boundary being crossed. This was only observed in Scenario 4 of our simulations. In this scenario we give a range of observed OF efficacy stopping rates for each test statistic, where the lower OF efficacy stopping rate shown corresponds to use of the JT safety boundary (most strict safety boundary) and the higher OF efficacy stopping rate shown corresponds to the OF safety boundary (least strict safety boundary).

In Scenarios 8 and 9, where a cure pattern emerges during the trial, the proposed test statistic has approximately 2% and 10% higher power than the logrank and RMS tests, respectively (Scenario 8, Table 1). Safety profiles shown for Scenario 9 in Table 1 likewise reflect a slight improvement over the logrank test and a large improvement over the RMS test. AST, ASN, and ANE results, however, show only minimal differences (Scenario 8 and 9, Table 2).

6 | EXAMPLE

Fischl et al. (1990), on behalf of the AIDS Clinical Trials Group (ACTG), randomized 524 patients to high-dose ($n = 262$) versus low-dose ($n = 262$) azidothymidine (AZT). The standard, higher AZT dose succeeded in reducing mortality but came with substantial toxicity. Investigators hoped that the lower dose would reduce toxicity while maintaining the survival benefit. Figure 4(a) displays the average number of additional days lived per year when taking low versus high dose AZT, estimated using our methodology with $\tau = 1$ year, at analysis times in 1987, 1988, 1989 and 1990. Although validity of our testing procedure does not require a stable treatment effect over time, the low-dose AZT benefit appears approximately stable at each analysis. Using our proposed group sequentially monitored test statistic, the OF efficacy boundary is crossed at the 1990 analysis with the low dose group living an estimated 10.7 days longer per year than the high dose group. The JT safety boundary ensures early trial termination if the experimental low-dose trends towards higher mortality, but this boundary was not crossed. Web Appendix E of supplementary materials summarizes how group sequentially monitored logrank and RMS tests performed in this case. As seen in Figure 4(b), there was a delayed treatment effect that perhaps favored our methodology as compared to the logrank and RMS methods. Neither of these competitors recommended stopping at the 1990 analysis time.

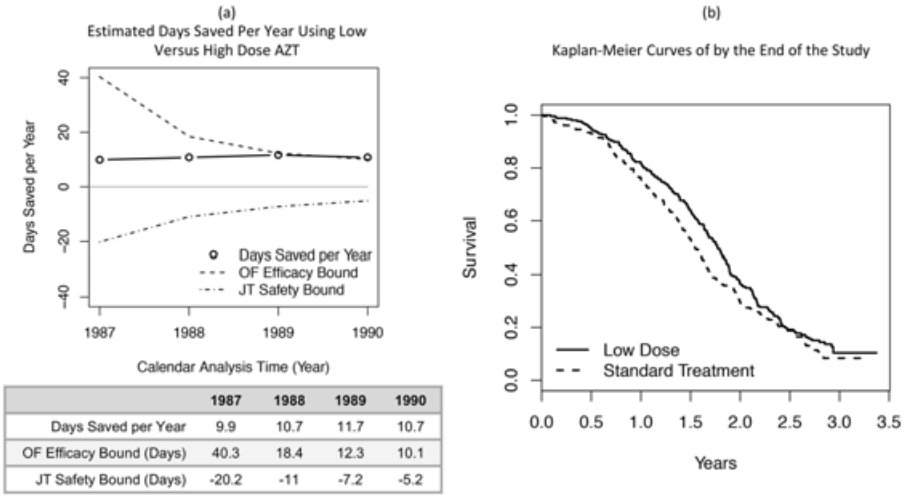
7 | DISCUSSION

There are a good many nonparametric group sequential monitoring methods available for censored time-to-event outcomes in clinical trials, the logrank test and the RMS test among

TABLE 2 Average study time (AST) in years, average sample number (ASN), and average number of events (ANE) in Scenarios 1–9.

		AST			ASN			ANE		
Scenario	**Test statistic**	**JT**	**P**	**OF**	**JT**	**P**	**OF**	**JT**	**P**	**OF**
	Proposed	4.7	4.9	5.0	195	199	200	156	163	164
1	RMS	4.6	4.9	5.0	195	199	200	156	163	164
	Logrank	4.7	4.9	5.0	195	199	200	156	164	165
	Proposed	3.7	3.7	3.7	185	185	185	143	143	143
2	RMS	3.7	3.7	3.7	184	184	184	142	142	142
	Logrank	3.7	3.7	3.7	185	185	185	143	143	143
	Proposed	2.1	3.0	3.6	151	169	184	93	121	142
3	RMS	2.1	3.1	3.6	151	170	183	93	123	141
	Logrank	2.0	3.0	3.6	149	169	184	90	121	141
	Proposed	3.7	3.8	3.8	187	188	188	132	134	134
4	RMS	4.2	4.3	4.3	193	195	196	143	146	147
	Logrank	4.0	4.1	4.1	191	192	193	140	142	143
	Proposed	2.8	3.8	3.9	169	185	189	108	132	137
5	RMS	3.4	4.3	4.3	180	193	195	123	146	147
	Logrank	3.0	4.1	4.2	173	189	193	113	140	144
	Proposed	3.7	3.7	3.7	184	184	184	143	143	143
6	RMS	3.6	3.6	3.6	183	183	183	141	141	141
	Logrank	3.7	3.7	3.7	184	184	184	143	143	143
	Proposed	2.2	3.2	3.8	153	171	185	96	126	145
7	RMS	2.2	3.2	3.6	153	171	183	96	126	142
	Logrank	2.1	3.1	3.7	150	170	184	92	124	143
	Proposed	3.4	3.4	3.4	181	181	181	128	128	128
8	RMS	3.7	3.7	3.7	183	183	183	132	132	132
	Logrank	3.5	3.5	3.5	182	182	182	130	130	130
	Proposed	2.1	3.0	3.5	152	169	182	92	113	130
9	RMS	2.2	3.2	3.7	154	172	184	94	118	133
	Logrank	2.0	3.0	3.6	150	169	183	90	113	131

the most popular, and so a natural question is what the proposed test statistic offers clinical trial researchers that the others do not. We see both philosophical and operational advantages to this statistic being used in practice. The philosophical argument hinges on the idea that times-to-event can be repurposed into a longitudinal data structure, with repeated measures within individual measured regularly throughout follow-up. Each τ-restricted time-to-event carved from the overall follow-up time can be thought of as a longitudinal measure in this philosophy. Tayob and Murray (2016) proposed an improved estimate of τ-restricted means based on this idea and showed that τ-length follow-up windows starting



	1987	1988	1989	1990
Days Saved per Year	9.9	10.7	11.7	10.7
OF Efficacy Bound (Days)	40.3	18.4	12.3	10.1
JT Safety Bound (Days)	-20.2	-11	-7.2	-5.2

FIGURE 4 Figures in example: (a) estimated days saved per year using low versus high dose AZT; (b) Kaplan–Meier curves of by the end of the study.

every $\tau/2$ time units apart give attractive efficiency in estimating restricted means without unduly increasing computational time in creating these longitudinal measures. They extended this idea to the parametric setting in Tayob and Murray (2017), multiply imputing censored event times and then using standard generalized estimating equation methods for analyzing the longitudinal restricted event times. There is great potential in shifting our thoughts on censored times-to-event towards longitudinal data structures and the available methodology this shift entails.

In this article, we develop a two-sample test statistic based on comparing τ-restricted means as introduced in Tayob and Murray and we further develop group sequential monitoring methodology for using the test statistic in standard clinical trial settings where interim monitoring is common. The validity of the proposed testing procedure does not hinge upon stability of τ-restricted means in the different follow-up windows; the type I error is preserved regardless of the true event-time distribution. In scenarios where τ-restricted means are not stable over time, the test statistic compares overall τ-restricted means of mixture distributions that result from combining information from overlapping follow-up windows.

Event rates that shift year-by-year affect the power of all two sample testing procedures. It is well known that non-proportional hazards plague the power of the logrank test. Restricted mean differences also change as the period of follow-up lengthens, with differences becoming larger or smaller as event rates shift over time. As with all two-sample tests, as data accumulates, so does our interpretation of the data and the power of the testing procedure. The main concern in choosing any two-sample test statistic is whether authentic treatment differences can be detected with high power.

Our proposed method performs well not only in scenarios where short-term differences are anticipated to be stable, but also in settings that it may be hard to anticipate in the design stage of a clinical trial. We find it comforting that our methodology compares favorably to its competitors in proportional hazards settings, and has a notably improved performance in settings where treatment differences emerge only after a certain period of time or in settings where there is potential for cure.

Simulations suggest that shifting towards a longitudinal view of censored survival outcomes has practical advantages in group sequential monitoring of clinical trials. The feature of overlapping follow-up windows used in creating repeated τ-restricted event times subject to censoring is reminiscent of smoothing methods in graphical displays of longitudinal data.

An additional contribution of this article is an updated look at safety boundaries in the group sequential setting and a new recommendation for the shape parameter ω used with the JT spending function. Our recommended shape boundary allows the first interim analysis to reject if the standardized normal test statistic exceeds the safety boundary of -1.96, which clinical investigators have been hard-wired to associate with statistical significance. Data and safety monitoring committees are likely to feel uncomfortable continuing a trial that exceeds this critical value and yet for many years biostatisticians have taught investigators the consequences of using traditional significance levels in the group sequential setting in terms of inflated type I errors. This article emphasizes the idea that type I error inflation has different consequences for the efficacy boundary as opposed to the safety boundary. We argue that is is possible to maintain an overall false positive trial result to an $\alpha/2$ level using an appropriate efficacy boundary and separately strategize a stopping rule that protects safety without unduly reducing power of the study. Our recommended variant of the JT safety bound achieves this goal with remarkable effectiveness as seen in simulation. We ultimately recommend use of our proposed test statistic in the group sequential setting using OF efficacy and our JT safety boundaries.

ORCID

Meng Xia http://orcid.org/0000-0002-9711-6215
Nabihah Tayob http://orcid.org/0000-0001-6088-167X

REFERENCES

Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika* 57, 579–594.

DeMets, D. L. and Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Stat Med* 13, 1341–1352.

DeMets, D. L. and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 69, 661–663.

Fischl, M., Parker, L., Petinelli, C., et al. (1990). A randomized controlled trial of a reduced daily dose of zidovudine in patients with the aquired immunodeficiency syndrome. *N Engl J Med* 323, 1009–1014.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., and Granger, C. B. (2015). *Foundamentals of Clinical Trials*. Springer.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52452, 203–223.

Harrington, D. P. (2012). *Design for Clinical Trials*. Springer.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* 69, 553–566.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70, 659–663.

Li, Z. (1999). A group sequential test for survival trials: An alternative to rank-based procedures. *Biometrics* 55, 277–283.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *J Am Stat Assoc* 58, 690–700.

Mantel, N. (1966). Evaluation of survival data and two new rank-order statistics arising in its consideration. *Cancer Chemother Rep* 50, 163–170.

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

O'Brien, P. and Fleming, T. (1979). A multiple testing procedure for clinical trials. *Biometrics* 35, 549–556.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J Stat Plan Inference* 42, 19–35.

Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics* 45, 497–507.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser A* 135, 185–207.

Pocock, S. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 191–199.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* 65, 167–179.

Proschan, M. A., Lan, K. K. G., and Wittes, J. T. (2006). *Statistical Monitoring of Chinical Trials: A Unified Approach*. Springer.

Tayob, N. and Murray, S. (2016). Nonparametric restricted mean analysis across multiple follow-up intervals. *Stat Probabil Lett* 109, 152–158.

Tayob, N. and Murray, S. (2017). Statistical consequences of a successful lung allocation system – Recovering information and reducing bias in models for urgency. *Stat Med* 36, 2435–2451.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* 68, 311–315.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J Am Stat Assoc* 77, 855–861.

Ware, J. H., Muller, J., and Braunwald, E. (1985). The futility index. An approach to the cost-effective termination of randomized clinical trials. *Am J Med* 78, 635–643.

SUPPORTING INFORMATION

Web Appendices A–E are available at the Biometrics website on Wiley Online Library. An R package implementing the proposed test is available at https://github.com/summerx0821/Nonparametric-GS-Methods-for-Evaluating-Survival-Benefit as well as the Biometrics website on Wiley Online Library.

How to cite this article: Xia M, Murray S, Tayob N. Nonparametric group sequential methods for evaluating survival benefit from multiple short-term follow-up windows. *Biometrics.* 2019;75:494–505. https://doi.org/10.1111/biom.13007