Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Nonparametric Group Sequential Methods for Recurrent and Terminal Events from Multiple Follow-up Windows

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

[1]University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109, USA
[2]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA

Correspondence
Susan Murray, Department, University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109, USA
Email: skmurray@umich.edu

Summary

Few methods are currently available for group sequential analysis of recurrent events data subject to a terminal event in the clinical trial setting. This research helps fill this gap by developing a completely nonparametric group sequential monitoring procedure for use with the two-sample Tayob and Murray [1] statistic. Advantages of the Tayob and Murray statistic include high power to detect treatment differences when there is correlation between recurrent event times or between recurrent and terminal events in an individual. This statistic does not suffer bias from dependent censoring, regardless of the correlation between event times in an individual. This manuscript briefly reviews the Tayob and Murray statistic, develops and describes how to use methods for its group sequential analysis, and through simulation compares its operating characteristics with those of Cook and Lawless [2], which is currently in use as the only available nonparametric method for group sequential analysis of recurrent event data. The merits of our proposed approach are most clearly demonstrated when gap times between recurrent events are correlated; when gap times between events are independent the Cook and Lawless method is difficult to beat. Simulations demonstrate that as correlation between recurrent event times grows, the reduction in power using the Cook and Lawless approach is substantial when compared to our method. Finally, we use our method to analyze recurrent acute exacerbation outcomes from the Azithromycin in COPD Trial.

KEYWORDS:
Group Sequential Methods, Nonparametric Test, Recurrent and Terminal Events

1 | INTRODUCTION

Consider the typical setting for a two-arm clinical trial of a chronic, slowly progressing terminal disease. Several lung diseases fall into this category including Interstitial Pulmonary Fibrosis (IPF), Chronic Obstructive Pulmonary Disease (COPD) and Cystic Fibrosis (CF), among others. Pulmonary exacerbations are a common recurrent event in these patients, with some patients also experiencing terminal events. The Azithromycin in COPD Trial (NACT) is one of many clinical trials following this pattern. More generally, patients may experience a mixture distribution of important, potentially recurring signals of disease progression during follow-up. In IPF studies, for example, patients are considered progressors if they experience an acute exacerbation, a 10% decline in forced vital capacity (FVC), a 15% decline in diffusing capacity of the lung for carbon monoxide (DLCO), lung transplantation or death, where these latter two events are each considered terminal for lung outcome follow-up. Clinical research design is often based on time to the first occurrence of a recurrent event or the first event from a list of potentially recurring progression outcomes.

There are advantages and disadvantages to following only the first time-to-event. The most obvious advantage is the existence of several methods for group sequential clinical trial design and analysis of censored survival data that are applicable to a single time-to-event or time-to-combined-endpoint[3,4,5,6,7]. An obvious disadvantage, however, is the loss of information from ignoring progression events after the first that occurs for each patient. Consider Figure 1, which shows progression endpoints from an IPF patient followed as part of the COMET study (Correlating Outcome Measures to Estimate Time to progression in IPF[8]). This patient's first observed progression endpoint involves a decline in DLCO. An analysis based only on the first time-to-combined endpoint will ignore information on the subsequent progression endpoints, acute exacerbation and death.

Although there are several available methods for conducting two-sample tests of recurrent event data when a single analysis is conducted (based on, for example, Andersen and Gil (1982)[9], Lin et al. (2000)[10], Prentice et al. (1981)[11], Ghosh and Lin (2000)[12] or Tayob and Murray (2014)[1]), there is little available methodology for conducting group sequential analysis in this setting. The most highly cited method was introduced by Cook and Lawless[2], who developed nonparametric group sequential methods for pseudo-score statistics monitored over time. Their method does not introduce any assumptions regarding the dependence structure between recurrent event times and is framed to perform well when the cumulative mean number of events is proportional over time. Cook et al.[13] later extended this method to settings with multiple treatment periods. A parametric group sequential data analysis approach was put forward by Jiang[14], who assumed local Poisson processes that allow event rates to change over time as well as a frailty parameter to address correlation between event times.

Recognizing that nonparametric methods are vastly preferred in clinical trial settings subject to approval by regulatory agencies, and following the example of Cook and Lawless in this regard, this manuscript aims to contribute new group sequential methodology for the recurrent event setting without introducing assumptions that could adversely affect the interpretation of the observed data. In particular, we develop nonparametric group sequential methods for monitoring the Tayob and Murray statistic[1] in the recurrent events setting subject to a terminating event. In framing their statistic, recurrent event outcomes are restructured into a series of censored longitudinal times-to-first-event in regularly spaced short-term (length τ) follow-up windows for each patient. Their test then compares the difference between overall τ-restricted mean event-times between groups. In the case of a single analysis, Tayob and Murray demonstrated nice operating characteristics of their statistic in analyzing a mixture of recurrent and terminal events, with superior performance to methods of Lin et al.[10] and Ghosh and Lin[12] when recurrent and terminal events were correlated. The development of group sequential methods for this nonparametric statistic will improve the current arsenal of statistical methods for clinical trial monitoring.

The remainder of this manuscript is organized as follows. Section 2 defines notation required to repurpose traditional recurrent events data available at analysis time s into a series of censored longitudinal times-to-first event in regularly spaced short-term (length τ) follow-up windows for each patient. Section 3 briefly reviews the Tayob and Murray two-sample testing procedure in the case of a single analysis. Section 4 extends methodology to the group sequential setting. Section 5 describes simulated operating characteristics of our method compared to that of Cook and Lawless[2]. We demonstrate the method using data from the Azithromycin in COPD Trial. Discussion follows in Section 7.

2 | NOTATION

We borrow notation from Tayob and Murray[1], additionally embedding a 'calendar time' scale parameter, s, to allow for terms that change according to analysis time. For simplicity, we assume that s indexes time from initiation of the overall study rather than an actual calendar date. Patient entry times and interim analysis times are both described on this time scale. A separate 'study time' scale, indexed by t, denotes time from a participant's entry to the study. Participants' time at risk, duration of follow-up as well as times to recurrent and terminating events are measured on this time scale.

We temporarily submerge notation corresponding to treatment group g, initially focusing on the one-sample case. Suppose $i = 1, \dots, N$ patients enter a clinical trial at calendar times $E_1, E_2, ..., E_N$. Interim analyses of accumulated data are planned at calendar times, $s = s_1, s_2, ..., s_K$. Let $n(s) = \sum_{i=1}^{N} I(E_i \leq s)$ index the number of accrued individuals at interim analysis time, s, with $n(s) = N$ for $s \geq \max(E_1, \dots, E_N)$.

Recurrent events for individual i occur at times $T_{i1} < T_{i2} < \cdots < T_{iJ_i-1}$ on the study time scale, with a terminating event at time T_{iJ_i}. For each individual, i, V_i is a loss-to-follow-up time measured from study entry. The censoring random variable that also incorporates administrative censoring, $C_i(s) = \min(V_i, s - E_i)$, updates at each analysis time, s. Recurrent and terminal events for participant i are subject to independent censoring by $C_i(s)$. However, an arbitrary dependence structure is allowed between all events T_{ij_1} and $T_{ij_2}, j_1 \neq j_2$, taken from patient i. In particular, the multivariate distribution of gap times for each patient i, $\{T_{i1}, T_{i2} - T_{i1}, \dots, T_{iJ_i-1} - T_{iJ_i-2}\}$, is not constrained to an independent covariance structure.

Traditionally observed data for patients accrued prior to analysis time s is recorded as $X_{ij}(s) = min\{T_{ij}, C_i(s)\}$, $j = 1, \dots, \tilde{J}_i(s)$ and $\delta_{ij}(s) = I\{T_{ij} \leq C_i(s)\}$, $j = 1, \dots, \tilde{J}_i(s)$, where $\tilde{J}_i(s) \leq J_i$ is the number of observed event times. However, the Tayob and Murray statistic reorganizes the observed data into τ-length, potentially overlapping, follow-up windows starting at regularly-spaced study times $t \in \{t_1, t_2, \dots, t_b\}$ with $t_1 = 0$ and b equal to the ceiling of s/a, so that t_b does not exceed the available follow-up at analysis time s. Within each τ-length follow-up window, the first τ-restricted time-to-event is recorded, along with the corresponding censoring indicator.

For each individual i, a notational bookkeeper that updates at each analysis time s, $\eta_i(s,t) = min\{j = 1, \dots, \tilde{J}_i(s) : X_{ij}(s) \geq t\}$, indexes the time-to-first-event in a follow-up window starting at t from the original sequence of observed events. Using this index simplifies notation for the time-to-first event in this window at analysis time s, $X_i(s,t) = X_{i\eta_i(s,t)}(s) - t$ and its corresponding failure indicator $\delta_i(s,t) = \delta_{i\eta_i(s,t)}(s)$.

Tayob and Murray discuss advantages of this data restructuring at length. In short, a rather complex correlated gaptime data structure that is subject to dependent censoring by $C_i(s)$ is converted to a well-behaved longitudinal outcomes dataset that is subject to independent censoring by $C_i(s)$. One feature that emerges as a consequence of this data restructuring is the possibility that a recurrent event is tagged in more than one follow-up window for analysis. Hence careful attention to the correlation structure that takes this additional complexity into account is implemented. There is also the possibility of a recurrent event being excluded from the analysis, which can be mitigated by more frequently spaced window start times, t.

In a special case with exponentially distributed gap times between events, Xia and Murray [15] quantified the average proportion of recurrent events captured in at least one follow-up window when traditional recurrent event data is restructured in the manner of Tayob and Murray. This proportion approaches one as the equal spacing between follow-up window start times, $a = t_j - t_{j-1}, j = 2, \dots, b$, approaches zero. However, the computational burden associated with very small a led to their recommendation that a be a fraction of the anticipated mean recurrent event time in the control group. In particular, their rule of thumb suggested $a = 1/2$ or $1/3$ of the control group mean recurrent event time would tend to capture 80% and 90% of the events, respectively, in the case of exponentially distributed gap times between events.

To solidify some of the notation presented above, consider Figures 2 and 3. In Figure 2, which is indexed by study time, different spacing of follow-up windows ($a = 50, 100$ and 200 days) are shown for the example COMET patient previously mentioned in the introduction. The choice of $a = 200$ days results in two observed events being included in the analysis, the DLCO decline at 105 days and the acute exacerbation at 298 days. However the death at 331 days is overlooked in the analysis since it is not the first event to be observed in either of the follow-up windows starting at zero or 200 days. Both $a = 50$ and $a = 100$ days capture all three events in the analysis.

Moving forward with $a = 100$ in Figure 3, and superimposing calendar time s in addition to study time t, we see the patient entering the study at $E_i = 15$ days from the initiation of the study in calendar time. The first interim analysis is conducted at $s_1 = 157$ days in calendar time, at which time only a single event has been observed at $T_{i1} = 105$ days from study entry. The patient's data is administratively censored at $C_i(157) = 142$ days. The traditional version of the recurrent events data at this analysis time is $\{[X_{i1}(157) = 105, \delta_{i1}(157) = 1]; [X_{i2}(157) = 142, \delta_{i1}(157) = 0]\}$, so that $\tilde{J}_i(157) = 2$. At analysis time $s_1 = 157$ days, the longitudinal data structure imposed by Tayob and Murray has two data triplets from follow-up windows starting at $t = 0$ and $t = 100$: $\{\eta_i(157,0) = 1, X_i(157,0) = 105, \delta_i(157,0) = 1\}$ and $\{\eta_i(157,100) = 1, X_i(157,100) = 5, \delta_i(157,100) = 1\}$, so that the $T_{i1} = 105$ event is captured as the first observed event in each of these two follow-up windows.

At the second analysis time at $s_2 = 369$ days, administrative censoring for patient i is updated to $C_i(369) = 354$. The traditional recurrent events data becomes $\{[X_{i1}(369) = 105, \delta_{i1}(369) = 1]; [X_{i2}(369) = 298, \delta_{i2}(369) = 1]; [X_{i3}(369) = 331, \delta_{i3}(369) = 1]\}$, so that $\tilde{J}_i(369) = 3$. The restructured longitudinal dataset includes data from 4 follow-up windows starting at $t = 0, 100, 200$ and 300 yielding the data triplets $\{[\eta_i(369,0) = 1,\ X_i(369,0) = 105,\ \delta_i(369,0) = 1]; [\eta_i(369,100) = 1, X_i(369,100) = 5, \delta_i(369,100) = 1]; [\eta_i(369,200) = 2, X_i(369,200) = 98, \delta_i(369,200) = 1]; [\eta_i(369,300) = 3, X_i(369,300) = 31, \delta_i(369,300) = 1]\}$.

We now define the counting and at risk processes corresponding to the restructured longitudinal dataset at interim analysis time, s. For a follow-up window starting at time t, u indexes time from t in that window. For any individual i with $E_i < s$, $N_i(s,t,u) = I\{X_i(s,t) \leq u, \delta_i(s,t) = 1\}$ is the event counting process for the time to first event in the follow-up window starting at time t. The corresponding at risk process is $Y_i(s,t,u) = I\{X_i(s,t) \geq u\}$. Let $N(s,t,u) = \sum_{i=1}^{n(s)} N_i(s,t,u)$ and $Y(s,t,u) = \sum_{i=1}^{n(s)} Y_i(s,t,u)$ sum these processes across individuals entered by interim analysis time s.

At interim analysis s, let $N_i(s,u) = \sum_{j=1}^{b} N_i(s,t_j,u)$ count the observed times-to-first-event across the b follow-up windows attributed to individual i that are seen prior to window time u; the corresponding at risk process is $Y_i(s,u) = \sum_{j=1}^{b} Y_i(s,t_j,u)$. Pooling time-to-first event data across all follow-up windows and all individuals observed at interim analysis time s, we define $N(s,u) = \sum_{i=1}^{n(s)} N_i(s,u)$ and $Y(s,u) = \sum_{i=1}^{n(s)} Y_i(s,u)$.

It will be convenient to also index hazard functions according to the three time indices $\{s,t,u\}$. At analysis time s, let hazard function

$$\lambda(s,t,u) = \lim_{\Delta u \to 0} [Pr\{u \le X_i(s,t) < u + \Delta u, \delta_i(s,t) = 1 | X_i(s,t) \ge u\}/\Delta u].$$

The index, s, can be dropped as superfluous in the first term, i.e., $\lambda(s,t,u) = \lambda(t,u)$. This is not true for the hazard function corresponding to the mixture distribution of times-to-first event contributed from the various follow-up windows from individuals at analysis time s, $\lambda^W(s,u)$.

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

Because $\lambda^W(s,u)$ is a function of $Pr\{X_i(s,t) \ge u\}$, this term can potentially change as more follow-up information accumulates at later interim analyses.

3 | NONPARAMETRIC TWO-SAMPLE TESTS FOR RECURRENT EVENTS AND TERMINAL EVENTS AT SINGLE ANALYSIS TIME

In this section, we review the Tayob and Murray test statistic, introducing additional notation for when a single analysis is performed at, say, calendar time s. Subscripts $g = 1, 2$, indicate treatment group when used with notation from the last section. Throughout the following, random variables from different treatment groups are assumed to be independent of one another. Later in section 4, we extend these methods to the case where more than one analysis is performed at calendar times $s_1, s_2, \dots, s_K$ in the group sequential clinical trial setting.

The estimated overall τ-restricted mean time-to-first-event for treatment group g based on the restructured longitudinal dataset available at analysis time s is

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

which consistently estimates the mean of this mixture distribution of τ-restricted times-to-first-event, i.e., $\mu_g(s,\tau) = \int_0^{\tau} exp\left\{ - \int_0^{u_2} \lambda_g^W(s,u_1) du_1 \right\} du_2$.

Let $\pi_g(s)$ be the proportion of individuals in group g at analysis time s, with consistent estimate $\hat{\pi}_g(s) = n_g(s)/\{n_1(s)+n_2(s)\}$. At analysis time, s, the Tayob and Murray statistic tests the null hypothesis, $H_0 : \mu_1(s,\tau) = \mu_2(s,\tau)$, using

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

which under H_0 converges asymptotically to a mean zero Normal distribution with variance

$$\pi_2(s)\sigma_1^2(s) + \pi_1(s)\sigma_2^2(s),$$

where

$$\hat{\sigma}_g^2(s) = \sum_{i=1}^{n_g(s)} [z_i\{\hat{\mu}_g(s,\tau)\} - \bar{z}\{\hat{\mu}_g(s,\tau)\}]^2/[n_g(s) - 1],$$

$$z_i\{\hat{\mu}_g(s,\tau)\} = \sum_{l=1}^{b} z_{il}\{\hat{\mu}_g(s,\tau)\},$$

$$\bar{z}\{\hat{\mu}_g(s,\tau)\} = \sum_{i=1}^{n_g(s)} z_i\{\hat{\mu}_g(s,\tau)\}/n_g(s)$$

and $z_{il}\{\hat{\mu}_g(s,\tau)\} =$

$$\int_0^{\tau} exp\left\{-\int_0^{u_2}\frac{dN_g(s,u_1)}{Y_g(s,u_1)}\right\}\left\{\int_0^{u_2}\frac{dN_{gi}(s,t_l,u_1)-Y_{gi}(s,t_l,u_1)\frac{dN_g(s,u_1)}{Y_g(s,u_1)}}{Y_g(s,u_1)/n_g(s)}\right\}du_2. \qquad (1)$$

An approximate $1-\alpha$ level confidence interval for the average treatment difference in τ-restricted times-to-first-event, $\mu_1(s,\tau)-\mu_2(s,\tau)$, becomes

$$\{\hat{\mu}_1(s,\tau)-\hat{\mu}_2(s,\tau)\} \pm \mathcal{Z}_{1-\alpha/2} \times \sqrt{\hat{\sigma}_1^2(s)/n_1(s)+\hat{\sigma}_2^2(s)/n_2(s)},$$

where $\mathcal{Z}_{1-\alpha/2}$ is the $100\times(1-\alpha/2)\%$ quantile of the standard Normal distribution. For finite sample sizes and a single planned analysis at time s, the standardized test statistic

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

follows an approximate Normal(0,1) distribution, with critical values of ± 1.96 conferring an overall type I error of 5%. In the special case where only the first time-to-event is used in the analysis, a test statistic and corresponding group sequential monitoring procedure was developed by **(author?)**[7]. However, there is no group sequential method available for the setting with recurrent events available, which is what we develop in the following section.

4 | MORE THAN ONE ANALYSIS AT CALENDAR TIMES, $S_1, S_2, \ldots, S_K$

In this section, we extend methodology for the Tayob and Murray statistic to the group sequential setting. At each analysis time s the standardized test statistic, $\tilde{\mathscr{T}}(s)$, is evaluated and a decision to either end the trial early or continue is made based on upper and lower critical values, $c_L(s)$ and $c_U(s)$, respectively. With $K > 1$ planned analyses, critical values $\left\{c_L(s_1), c_U(s_1)\right\}, \ldots, \left\{c_L(s_K), c_U(s_K)\right\}$ corresponding to test statistics, $\tilde{\mathscr{T}}_K = \left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_K)\right\}$, must be carefully chosen to preserve an overall type I error of α[16,17]. Type I error spending functions are the most common approach for designating type I error to be used at interim analyses so that no more than α type I error is used throughout the clinical trial[18,19]. The O'Brien-Fleming (OF) spending function, $\alpha_{OF}(\gamma) = 2 - 2\Phi(\mathcal{Z}_{1-\alpha/2}/\sqrt{\gamma})$, proposed by Lan and DeMets is the most common spending function used in practice, although the only requirement for a spending function, $\alpha(\gamma)$, is that it be monotonically increasing over $(0, \alpha)$ as γ increases from zero to one.

Information-based type I error spending takes the spending function parameter, γ, to be the proportion of statistical information available at interim analysis time s_k relative to the information that will be available at the final analysis at time s_K, $k = 1, \ldots, K$. To our knowledge, the two-sample logrank test is the only group sequentially monitored statistic for time-to-event data where this information proportion reduces to a simple calculation; in this case γ is a ratio of observed events at s_k to the number of events used in powering the study. For the Tayob and Murray statistic, the proportion of information at analysis time s_k is $Var\mathscr{T}(s_K)/Var\mathscr{T}(s_k)$, where $Var\mathscr{T}(s_K)$ can be estimated via simulation using distributional and design assumptions used in powering the trial.

A common simplistic surrogate for statistical information is to use the proportion of calendar time that has passed at analysis time s relative to the planned duration of the trial. The method for estimating γ at each analysis time may affect study power, but typically to a less extent than the choice of spending function[20]. For simplicity, we use the calendar time surrogate for statistical information in our simulation and example sections. The type I error level is maintained for any spending function where at the final analysis, $\gamma = 1$.

Derivation of critical values for the k^{th} interim analysis also requires knowledge of the multivariate distribution of $\tilde{\mathscr{T}}_k = \left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_k)\right\}$, $(k = 1, \ldots, K)$. Let Σ_k be the $k \times k$ covariance matrix for $\tilde{\mathscr{T}}_k$, so that the k_1^{st}, k_2^{nd} element $\sigma_{k_1 k_2}$ of this matrix is $Cov\left\{\tilde{\mathscr{T}}(s_{k_1}), \tilde{\mathscr{T}}(s_{k_2})\right\}$, $k_1, k_2 \leq k$. Because each test statistic has already been standardized to have variance 1.0, Σ_k

is also a correlation matrix for $\tilde{\mathscr{T}}_k$. In Appendix A of Supplementary Materials, we prove that the multivariate distribution of $\tilde{\mathscr{T}}_k$ is a mean zero Normal distribution with elements $\sigma_{k_1k_2}$ of its covariance matrix Σ_k that can be estimated with

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

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, and are estimated here using data available at $s = s_{k_2}$. We also define $\hat{\psi}_g(s_{k_1}, s_{k_2}) = n_g(s_{k_1})/n_g(s_{k_2})$ and

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

So that we replace 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 $\bar{z}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ with $\tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ to obtain $\tilde{\sigma}_g^2(s_{k_1})$, $\tilde{z}_i\{\hat{\mu}_g(s_{k_1}, \tau)\}$ and $\bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1}, \tau)\}$. The purpose of these latter substitutions is to estimate quantities that are not parameterized for a particular analysis time s with the more complete data available at the latter analysis time, s_{k_2}.

Estimation of null hypothesis percentiles involved in critical value calculations can be accommodated using either numerical integration techniques applied to the joint null hypothesis distribution or simulation techniques based on multivariate replicates from this joint distribution. For instance, suppose an OF spending function is chosen with spending function parameters $(\gamma_1, \ldots, \gamma_{k-1})$ at analysis times $(s_1, \ldots, s_{k-1})$. At analysis time s_k, the upper critical boundary, $c_U(s_k)$, is based on the $1 - \frac{\alpha_{OF}(\gamma_k) - \alpha_{OF}(\gamma_{k-1})}{1-\alpha_{OF}(\gamma_{k-1})}$ percentile of the null hypothesis conditional distribution of $|\tilde{\mathscr{T}}(s_k)|$ given critical boundaries were not crossed at prior interim analyses by $\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_{k-1})$. For symmetric critical boundaries we use $c_L(s_k) = -c_U(s_k)$.

In simulation and example sections of this manuscript, critical boundaries are simulated. In particular, for critical values at analysis time s_k, we generate $H = 1$ million mean zero multivariate normal iterates, $\{Z_h(s_1), \ldots, Z_h(s_k)\}$, $h = 1, \ldots, H$, with correlation (covariance) matrix Σ_k. Among the subset, $S(s_{k-1})$, of these iterates that fail to reject the null hypothesis at previous analyses from s_1 to s_{k-1}, we estimate $c_U(s_k) = -c_U(s_k)$ with the $1 - \frac{\alpha_{OF}(\gamma_k) - \alpha_{OF}(\gamma_{k-1})}{1-\alpha_{OF}(\gamma_{k-1})}$ percentile of $|Z_h(s_k)|$. In our simulations, $H = 1$ million successfully estimated the very small percentiles used by the OF spending function.

5 | SIMULATIONS

Simulations were conducted to compare operating characteristics in the group sequential setting for (1) the Tayob and Murray[1] (TM) test using our proposed methodology with $\tau = 12$ months, (2) the Cook and Lawless[2] (CL) cumulative mean test and (3) a logrank (LR) analysis of the first time-to-event. Each tabulated result is based on 1000 iterations of the simulation approaches described below.

We assume a 48-month clinical trial with annual interim analyses scheduled at $s = \{12, 24, 36, 48\}$ months from the start of the study. One hundred patients per treatment group are enrolled, half at baseline, with the remainder accrued uniformly over the first 24 months. Participants are administratively censored according to the analysis time, with no additional loss-to-follow-up otherwise. An O'Brien-Fleming (OF) type I error spending function is used to determine group sequential stopping rules with an overall type I error of 0.05, where the spending function parameter, γ, was taken to be the proportion of calendar time used by analysis time s of the planned 48 months.

Within each patient, we generate a dependence structure between events using a Gaussian copula approach[21]. This approach induces correlation between gap times $T_{ij} - T_{ij-1}$ for $j = 2, \ldots, J_i - 1$ as well as correlation between each gap time and the terminating event T_{iJ_i}. We first simulate mean zero multivariate normal random variables $\{U_{i1}, U_{i2}, \dots, U_{i200}, V_i\}$, with covariance matrix satisfying $Var(V_i) = Var(U_{ij}) = 1$ for $j = 1, \dots, 200$, with ρ_1 parameterizing the correlation between U_{ij} and $U_{ij\prime}$, for $j \neq j\prime$, and ρ_2 parameterizing the correlation between U_{ij} and V_i for $j = 1, \dots, 200$. In addition to the setting with independence between all recurrent and terminal events ($\rho_1 = \rho_2 = 0$), low (0.3), medium (0.5) and high (0.7) values of ρ_1 and ρ_2 are explored. We then use the probability integral transform method to convert the multivariate normal random variables to correlated Uniform(0,1) random variables and then to correlated exponential random variables. The simulated exponentially distributed random variable originating from V_i becomes the terminal event and the remaining exponentially distributed events become gap times between recurrent events, with $J_i - 1$ counting the recurrent events prior to the terminating event for individual i; simulated events that occur beyond the terminal event for a participant are discarded.

For the control group, recurrent events are simulated to occur every 3 months on average, subject to a terminal event with a mean of 36 months. Following the rule of thumb from Xia and Murray (2018)[15] for this control group event rate, follow-up windows for the TM method are initiated every 1.5 months so that $t_1 = 0, t_2 = 1.5, t_3 = 3, t_4 = 4.5, \dots, t_b = s$ months. The experimental group experiences a treatment benefit in terms of both the terminal and recurrent event rates, with recurrent events occurring every 4.3 months on average and a mean time to terminating event of 51.4 months.

Under the null hypothesis, for all group sequentially monitored test statistics and all correlation structures, simulated overall type I error was within expected simulation error of the desired 0.05 level. With independently generated event times, overall type I errors were 0.054, 0.054 and 0.041 for the group sequentially monitored TM, CL and LR statistics, respectively. Table 1 displays overall type I error simulation results assuming different combinations of low, medium and high correlation between an individual's event times.

Cumulative power for detecting the alternative hypothesis at each analysis time, in the special case of independently generated recurrent and terminal event times, is shown in Appendix B Figure S1 of Supplementary Materials. Simulated power for the group sequentially monitored CL statistic (triangles) was highest in this case, followed closely by the TM statistic (circles) and distantly by the LR method (+).

For correlated recurrent and terminal event settings simulated assuming the alternative hypothesis, Figure 4 displays power for group sequentially monitored TM, CL and LR statistics. Panels moving from top to bottom in this figure correspond to increasing levels of correlation between recurrent events in an individual. Panels moving from left to right in this figure correspond to increasing levels of correlation between recurrent and terminal events. For any particular panel, simulated power is displayed on the vertical axis; the horizontal axis is interim analysis time ($s = 12, 24, 36$ or 48 months). For each of these correlation structures, the power of the group sequentially monitored TM statistic approximates or exceeds the power of the CL and LR methods.

The group sequentially monitored logrank test only uses the first time-to-event in each individual, and therefore is not affected by correlation between event times as simulated in the various panels of Figure 4. Because the logrank test's simulated power dynamic is similar from panel to panel of Figure 4, merely reflecting simulation variability across the scenarios, it is helpful in spotting changes in the behavior of the group sequentially monitored TM and CL methods. The power dynamics of these latter group sequentially monitored statistics change according to the degree of statistical information gained from the additionally incorporated recurrent and terminal events.

For the TM statistic, only modest changes in power dynamics are seen within any row of Figure 4, likely because of the small relative role terminal events (4.6-8.1% of simulated events) play in these analyses compared to the role of the recurrent events (91.9-95.4% of simulated events). As correlation between recurrent events increases, the statistical information in the longitudinally constructed censored event times used by the TM method decreases. Hence the power of the TM statistic decreases when moving from top to bottom panels in Figure 4.

The power dynamic of the group sequentially monitored CL statistic is strongly impacted by the correlation structure between events. Whereas in Supplemental Figure S1 (with all independent events), the CL test statistic has the largest power of the methods shown, power for the CL statistic erodes substantially as correlation between recurrent events increases. In the bottom row panels of Figure 4, the LR test outperforms the CL test even though the LR test is only using the first observed event-time per individual. Upon further exploration of the simulated CL test statistics, the explanation for this power dynamic rests in the variability of the number of events per individual that the CL test statistic is built from. The patient to patient variability in the observed number of events increases as the correlation between recurrent events increases, causing the variance of the mean number of cumulative events to increase, and the CL test to lose power. Intuitively, increasing correlation drives the total number of observed events higher in patients with a tendency for short times-to-event. Similarly, increasing correlation drives the total number of observed events lower for individuals with a tendency towards long times-to-event. Taking both of these patterns into account, the range of the observed number of events widens as correlation between events increases.

Panels in the middle row of Figure 4 show power for the CL test improving from the worst of the three methods (in the case with medium correlation between recurrent events and low correlation between recurrent and terminal events) to power nearly identical to the TM method (in the case with medium correlation between recurrent events and high correlation between recurrent and terminal events). Moving left to right the variability in the number of observed events per individual is stabilizing in this row of figures. Those with a tendency towards short times-to-event are experiencing a terminal event before their total count gets very high. Similarly, those with a tendency towards longer times-to-event are experiencing longer times to accumulate these event counts before a terminal event. A similar pattern is observed, to a lesser extent, in the lower right panel of Figure 4, where the power of the CL method increases a bit compared to its power dynamics as shown in panels to its left.

Additional simulation results are available in Appendix C of Supplementary Materials. The special case with a single time-to-event, studied via the censored longitudinal framework, is shown for comparison in Figure S2. Group sequential monitoring in this special case was previously developed by Xia et al.[7]. When considering only single time-to-event group sequential methods, Xia et al.[7] gives a good summary of pros and cons to using the censored longitudinal data paradigm for analyses. When recurrent event data is available, the analyses that use this extra information outperform methods that use only a single time-to-event. In Figure S3, we evaluate the impact of choosing different window lengths, τ, on the performance of our group sequentially monitored test statistic. In this figure, the choice of τ shows minimal impact on study power. We recommend that τ be chosen to give interpretations of interest particular to the research setting where the method is applied. For example, $\tau = 6$ or 12 months would address the average recurrence-free time during those respective lengths of follow-up.

6 | EXAMPLE

The Azithromycin in COPD Trial[22] randomized 1117 patients with a history of acute exacerbations to 250 mg daily of azithromycin or placebo. The original group sequential monitoring plan for this study was based on a logrank analysis of the time-to-first acute exacerbation or death, with few of these events anticipated to be deaths (around 4.4% of enrolled patients at approximately one year of follow-up). Interim analyses were conducted every 6 months with overall type I error for the trial controlled via an O'Brien-Fleming spending function. Conditional power analyses were additionally provided to the Data and Safety Monitoring Committee. To make this example more interesting, we restrict attention to 381 patients accrued during the first year of follow-up. In constructing the TM statistic, we use $\tau = 6$ months and, following Xia and Murray (2018)[15], initiate follow-up windows every 2 months (approximately one third of the historic mean time to exacerbation in this population).

Figure 5 shows the estimated days free of acute exacerbation or death per 6-months of follow-up, based on the TM statistic, at each of the interim analysis times. Group sequential boundaries based on the O'Brien-Fleming spending function are superimposed with an overall type I error of 5%. These boundaries are presented on the scale of the observed effect size needed for the trial to stop early, which can be calculated as $c_U(s)\sqrt{\hat{\sigma}_1^2(s)/n_1(s) + \hat{\sigma}_2^2(s)/n_2(s)}$ for upper bound and for $c_L(s)\sqrt{\hat{\sigma}_1^2(s)/n_1(s) + \hat{\sigma}_2^2(s)/n_2(s)}$ lower bound, where $c_U(s)$ and $c_L(s)$ are critical values for the standardized test statistics as described in Section 4. A recommended stopping boundary for safety with spending function, $\alpha_{JT}(\gamma) = 0.2\gamma^{1.5}$, is superimposed in Figure 5. This boundary is a special case of a Jennison and Turnbull[23] boundary that we have personalized to stop at the first interim analysis if the standardized test statistic exceeds a 1.96 critical boundary in favor of the placebo group. The overall probability of stopping for a safety signal based on this boundary is 20% under the null hypothesis of no treatment effect.

The TM test statistic recommends stopping the trial in favor of the azithromycin arm at the 3rd interim analysis (18 months into the study). For comparison, standardized TM, CL and logrank test statistics and corresponding stopping boundaries are displayed in Figure 6. The CL stops at the 4th interim analysis (2 years into the study) with 59 additional acute exacerbations and 4 additional deaths observed compared to the TM-based group sequential analysis. The logrank analysis of time-to-first event does not detect a significant benefit of azithromycin in this subset of patients from the original study.

7 | DISCUSSION

In this paper, we develop a new nonparametric tool for group sequentially monitoring clinical trials based on recurrent event outcomes subject to a terminal event. Our method is appropriate and robust for events that are correlated within individual or for completely independent event times. Treatment effects observed across analysis times are simple to interpret. In addition to plots showing stopping boundaries based on standardized test statistics, we display observed data and stopping boundaries on the scale of the needed effect size for the trial to stop.

Statistical literature for nonparametric group sequential monitoring of clinical trials is currently dominated by single time-to-event analyses. In the recurrent events setting, many researchers still design their trials using only the first time-to-event because of the availability of software, or in some cases because of concern that strong assumptions are required for recurrent event analyses to be valid.

This, of course, is a shame because (1) there is quite a nice existing nonparametric method for group sequential monitoring of recurrent events data available from Cook and Lawless (1996)[2] that is being under-utilized in clinical trial design in our opinion. This method is also appropriate for correlated events within an individual and performs particularly well when events from the same individual are independent. (2) Clinical trial designs that do not take advantage of events that occur after the first observed event are statistically inefficient, which has financial implications for the overall cost of a clinical trial.

In developing group sequential methodology relating to the Tayob and Murray statistic, we hope to enrich needed literature in this area. Our method performs particularly well when events times within an individual are correlated, and is competitive with the Cook and Lawless method when events are independent.

With continually improving treatments for those with chronic disease, trials are becoming more dependent on surrogate outcomes and combined endpoints rather than mortality alone. Many of these events are recurrent in nature. This trend is likely to continue as lifetimes are successfully extended and as time pressure for faster drug approval increases. We strongly believe that in settings of chronic disease, clinical trial design and analysis should move towards recurrent events methods that incorporate a mixture of disease progression events over time; that this should be the default design choice in understanding a patient's disease burden.

Choice of progression events to include in the (recurrent) composite endpoint must be done with care, as the TM statistic currently gives similar importance to each type of progression when assessing treatment benefit. A future research direction is how one might take into account the very real scenario where progression events differ in severity. Of course, in settings where serious mortality rates continue to be seen, mortality analyses should continue as the preferred analysis for making therapeutic recommendations. However, in settings where mortality is less commonly seen, our method is an attractive choice for pursuing therapies that extend time free from progression events.

In any clinical trial setting, the treatment effect may change over time; this is a common scenario that a group sequential monitoring tool should be able to address. The TM statistic is based on restricted means with no expectation that the treatment differences will follow any particular parametric or semiparametric pattern over time. In the original Tayob and Murray (2014)[1] article, simulations show how the statistic performs when there are delayed treatment effects, short-term treatment effects and Weibull distributed crossing hazards. Supplementary materials of the Tayob Murray article (Figure 2) show how one might graphically display $\tau-$restricted means over the follow-up period when there are crossing hazards. These graphical displays could be provided to data monitoring committees along with the group sequential boundaries we provide in this manuscript.

We end this manuscript with a reminder that when reporting an observed treatment effect or confidence interval from a group sequentially monitored trial that stopped early for perceived benefit, the estimate of treatment benefit from the trial may be inflated in finite samples. While several authors have pursued bias correction methods, for example Tsiatis et al. (1984)[24], Whitehead (1986)[25], Emerson and Fleming (1990)[26], Liu and Hall (1999)[27], Molenberghs et al. (2014)[28], others have argued that higher mean squared error of bias-corrected estimates make bias correction unappealing (and unnecessary) in practice[29], particularly if using a conservative O'Brien-Fleming type I error spending approach.

ACKNOWLEDGEMENTS

The authors would like to thank the COPD Clinical Research Network, in particular our University of Michigan collaborator, Meilan Han, for the use of their data. *Conflict of Interest*: None declared.

DATA AVAILABILITY STATEMENT

Statistics in Medicine encourages authors to share the data and other artefacts supporting the results in the paper by archiving it in an appropriate public repository. The datasets generated during and analysed during the current study are available online at https://github.com/summerx0821/Nonparametric-GS-Methods-for-Recurrent-Terminal-Events. Software in the form of R code is also available at this website. Shared data should be cited.

References

1. Tayob N, Murray S. Nonparametric tests of treatment efffect based on combined endpoints for mortality and recurrent events. *Biostatistics* 2014.
2. Cook RJ, Lawless JF. Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics* 1996; 52: 1311–1323.
3. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* 1982; 77: 855-861.
4. Murray S, Tsiatis AA. Sequential Methods for Comparing Years of Life Saved in the Two-Sample Censored Data Problem. *Biometrics* 1999; 55(4): 1085–1092.
5. Li Z. A group sequential test for survival trials: an alternative to rank-based procedures. *Biometrics* 1999; 55(1): 277–283.
6. Logan BR, Mo S. Group sequential tests for long-term survival comparisons. *Lifetime data analysis* 2015; 21(2): 218–240.
7. Xia M, Murray S, Tayob N. Nonparametric Group Sequential Methods for Evaluating Survival Benefit from Multiple Short-Term Follow-up Windows. *Biometrics* 2018.
8. Ashley SL, Xia M, Murray S, O'Dwyer DN, Grant E. Six-SOMAmer Index Relating to Immune, Protease and Angiogenic Functions Predicts Progression in IPF. *PloS One* 2016.
9. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat* 1982.
10. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of Royal Statistical Society: Series B* 2000; 62(4): 711-730.
11. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika* 1981; 68: 373–79.
12. Ghosh D, Lin D. Nonparametric analysis of recurrent events and death. *Biometrics* 2000; 56(2): 554-562.
13. Cook RJ, Yi GY, Lee KA. Sequential Testing with Recurrent Events over Multiple Treatment Periods. *Stat Biosci* 2010; 2: 137–153.
14. Jiang W. Group sequential procedures for repeated events data with frailty. *J Biopharm Stat* 1999; 9: 379–399.
15. Xia M, Murray S. Commentary on Tayob and Murray (2014) with a useful update pertaining to study design. *Biostatistics* 2018.
16. Pocock S. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; 64(2): 191-199.
17. O'Brien P, Fleming T. A Multiple Testing Procedure for Clinical Trials. *Biometrics* 1979; 35: 549–556.
18. Gordon Lan K, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; 70(3): 659–663.
19. DeMets DL, Lan KKG. Interim Analysis: The Alpha Spending Function Approach. *Statistics in Medicine* 1994; 13: 1341-1352.
20. Lan KG, DeMets DL. Group sequential procedures: calendar versus information time. *Statistics in Medicine* 1989; 8(10): 1191–1198.
21. Li DX. On default correlation: A copula function approach. 1999.

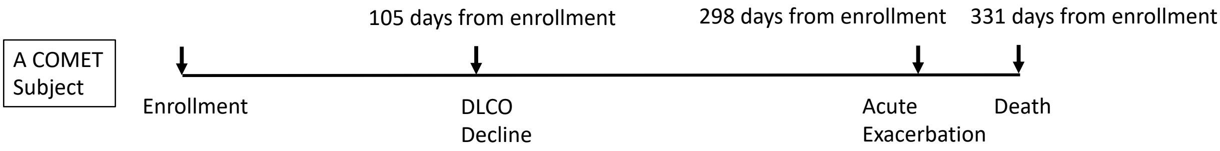

FIGURE 1 An Example IPF Patient from COMET Study.

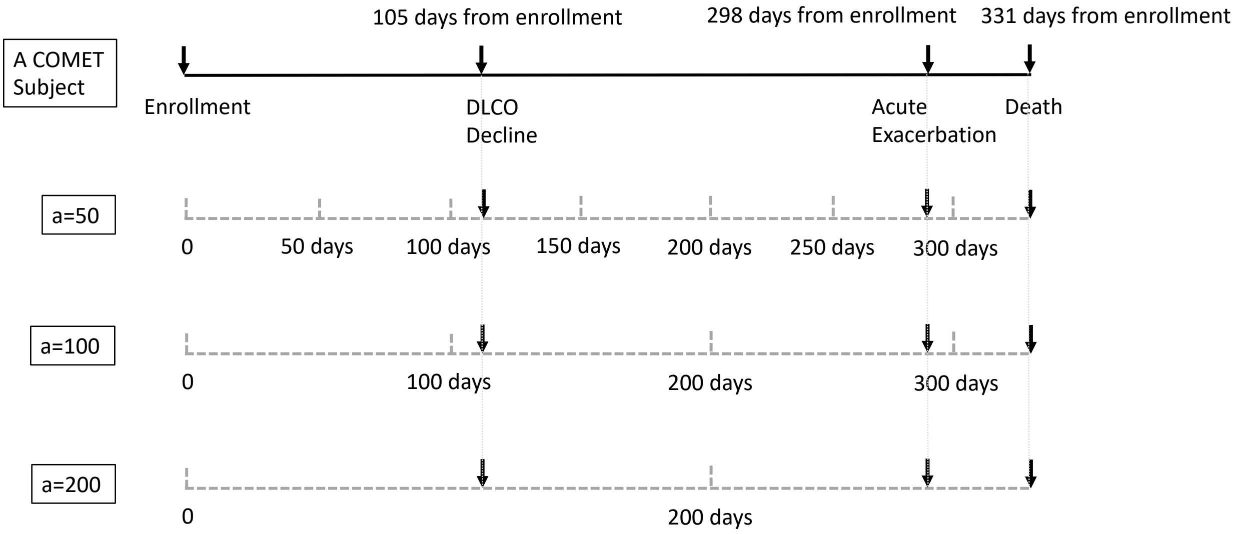

FIGURE 2 An Example IPF Patient from COMET Study with Different Setups of Follow-up Windows ($a = t_j - t_{j-1}, j = 2, \dots, b,$).

22. Albert RK, Connett J, Bailey WC, et al. Azithromycin for prevention of exacerbations of COPD. *New England Journal of Medicine* 2011; 365(8): 689–698.

23. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall . 2000.

24. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984: 797–803.

25. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986; 73(3): 573–581.

26. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; 77(4): 875–892.

27. Liu A, Hall W. Unbiased estimation following a group sequential test. *Biometrika* 1999; 86(1): 71–78.

28. Molenberghs G, Kenward MG, Aerts M, et al. On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research* 2014; 23(1): 11–41.

29. Milanzi E, Molenberghs G, Alonso A, et al. Estimation after a group sequential trial. *Statistics in biosciences* 2015; 7(2): 187–205.

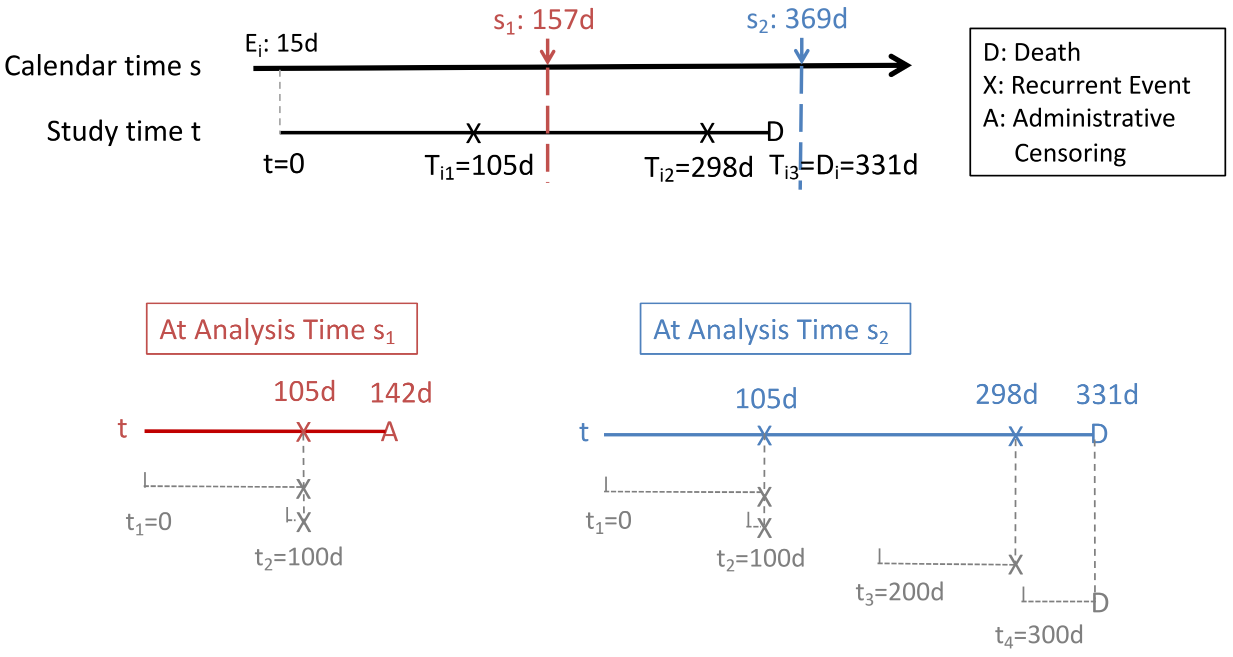

FIGURE 3 Notation for An Example Individual, with Random Variables Given in Detail.

FIGURES AND TABLES

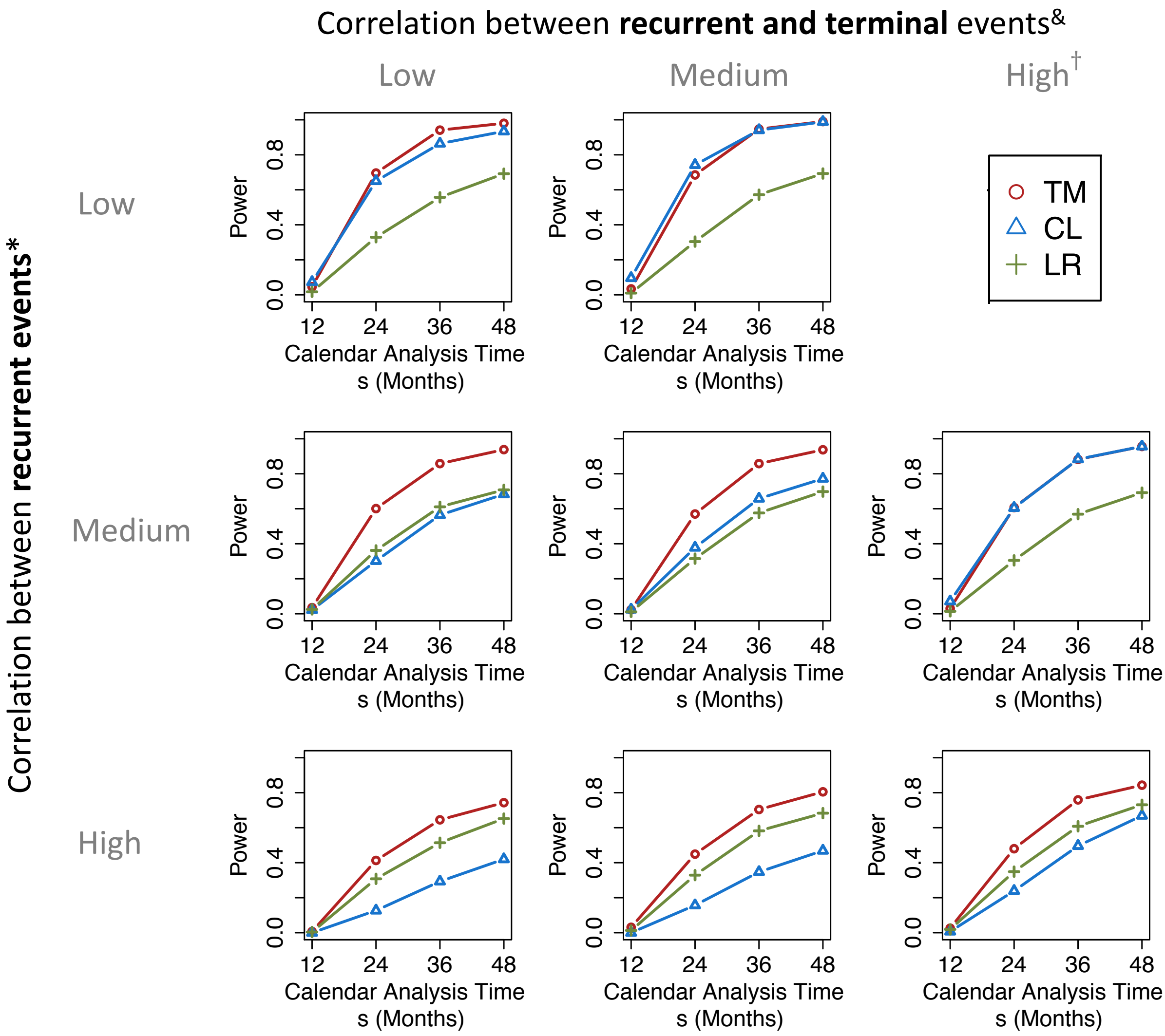

FIGURE 4 Cumulative Power at Each Analysis Time by Varying Levels of Correlation Between Recurrent Events (Rows) and Correlation between Recurrent and Terminal Events (Columns).

(TM: Tayob and Murray (2014) test; CL: Cook and Lawless (1996) test; LR: log-rank test.)

† Data is not shown for the case with low ρ_1 and high ρ_2 since this covariance structure was difficult to construct. Intuitively, it is difficult to have gap times weakly correlated with one another and at the same time all highly correlated with the terminal event time.

* Low, medium to high correlations between recurrent events are generated from $\rho_1 = 0.3, 0.5$ and 0.7, respectively.

& Low, median to high correlations between recurrent and terminal events are generated from $\rho_2 = 0.3, 0.5$ and 0.7, respectively.

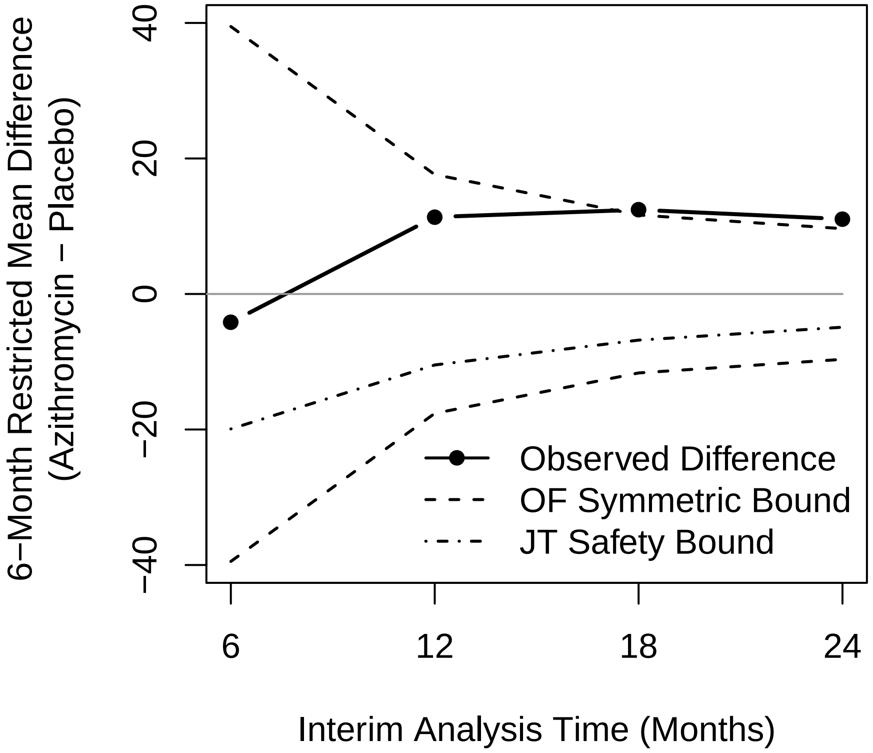

FIGURE 5 Additional Days Free of Acute Exacerbation or Death per 6-months of Follow-up When Using Azithromycin versus Placebo, Based on the Tayob and Murray (2014) (TM) Statistic. (OF: O'Brien-Fleming; JT: Jennison and Turnbull)

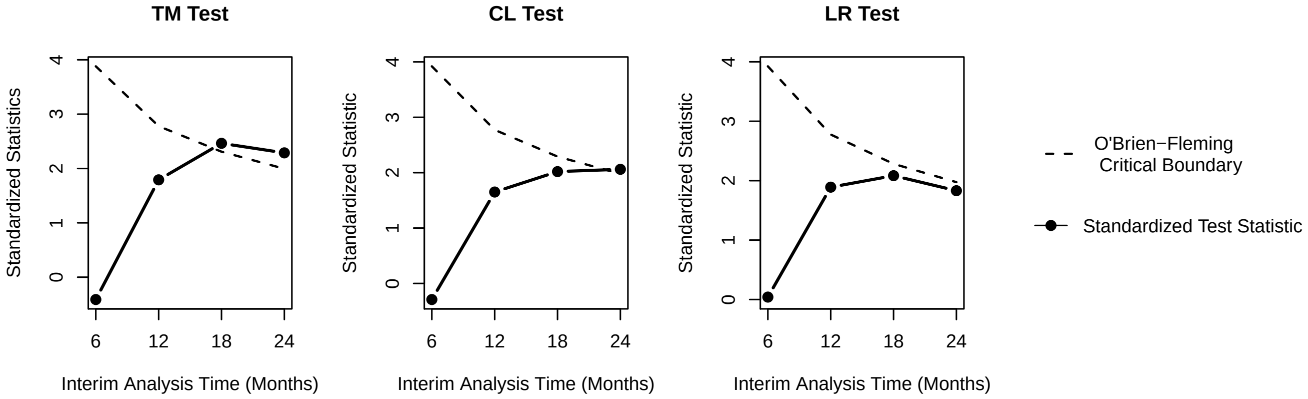

FIGURE 6 Standardized Test Statistics and Critical Boundaries for NACT Example. (Lower (symmetric) O'Brien-Fleming boundary and Jennison & Turnbull (JT) boundary are not displayed. TM: Tayob and Murray (2014) test; CL: Cook and Lawless (1996) test; LR: log-rank test.)

TABLE 1 Overall type I error by varying levels of correlation between recurrent events (rows) and correlation between recurrent and terminal events (columns).

			Correlation between recurrent and terminal events&		
		Test	**Low**	**Medium**	**High**
Correlation between recurrent events*	**Low**	TM	0.053	0.050	NA†
		CL	0.050	0.050	
		LR	0.048	0.041	
	Medium	TM	0.055	0.058	0.044
		CL	0.039	0.045	0.038
		LR	0.051	0.055	0.057
	High	TM	0.058	0.056	0.051
		CL	0.040	0.048	0.045
		LR	0.054	0.048	0.058

(TM: Tayob and Murray (2014) test; CL: Cook and Lawless (1996) test; LR: log-rank test.)

† Data is not shown for the case with low ρ_1 and high ρ_2 since this covariance structure was difficult to construct. Intuitively, it is difficult to have gap times weakly correlated with one another and at the same time all highly correlated with the terminal event time.

* Low, medium to high correlations between recurrent events are generated from $\rho_1 = 0.3, 0.5$ and 0.7, respectively.

& Low, median to high correlations between recurrent and terminal events are generated from $\rho_2 = 0.3, 0.5$ and 0.7, respectively.