Study Design and Analysis of Censored Longitudinal Time-to-Event and Recurrent Event Data

by

Meng Xia

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Professor Susan Murray, Chair
Professor Bethany B. Moore
Professor Jeremy M.G. Taylor
Associate Professor Min Zhang

Meng Xia

summerx@umich.edu

ORCID iD: 0000-0002-9711-6215

For my mother, Jing Su. Thank you for your unconditional love.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Susan Murray, who delights my PhD life and help me make it through. There were thousands of times that I regretted my decision to pursue a PhD in Biostatistics, however, Susan is my only reasons to persist. You are the best advisor. You have been extremely patient with me in guidance and you always have a solution to lead me out of difficulties. Every time when I got frustrated and upset from my work, you gave me a comforting hug, without a word I know that I'm not that terrible and I've been cared. You keep telling culture-related jokes that for most times I don't understand, but I can't help laughing because I feel your humor and happiness. You are not only my advisor of Biostatistics, but also my mentor of American cultures. I treasure those times that we work together as officemates, discussing statistical questions while dragging socks from your puppy. You may not be the best host, but you are the best advisor ever.

I would like to thank all our collaborators in the pulmonary division in University of Michigan Hospital. Dr. Beth Moore, who kindly agrees to be my committee member, shows me the scientific thinking from the perspective of clinical and biological consideration. I really enjoyed working with you and Shanna Ashley, Dave O'Dwyer, Xiaofeng Zhou these years, and I'm so proud of the research questions that we have conquered together. I also appreciate the patience of Dr. Kevin Flaherty, Margaret Salisbury and Eric White when I just started the GSRA work. You tolerated me as a freshman and helped me to become better. I would also like to thank Dr. Meilan Han and Dr. Vibha Lama who provided me with brilliant projects to work on and

supported my education as well as practice. And many thanks to Michael Combs, who helps me a lot in quality check of analysis, and other doctor friends, Wassim Labaki, Jamie Sheth, Beth Belloli, and Bonnie Wang.

I would also like to thank my dissertation committee member, Dr. Jeremy Taylor and Dr. Min Zhang for your constructive input and precious time on this research. Besides, I must also thank that Dr. Jeremy Taylor shared me with a lot of related research topics and Dr. Min Zhang taught me advanced survival analysis which laid a base of my dissertation work. Many thanks to all the faculty and staff members in the Biostatistics Department for the guidance and assistance, to all my friends for our great moments together in Ann Arbor, and to my beloved University of Michigan. Go BLUE forever!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation pursues a paradigm shift from traditionally recorded censored time-to-event and time-to-recurrent event data and corresponding analyses. Instead these data are repurposed into censored short-term outcomes measured longitudinally over potentially overlapping follow-up periods of length τ. Previous work by *Tayob and Murray* (2014, 2016, 2017) exploited this framework, with univariable and multivariable methods for estimating behavior of τ-restricted outcomes drawn from a single time-to-event and with a two-sample test developed for comparing censored longitudinal outcomes drawn from the recurrent events setting. This thesis considers three practical settings that can benefit statistically from repurposing traditional data into this censored longitudinal data structure.

Chapter II addresses the first research setting. This chapter develops a two-sample test and corresponding group sequential methodology for comparing overall τ-restricted means between treatment groups, where each patient contributes many overlapping τ-length follow-up windows of information during the course of the clinical trial. Operating characteristics explored through simulation compare favorably with existing nonparametric methods for group sequentially monitored test statistics in this setting, including the traditional restricted mean test and the logrank test. The proposed method performs especially well in cases where there is a delayed treatment effect and/or a subset of cured patients. This chapter considers symmetric and asymmetric error spending approaches and makes recommendations for how to choose appropriate group sequential stopping boundaries in a variety of settings.

Chapter III addresses the second research setting. Very few methods are currently available for group sequential analysis of recurrent events data subject to a terminal event in the clinical trial setting. Chapter 3 helps fill this gap by developing methods for sequentially monitoring the nonparametric, two-sample *Tayob and Murray* (2014) (TM) statistic. This chapter briefly reviews the TM statistic, develops and describes how to use the proposed group sequential analysis methods, and through simulation compares its operating characteristics with those of *Cook and Lawless* (1996), as well as a time-to-first-event analysis based on the logrank test. Our advantages include high power to detect treatment differences when there is correlation between event times in an individual and elegantly avoiding dependent censoring bias.

One important component of using the TM statistic, as well as Chapter III methodology for group sequential monitoring, is to wisely construct the censored longitudinal data framework for the recurrent events. Chapter IV formalizes the corresponding guidance. A useful metric, the expected proportion of recurrent events captured as the first event in at least one follow-up window, is derived, and operating characteristics of the TM statistic are summarized. For design and analysis purposes, we formulate recommendations based on the special case with independent exponentially distributed gap times.

Chapter V develops multivariable restricted time models appropriate for analysis of recurrent events data, where data is repurposed into censored longitudinal outcomes in τ-length follow-up windows. This chapter develops two approaches for addressing the censored nature of the outcomes: a pseudo-observations (PO) approach and a multiple imputations (MI) approach. Each of these approaches allows for complete data methods, such as generalized estimating equations, to be used for the analysis of the newly constructed correlated outcomes. Through simulation, this chapter assesses the performance of the proposed PO and MI methods. Both PO and MI approaches show attractive results with either correlated or independent gap times.

CHAPTER I

Introduction

Restricted mean survival time methodology for censored survival data has grown in popularity as an alternative to hazard-oriented methods. It has advantages of free from assumption of proportionality and clinically meaningful interpretation. This dissertation work evaluates the restricted time after repurposing the traditional recorded time-to-event or recurrent event data into censored longitudinal outcomes from overlapping follow-up windows. Previous work by Tayob and Murray explored the use of repeated and overlapping follow-up windows to supplement restricted mean estimation (*Tayob and Murray*, 2016) and regression analysis (*Tayob and Murray*, 2017) in single time-to-event setting, as well as two-sample testing in recurrent event setting subject to terminal events and censoring (*Tayob and Murray*, 2014). However, these methods are only ready for single time analysis. In this dissertation, we develop group sequential methods for two-sample tests with a similar technique for constructing follow-up windows in (Chapter II) the standard censored survival endpoint setting as well as (Chapter III-IV) the recurrent events setting. In these chapters, group is the only covariate. For the setting with multiple, possibly time-dependent covariates and recurrent event outcomes, we develop multivariable restricted time regression methodology using multiple follow-up windows (Chapter V).

Throughout this dissertation, we use repeated and overlapping follow-up windows

to improve the restricted mean survival time estimation and to extend the restricted time methodology to recurrent event setting. The idea is to repurpose the data into a regularly spaced longitudinal form. The endpoints are based on τ-length follow-up windows that start at evenly spaced times. In each of these follow-up windows, the observed endpoint is the time from the beginning of the window to the first event that occurs in that window, or τ if no event occurs during that window. These endpoints are subject to the usual independent right censoring that occurs in the clinical trial setting. The restricted time estimated from the longitudinal censored survival endpoints can be readily interpreted as the survival time or time free from recurrent events over the next τ-length period. The follow-up windows are constructed upon each patient's entry and are universal across all enrolled patients. Specification of the follow-up windows is predetermined by investigators at the phase of study design: the length of windows, τ, is usually decided by the clinical interest and the spaced time between adjacent windows is determined by testing/modeling efficiency which will be discussed more in Chapter IV.

In Chapter II, we take a fresh look at group sequential methods applied to two-sample tests of standard censored survival data and proposes an alternative method of defining and evaluating treatment benefit. Our method repurposes traditional censored event time data into a sequence of short-term outcomes taken from the (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, to ensure an attractive safety profile while allowing for additional follow-up of auxiliary data for future research use. Recommendations for how to choose proper group sequential stopping boundaries are given, with supporting simulations and an example from the AIDS Clinical Trial Group.

In Chapter III, we generalize the group sequential testing procedure to recurrent event setting. Very few methods are currently available for group sequential analysis of recurrent events data subject to a terminal event in the clinical trial setting. Our research helps fill this gap by developing methods for sequentially monitoring the nonparametric, two-sample *Tayob and Murray* (2014) statistic. Advantages of the Tayob and Murray statistic include a high power to detect treatment differences when there is a correlation between recurrent event times or between recurrent and terminal events in an individual. This statistic does not suffer bias from dependent censoring potentially caused by the terminal events, regardless of the correlation between event times in an individual. Nor does the statistic assume the proportionality between groups of the cumulative mean number of events over time. This chapter 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* (1996), which is currently in use, as well as a time-to-first-event analysis using the logrank test. We further illustrate our method using data from the Azithromycin in COPD Trial.

As described previously, how to wisely place the τ-length follow-up windows is predetermined by the investigators and is related to the testing efficiency. Chapter II uses follow-up windows initiated every $\tau/2$ suggested for the special case with a single time-to-event. In terms of the recurrent event setting in Chapter III, we will prefer, intuitively, smaller values of spaced time which will create more follow-up windows

that capture more recurrent events, however, at the cost of computational efficiency. In Chapter IV, we give improved guidance on the choice of the spaced time between adjacent windows. Our recommendation is framed in terms of the average proportion of recurrent events captured in at least one follow-up window for individuals. We study how the average proportion influences our choice of spaced time and therefore power through simulations.

Following the recurrent event setting as Chapter III, the nonparametric two-sample test may not be adequate in the analysis of data where more than one factors are interested and potentially associated with the recurrent event times. In Chapter V, we target on predicting the time free from recurrent events over a prespecified follow-up period with multivariable regression model. We embrace the similar philosophy of transforming the recurrent event data structure into censored longitudinal time-to-first-event outcomes via repeated τ-length follow-up windows. Two approaches are developed to account for censoring issue: pseudo-observations (PO) and multiple imputations (MI). Generalized estimating equation can be then utilized to illustrate the mean structure with complete data and correlated outcomes. We assess the performances of PO and MI methods against the uncensored case by simulation under scenarios of independent gap times or correlated gap times. Both PO and MI approaches show attractive results. We also demonstrate how to apply the proposed methods in the data from Azithromycin in COPD Trial.

Before we start the next chapter, to make sure that all readers are on the same page, we briefly review some concepts and related context for future use.

1.1 Restricted Mean Survival Time

The restricted mean survival time (RMST) was first proposed by *Irwin* (1949) since the mean survival time is not estimable in the presence of censoring. Under the standard survival setting where we use T as the time to an event, a τ-restricted

mean survival time, denoted by $\mu(\tau)$, is the mean of time to events truncated by a prespecified time τ, i.e. $\mu(\tau) = E[min(T, \tau)]$. Let $S(t)$ indicate the survival function of T. The RMST is also the area under the survival curve of T up to time τ, thus $\mu(\tau)$ can be estimated well with the area under the corresponding Kaplan-Meier estimator or Nelson-Aalen estimator from time zero to τ:

$$\hat{\mu}(\tau) = \int_0^\tau \hat{S}(t)dt.$$

Analysis based on the RMST includes two-sample tests of the difference in RMST between two groups (*Pepe and Fleming*, 1989; *Karrison*, 1997; *Zhao et al.*, 2016) and modified two-sample tests to be adjusted by covariates (*Karrison*, 1987, 1997; *Zucker*, 1998), to be applied in group sequential design (*Murray and Tsiatis*, 1999; *Li*, 1999b) or to fulfill other particular requirements (*Zhao and Tsiatis*, 1997; *Chen and Tsiatis*, 2001; *Schaubel and Wei*, 2011). Regression model based on pseudo-observations is proposed by *Andersen et al.* (2004) and then generalized to be applied in more scenarios with special data structures (*Klein and Andersen*, 2005; *Andersen and Klein*, 2007; *Andrei and Murray*, 2007; *Graw et al.*, 2009; *Xiang and Murray*, 2012; *Nicolaie et al.*, 2013; *Tayob and Murray*, 2017). Besides the pseudo-observations approach, multivariable regression analysis on RMST are also developed with other strategies accounting for censoring, for example, with multiple imputations (*Liu et al.*, 2011; *Xiang et al.*, 2014; *Tayob and Murray*, 2017), with inverse probability weighting (*Zhang and Schaubel*, 2011, 2012; *Tian et al.*, 2013; *Wang and Schaubel*, 2018), parametric modeling (*Royston and Parmar*, 2011), etc.

The RMST analysis has many attractive properties. First, the estimation and comparison of RMST is valid under any distribution of time to event and does not require any assumption of proportionality between treatment groups in the clinical trial setting. It can summarize the difference in survival when the difference between

groups is not constant over time, for example cases where there is a delayed treatment effect or a crossing hazard. Besides, we think this measure has a relatively meaningful interpretation for both clinicians and patients. It directly illustrates the gain or loss of life expectancy. A certain hazard ratio reflects no information about the actual progression status of the patient cohort, while the RMST can provide both absolute and relative measures of risk. Therefore, more and more literature suggest moving beyond the hazard ratio in quantifying the survival difference and recommend RMST methodology as a default alternative option in the clinical trial design and analysis (*Royston and Parmar*, 2013; *Uno et al.*, 2014; *Kim et al.*, 2017; *Calkins et al.*, 2018).

In this dissertation, we take advantage of the appealing properties of RMST analysis and add to the field a new two-sample test with favored operating characteristics and a modified version for group sequentially monitoring. We also extend the restricted time analysis to the setting with recurrent events and work on the restricted time free from recurrent events. Besides providing a group sequential test for monitoring the recurrent event data subject to terminal events and censoring, we generalize the regression with pseudo-observation and multiple imputation techniques for single time-to-event data into the setting with recurrent events.

1.2 Group Sequential Design

Group sequential monitoring has a long and respected history in clinical trial design. It is a study design where data are evaluated as they are collected. Further sampling is stopped in accordance with pre-defined stopping rules. Multiple reasons can cause an early stop: 1) significant efficacy has been detected at an interim analysis, in which case the investigators would like to stop the current trial and move to the next stage so that the new treatment can be put into market and help the patients as soon as possible; 2) perceived harm of the investigational treatment goes out of tolerance, in which case it is morally unacceptable to keep the assigned patients in

high risk or to keep recruiting more patients to the harmful treatment group; 3) evidence has shown that there is barely any chance of getting a significantly superior results by the end of the study, in which case financial and human resources can be saved by stopping the trial early.

Based on the many obvious benefits, group sequential design becomes popular in clinical trials, which is also why we think it is worthwhile to develop methods that can be implemented in group sequential context. However, a statistical price must be paid for this monitoring process. By repeatedly looking at the data and conducting significance testing, we inflate the type I error probability beyond the desired level. As a remedy, statistical methods controlling type I error throughout a trial have been developed by many classic group sequential researches (*Pocock*, 1977; *Lan and DeMets*, 1983; *Tsiatis*, 1981, 1982). To maintain the overall type I error despite the repeated significance tests, we have to adjust the critical values $c_1, \ldots, c_K$ of the K number of planned analysis such that

$$Pr\{|\mathscr{T}_1| < c_1, \ldots, |\mathscr{T}_K| < c_K\} = 1 - \alpha$$

under the null hypothesis, where $\mathscr{T}_k$ is the test statistic based on data cumulated up to the k^{th} interim analysis. Given the critical values, the exit probability is defined as

$$Pr\{|\mathscr{T}_1| < c_1, \ldots, |\mathscr{T}_{k-1}| < c_{k-1}, |\mathscr{T}_k| > c_k\} = \pi_k$$

with the restriction $\sum_{k=1}^{K} \pi_k = \alpha$ (*Harrington*, 2012). Under this guidance, *Lan and DeMets* (1983) proposed alpha spending function, which is one of the most commonly used approaches, to provide options of how to spend the type I error, namely, how to distribute α to $\pi_1, \ldots, \pi_K$. Given the pre-specified $\pi_1, \ldots, \pi_K$ and joint distribution of $\mathscr{T}_1, \ldots, \mathscr{T}_K$ under the null hypothesis, we are able to determine the critical values to maintain the type I error probability at the level of α. More application based on the

alpha spending function to determine symmetric or asymmetric stopping boundaries will be discussed in this dissertation.

1.3 Recurrent Event Data

In clinical trial and medical research, investigators are often interested in studying processes which generate events repeatedly over time. Such processes are referred to as recurrent event processes and the data they provide is called recurrent event data (*Cook and Lawless*, 2007). Examples include recurrence of acute exacerbation of patients with chronic obstructive pulmonary disease (*Albert et al.*, 2011), recurrent ischemic cardiovascular events after acute coronary syndrome (*Schwartz et al.*, 2018), recurrent clostridium difficile infection (*Wilcox et al.*, 2017), repetitive head injuries in high-contact sports (*DeKosky et al.*, 2010), etc.

Poisson and negative binomial count models have been used to analyze recurrent event data per time at risk (*Frome et al.*, 1973; *Lawless*, 1987; *Lambert*, 1992; *Greene*, 1994). Besides, analysis strategies based on event times instead of event counts are favored and developed. The most commonly used methods are those based on the Cox proportional hazards model. *Andersen and Gill* (1982) proposed an extension of the original Cox model, assuming independence between the event gap times within the same patient. *Prentice et al.* (1981) proposed two stratified proportional hazard model considering the order of event and include intensity based on either the time from the beginning of study or gap times. *Wei et al.* (1989) formulated the marginal distribution of event times by Cox model based on time from the beginning of study. *Pepe and Cai* (1993) advocated the use of rate function of recurrence after the first event and *Lawless and Nadeau* (1995); *Lin et al.* (2000) later developed models for the cumulative mean number of events, assuming proportionality on the cumulative means over time. In terms of more complex recurrent event setting, frailty or random effect models were introduced to describe individual patients' heterogeneity or the

correlation between event times (*Aalen and Husebye*, 1991; *Hougaard*, 1995; *Rondeau et al.*, 2007; *Mazroui et al.*, 2013; *Rogers et al.*, 2016).

Another data structure of recurrent event in the presence of terminal events are also commonly seen in clinical research. For example in pulmonary disease studies, clinicians are interested in multiple progression endpoints including recurrent events of acute exacerbation, 10% decline in forced vital capacity and 15% decline in diffusing capacity of the lung for carbon monoxide; as well as terminal events of death and lung transplant. This type of data will be more and more common in future studies because treatments are continually improving and trials are becoming more dependent on surrogate outcomes or combined endpoints rather than mortality alone. This trend is likely to continue as lifetimes are successfully extended and as time pressure for faster drug approval increases. Primary methods are based on the analysis of marginal mean/rate function for the cumulative number of recurrent events with non-parametric (*Cook and Lawless*, 1997; *Ghosh and Lin*, 2000) or semi-parametric approaches (*Cook and Lawless*, 1997; *Ghosh and Lin*, 2002; *Schaubel and Zhang*, 2010).

Although many advanced tools have been developed to deal with the recurrent events in an appropriate manner, time-to-first-event analysis with traditional logrank test or Cox model are still very popular in clinical researches as those we cite at the beginning of the section (*Albert et al.*, 2011; *Schwartz et al.*, 2018; *Wilcox et al.*, 2017). To cater the taste of clinical researchers while making use of data more efficiently, in this dissertation we borrow the philosophy of time-to-first-event analysis but incorporate more than one event by constructing multiple follow-up windows. We believe our work will help the design and analysis in clinical studies, as well as better understanding the patients' disease burden.

CHAPTER II

Nonparametric Group Sequential Methods for Evaluating Survival Benefit from Multiple Short-Term Follow-up Windows

2.1 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*, 1999b; *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 chapter 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 chapter 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 (*Friedman et al.*, 2015; *Harrington*, 2012). These types of bounds also avoid the ethically uncomfortable scenario of trial termination only after statistical proof of increased mortality from the new treatment.

This chapter proceeds with a description of notation in Section 2.2. In Section 2.3 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 2.4; Derivations behind methods in Sections 2.3 and 2.4 are relegated to Appendix A. In Section 2.4, we also review symmetric versus asymmetric stopping boundaries, with a modified recommendation for safety monitoring. Section 2.5 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 2.6 and followed by discussion in Section 2.7.

2.2 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.

2.2.1 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 $X_i(s) = min\{T_i, C_i(s)\}$, 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 2.1 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 2.1 will need to account for potential correlation between these outcomes.

2.2.2 Counting Process Notation and Estimation

For an individual i who has entered the trial by interim analysis time s, $N_i(s,t,u) = I\{X_i(s,t) \leq u, \delta_i(s,t) = 1\}$ and $Y_i(s,t,u) = I\{X_i(s,t) \geq u\}$ 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 2.1, 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 $\{N_A(s_2,t_2,11^+) = 1, Y_A(s_2,t_2,11^+) = 0\}$.

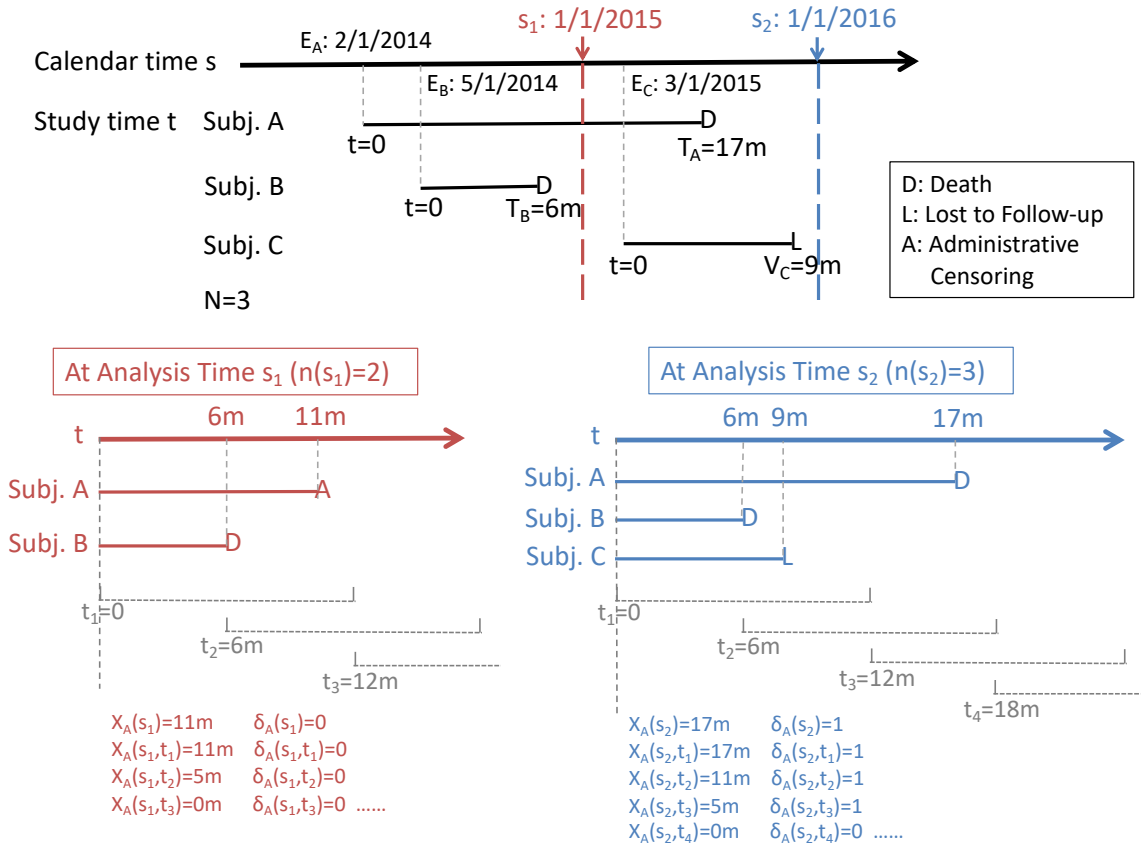


Figure 2.1: Notation for 3 Example Individuals, with Random Variables Specific to Subject A Given in Detail.

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 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.

2.3 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

$$\mathscr{T}(s) = \sqrt{\frac{n_1(s)n_2(s)}{n_1(s)+n_2(s)}}\{\hat{\mu}_1(s,\tau) - \hat{\mu}_2(s,\tau)\}. \tag{2.1}$$

Let $\hat{\pi}_g(s) = n_g(s)/\{n_1(s)+n_2(s)\}, g = 1, 2$. As shown in Appendix A.1, under the null hypothesis of $\mu_1(s,\tau) = \mu_2(s,\tau)$, the asymptotic limiting distribution of $\mathscr{T}(s)$ 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

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

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

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

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

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$, t_3 = 1$ year$, t_4 = 1.5$ years and $t_5 = 2$ years.

2.4 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, $\tilde{\mathscr{T}}(s)$, 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, $\tilde{\mathscr{T}}_K = \left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_K)\right\}$, must be carefully chosen

to preserve an overall type I error of α (*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 $\tilde{\mathscr{T}}_k$. As shown in Appendices A.1 and A.2, $\tilde{\mathscr{T}}_k$ 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

$$\begin{aligned}\hat{\sigma}_{k_1k_2} =& \{\hat{\pi}_2(s_{k_1})\tilde{\sigma}_1^2(s_{k_1}) + \hat{\pi}_1(s_{k_1})\tilde{\sigma}_2^2(s_{k_1})\}^{-\frac{1}{2}}\{\hat{\pi}_2(s_{k_2})\hat{\sigma}_1^2(s_{k_2}) + \hat{\pi}_1(s_{k_2})\hat{\sigma}_2^2(s_{k_2})\}^{-\frac{1}{2}} \\ & \times \sum_{g=1}^{2}\sqrt{\hat{\pi}_{3-g}(s_{k_1})\hat{\pi}_{3-g}(s_{k_2})\hat{\psi}_g(s_{k_1}, s_{k_2})}\Bigg(\sum_{i=1}^{n_g(s_{k_1})}\{n_g(s_{k_1}) - 1\}^{-1} \\ & \times [\tilde{z}_i\{\hat{\mu}_g(s_{k_1}, \tau)\} - \bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1}, \tau)\}]\,[z_i\{\hat{\mu}_g(s_{k_2}, \tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2}, \tau)\}]\Bigg)\end{aligned} \tag{2.2}$$

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 2.3 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

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

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 $\tilde{\mathscr{T}}_k$ are described further in Sections 2.4.1 and 2.4.2. Section 2.4.1 reviews how to calculate CVs based on symmetric type I error spending functions that are in common use. In Section 2.4.2 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.

2.4.1 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 $\alpha_{OF}(\frac{2}{3}) = 0.016$; at the final analysis time $\gamma = 1$ and $\alpha_{OF}(1) = 0.05$ by design. Since under the null hypothesis $\tilde{\mathscr{T}}(s_1)$ has an approximate Normal(0,1) distribution, and no type I error has been spent prior to s_1, $\{c_L(s_1), c_U(s_1)\} = \left\{\mathcal{Z}_{0.016/2}, \mathcal{Z}_{1-0.016/2}\right\}$. Calculation of $\{c_L(s_2), c_U(s_2)\}$ is not as straightforward due to stochastic dependence between $\tilde{\mathscr{T}}(s_1)$ and $\tilde{\mathscr{T}}(s_2)$. 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 ($\mathcal{Z}_{0.016/2} < \tilde{\mathscr{T}}(s_1) < \mathcal{Z}_{1-0.016/2}$) and need to satisfy:

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

Suppose the estimated correlation between $\tilde{\mathscr{T}}(s_1)$ and $\tilde{\mathscr{T}}(s_2)$, 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 $Pr\{\tilde{\mathscr{T}}(s_2) \notin (c_L(s_2), c_U(s_2)) \mid \mathcal{Z}_{0.016/2} < \tilde{\mathscr{T}}(s_1) < \mathcal{Z}_{1-0.016/2}, H_0\} = 0.035$ 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, $\{Z_m(s_1), \ldots, Z_m(s_k)\}$, 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 $c_L(s_k) = -c_U(s_k)$.

2.4.2 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); *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 (*Jennison and Turnbull*, 2000; *Proschan et al.*, 2006; *Demets and Lan*, 1994). 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 2.2 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.

2.5 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 Appendix A.4 we summarize results for our test statistic (2.1) 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

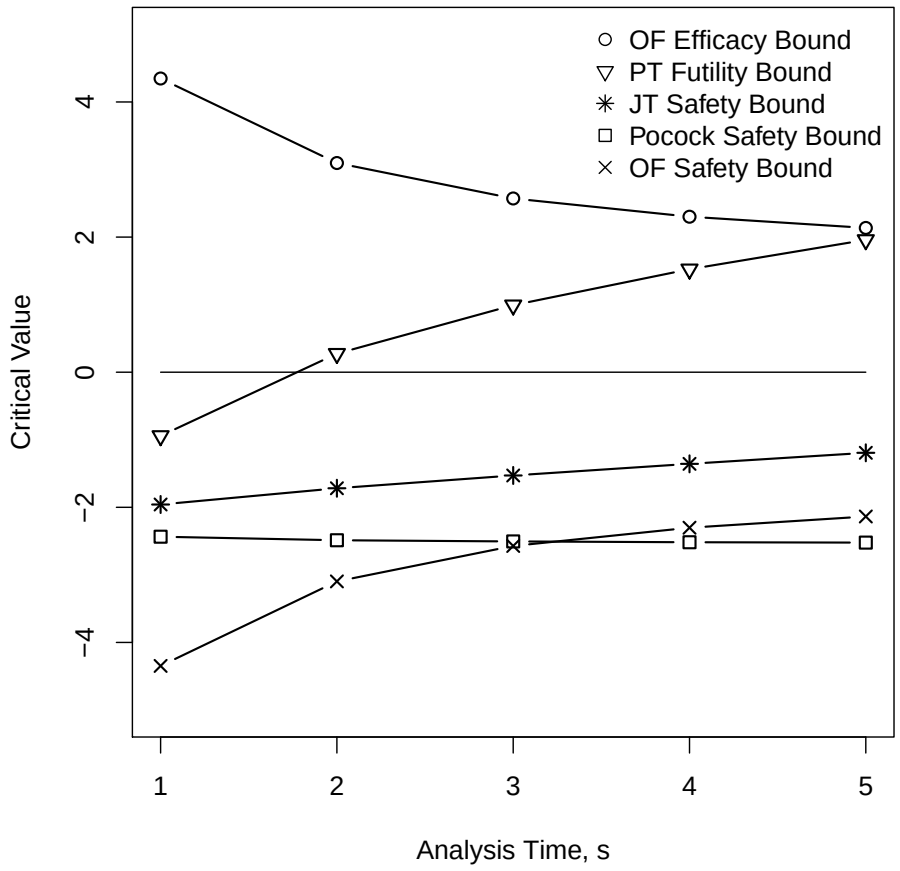


Figure 2.2: Example of Efficacy, Futility and Safety Boundaries
(OF:O'Brien and Fleming; PT: Pampallona and Tsiatis; JT: Jennison and Turnbull)

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 2.3 with piecewise hazards superimposed over the various survival curves, consider proportional hazard alternatives (Scenarios 2-3), delayed treatment effect alternatives (Scenarios 4-5), early treatment differences that attenuate over time (Scenarios 6-7) and alternatives subject to a cure pattern (Scenarios 8-9). Left and right panels of Figure 2.3 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 2.1 and 2.2 summarize group sequential operating characterics in Scenarios 1 though 9. Table 2.1 shows rates of stopping for perceived efficacy, i.e. study power, (column 3) or a perceived safety signal (columns 4-6). Table 2.2 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 1,000 iterations.

Table 2.1, 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

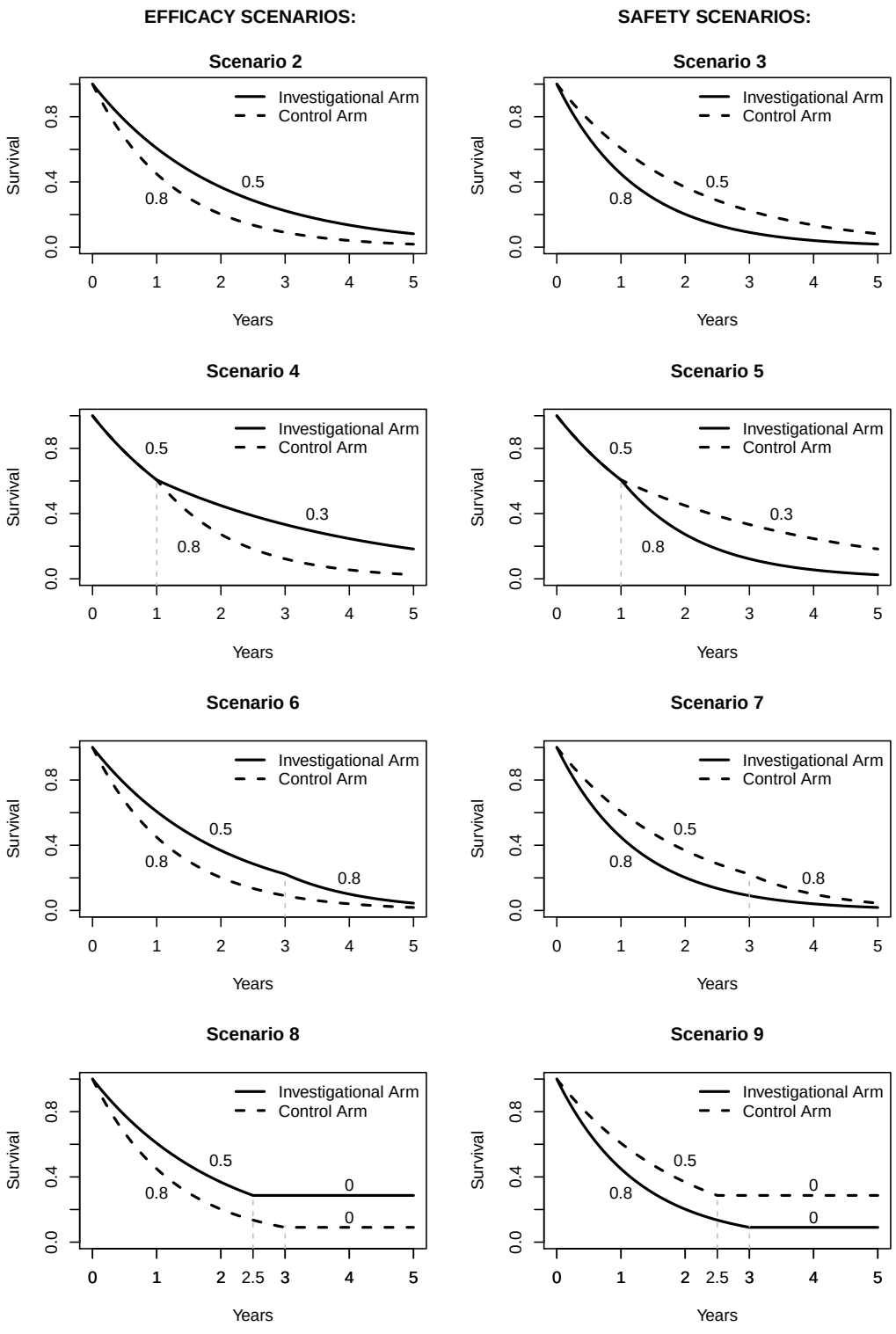


Figure 2.3: Survival Probabilities of the Efficacy and Safety Scenarios with Piecewise Hazards Superimposed over the Curves.

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

Table 2.1: Rates of Stopping for Efficacy (OF Efficacy), i.e. Study Power, or for Safety (JT Safety, P Safety, OF Safety)

† 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).

		AST			ASN			ANE		
Scenario	**Test Statistic**	**JT**	**P**	**OF**	**JT**	**P**	**OF**	**JT**	**P**	**OF**
1	Proposed	4.7	4.9	5.0	195	199	200	156	163	164
	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
2	Proposed	3.7	3.7	3.7	185	185	185	143	143	143
	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
3	Proposed	2.1	3.0	3.6	151	169	184	93	121	142
	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
4	Proposed	3.7	3.8	3.8	187	188	188	132	134	134
	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
5	Proposed	2.8	3.8	3.9	169	185	189	108	132	137
	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
6	Proposed	3.7	3.7	3.7	184	184	184	143	143	143
	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
7	Proposed	2.2	3.2	3.8	153	171	185	96	126	145
	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
8	Proposed	3.4	3.4	3.4	181	181	181	128	128	128
	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
9	Proposed	2.1	3.0	3.5	152	169	182	92	113	130
	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

Table 2.2: Average Study Time (AST) in Years, Average Sample Number (ASN) and Average Number of Events (ANE) in Scenarios 1 - 9

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 2.1) and earlier (AST in Table 2.2) 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 2.2).

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 2.1) and with a much smaller AST, ASN and ANE (Table 2.2). 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 percentage point of simulated power is lost when using the JT safety boundary compared to the other safety boundaries.

For proportional hazards scenarios (Table 2.1, Scenarios 2-3), 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 2.2, likewise, gives very similar AST, ASN and ANE results for the three test statistics.

In Scenarios 4-5, where there is a delayed treatment effect, the proposed statistic has at least 10% higher power (Scenario 4, Table 2.1) with a better safety profile (Scenario 5, Table 2.1) 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 4-5, Table 2.2).

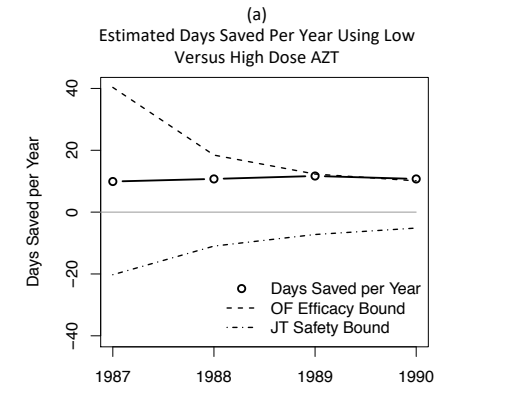
In Scenarios 6-7, where an early treatment difference emerges but becomes attenuated over time, power increases by approximately 2 percentages points when moving from the proposed to the logrank test, and from the logrank to the RMS test (Sce-

nario 6, Table 2.1). Safety profiles, AST, ASN and ANE likewise slightly favor the RMS procedure over the logrank and proposed test, respectively (Scenario 7, Tables 2.1 and 2.2).

In Scenarios 8-9, 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 2.1). Safety profiles shown for Scenario 9 in Table 2.1 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 8-9, Table 2.2).

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 2.4(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. Appendix A.5 summarizes how group sequentially monitored logrank and RMS tests performed in this case. As seen in Figure 2.4(b), there was a delayed treatment effect that perhaps favored our methodology as com-



	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

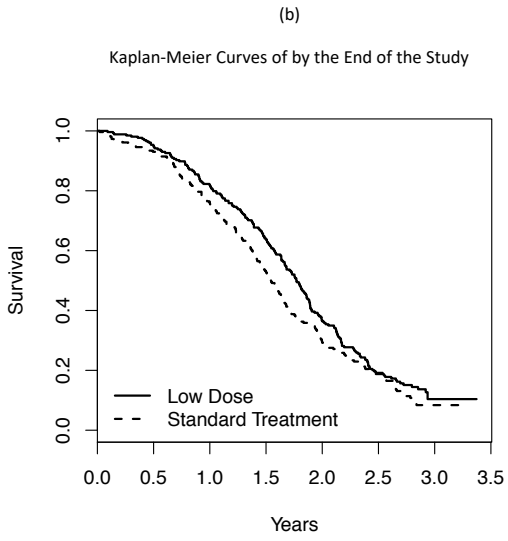


Figure 2.4: Figures in Example: (a) Estimated Days Saved Per Year Using Low Versus High Dose AZT; (b) Kaplan-Meier Curves of by the End of the Study

pared to the logrank and RMS methods. Neither of these competitors recommended stopping before the 1990 analysis time.

2.7 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 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 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 chapter, 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 chapter 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 chapter 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.

CHAPTER III

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

3.1 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 subject to termination by death in these patients. The Azithromycin in COPD Trial (NACT) is one of many clinical trials with follow-up of such events. 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.

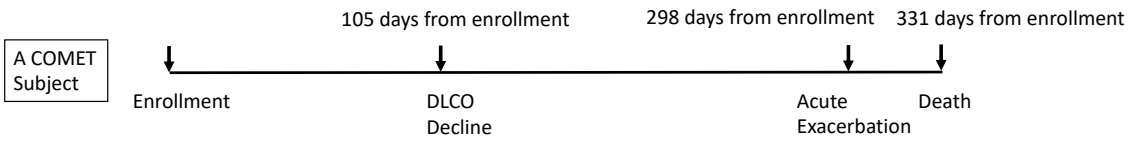


Figure 3.1: An Example IPF Patient from COMET Study.

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. (*Tsiatis*, 1982; *Murray and Tsiatis*, 1999; *Li*, 1999b; *Logan and Mo*, 2015) An obvious disadvantage, however, is the loss of information from ignoring progression events after the first that occurs for each patient. Consider Figure 3.1, 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) (*Ashley et al.*, 2016). 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, *Prentice et al.* (1981), *Andersen and Gill* (1982), *Lin et al.* (2000), *Ghosh and Lin* (2000) or *Tayob and Murray* (2014)), there is little available methodology for conducting group sequential analysis in this setting. *Cook and Lawless* (1996) developed group sequential methods for pseudo-score statistics monitored over time, which are framed to perform well when the cumulative mean number of events are proportional over time. *Cook et al.* (2010) later extended this method to settings with multiple treatment periods. *Jiang* (1999) developed group sequential analysis methods for recurrent events assuming local Poisson processes that allow event rates to change

over time, and incorporated a frailty parameter to address correlation between event times rather than accounting for this correlation nonparametrically.

In this chapter, we develop group sequential methods for monitoring the Tayob and Murray statistic (*Tayob and Murray*, 2014) for nonparametric analysis of recurrent events 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.* (2000) and *Ghosh and Lin* (2000) 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 chapter is organized as follows. Section 3.2 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 3.3 briefly reviews the Tayob and Murray two-sample testing procedure in the case of a single analysis. Section 3.4 extends methodology to the group sequential setting. Section 3.5 describes simulated operating characteristics of our method compared to that of *Cook and Lawless* (1996). We demonstrate the method using data from the Azithromycin in COPD Trial. Discussion follows in Section 3.7.

3.2 Notation

We borrow notation from *Tayob and Murray* (2014), 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 $i = 1, \ldots, N$ 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 $s \geq \max(E_1, \ldots, E_N)$.

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, $\{T_{i1}, T_{i2} - T_{i1}, \ldots, T_{iJ_i-1} - T_{iJ_i-2}\}$, is not constrained to an independent covariance structure.

Traditionally observed data for patients accrued prior to analysis time s is recorded as $X_{ij}(s) = min\{T_{ij}, C_i(s)\}$, $j = 1, \ldots, \tilde{J}_i(s)$ and $\delta_{ij}(s) = I\{T_{ij} \leq C_i(s)\}$, $j = 1, \ldots, \tilde{J}_i(s)$, 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 $t \in \{t_1, t_2, \ldots, t_b\}$ with $t_1 = 0$ and b chosen 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, $\eta_i(s,t) = min\{j = 1, \ldots, \tilde{J}_i(s) : X_{ij}(s) \geq t\}$, 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* (2018) 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, $a = t_j - t_{j-1}, j = 2, \ldots, b$, 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.

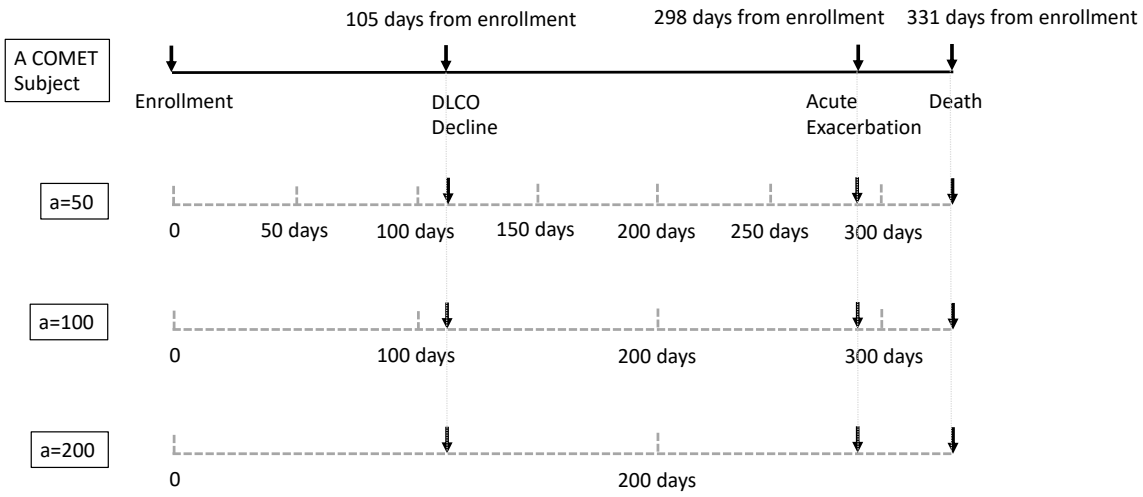


Figure 3.2: An Example IPF Patient from COMET Study with Different Setups of Follow-up Windows

To solidify some of the notation presented above, consider Figures 3.2 and 3.3. In Figure 3.2, 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 3.3, 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

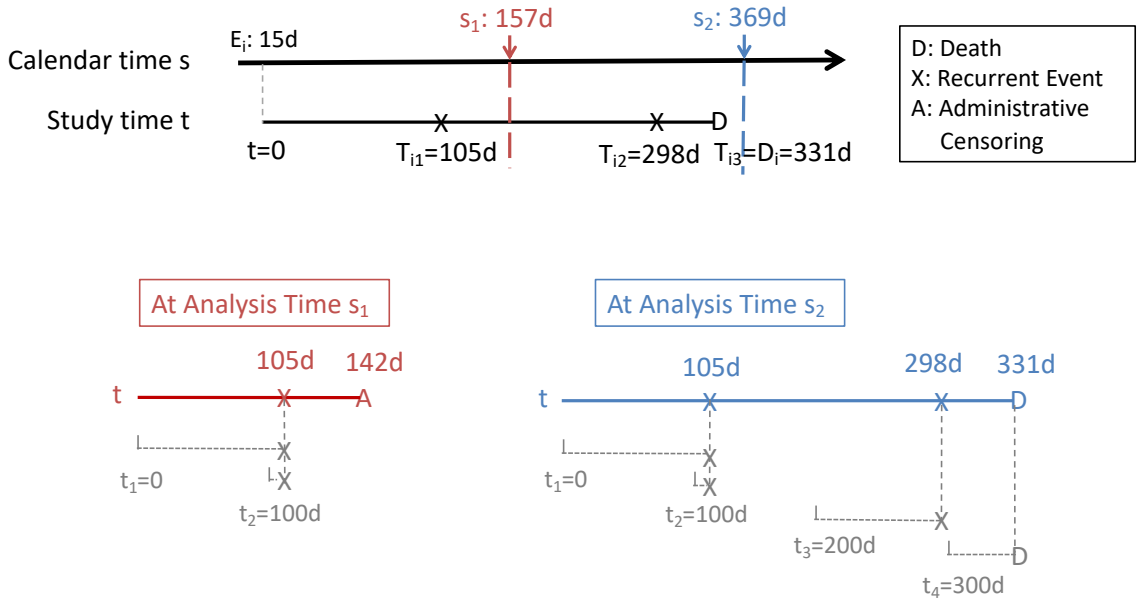


Figure 3.3: Notation for an Example Individual, with Random Variables Given in Detail.

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

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

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, $\lambda^W(s,u)$.

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

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

3.3 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 3.4, we extend these methods to the case where more than one analysis is performed at calendar times $s_1, s_2, \ldots, s_K$ 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 $z_{il}\{\hat{\mu}_g(s,\tau)\} =$

$$\int_0^{\tau} exp\left\{-\int_0^{u_2} \frac{dN_g(s,u_1)}{Y_g(s,u_1)}\right\}\left\{\int_0^{u_2} \frac{dN_{gi}(s,t_l,u_1) - Y_{gi}(s,t_l,u_1)\frac{dN_g(s,u_1)}{Y_g(s,u_1)}}{Y_g(s,u_1)/n_g(s)}\right\} du_2. \quad (3.1)$$

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 $\pm$ 1.96 conferring an overall type I error of 5%.

3.4 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 $\{c_L(s_1), c_U(s_1)\}, \ldots, \{c_L(s_K), c_U(s_K)\}$ 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 (*Pocock*, 1977; *O'Brien and Fleming*, 1979). 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 (*Lan and DeMets*, 1983; *Demets and Lan*, 1994). The O'Brien-Fleming (OF) spending function, $\alpha_{OF}(\gamma) = 2 - 2\Phi(\mathscr{Z}_{1-\alpha/2}/\sqrt{\gamma})$, 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 (*Lan and DeMets*, 1989). 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 B.1 we prove that the multivariate distribution of $\tilde{\mathscr{T}}_k$ is a mean zero Normal distribution with elements $\sigma_{k_1 k_2}$ of its covariance matrix Σ_k that can be estimated with

$$\begin{aligned}\hat{\sigma}_{k_1 k_2} =& \{\hat{\pi}_2(s_{k_1})\tilde{\sigma}_1^2(s_{k_1}) + \hat{\pi}_1(s_{k_1})\tilde{\sigma}_2^2(s_{k_1})\}^{-\frac{1}{2}}\{\hat{\pi}_2(s_{k_2})\hat{\sigma}_1^2(s_{k_2}) + \hat{\pi}_1(s_{k_2})\hat{\sigma}_2^2(s_{k_2})\}^{-\frac{1}{2}} \\ & \times \sum_{g=1}^{2}\sqrt{\hat{\pi}_{3-g}(s_{k_1})\hat{\pi}_{3-g}(s_{k_2})\hat{\psi}_g(s_{k_1}, s_{k_2})}\Bigg(\sum_{i=1}^{n_g(s_{k_1})}\{n_g(s_{k_1}) - 1\}^{-1} \\ & \times [\tilde{z}_i\{\hat{\mu}_g(s_{k_1}, \tau)\} - \bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1}, \tau)\}]\,[z_i\{\hat{\mu}_g(s_{k_2}, \tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2}, \tau)\}]\Bigg)\end{aligned} \tag{3.2}$$

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 3.3, 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 $\hat{\sigma}^2_g(s_{k_1})$, $z_i\{\hat{\mu}_g(s_{k_1},\tau)\}$ and $\bar{z}\{\hat{\mu}_g(s_{k_1},\tau)\}$ with $\tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$ to obtain $\tilde{\sigma}^2_g(s_{k_1})$, $\tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\}$ 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 chapter, 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, $\mathcal{S}(s_{k-1})$, 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.

3.5 Simulations

Simulations were conducted to compare operating characteristics in the group sequential setting for (1) the *Tayob and Murray* (2014) (TM) test using our proposed methodology with $\tau=12$ months, (2) the *Cook and Lawless* (1996) (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. (*Li*, 1999a) 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 $\{U_{i1}, U_{i2}, \ldots, U_{i200}, V_i\}$, with covariance matrix satisfying $Var(V_i) = Var(U_{ij}) = 1$ for $j = 1, \ldots, 200$, 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 $j = 1, \ldots, 200$. 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) for this control group event rate, follow-up windows for the TM method are initiated every 1.5 months so that $t_1 = 0, t_2 = 1.5, t_3 = 3, t_4 = 4.5, \ldots, t_b = s - 12$ 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 3.1 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.2. 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 3.4 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

			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

Table 3.1: Overall Type I Error by Varying Levels of Correlation between Recurrent Events (Rows) and Correlation between Recurrent and Terminal Events (Columns)

[†] 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 ρ_1=0.3, 0.5 and 0.7, respectively.

[&] Low, median to high correlations between recurrent and terminal events are generated from ρ_2=0.3, 0.5 and 0.7, respectively.

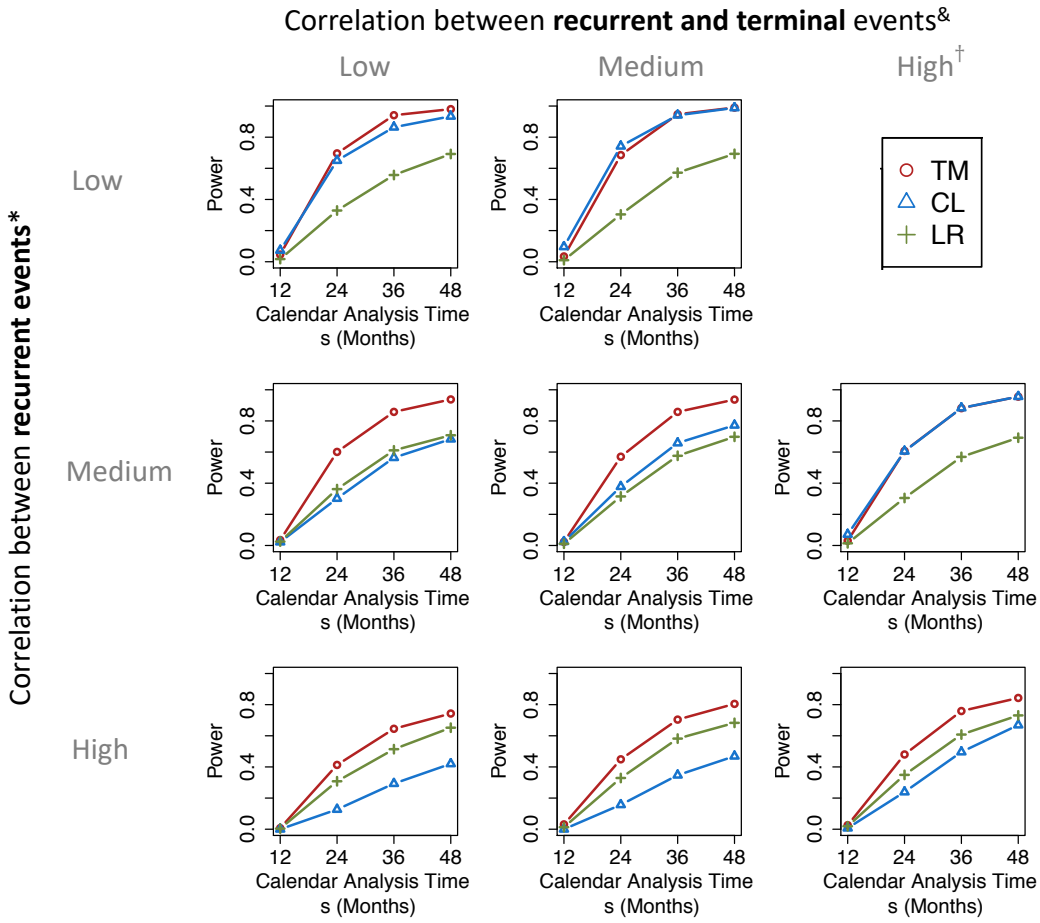


Figure 3.4: Cumulative Power at Each Analysis Time by Varying Levels of Correlation Between Recurrent Events (Rows) and Correlation between Recurrent and Terminal Events (Columns)

† 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 ρ_1 =0.3, 0.5 and 0.7, respectively.

& Low, median to high correlations between recurrent and terminal events are generated from ρ_2=0.3, 0.5 and 0.7, respectively.

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 3.4. Because the logrank test's simulated power dynamic is similar from panel to panel of Figure 3.4, 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 3.4, 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 3.4.

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 3.4, 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 3.4 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 3.4, where the power of the CL method increases a bit compared to its power dynamics as shown in panels to its left.

3.6 Example

The Azithromycin in COPD Trial (*Albert et al.*, 2011) 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. 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), initiate follow-up windows every 2 months (approximately one third of the historic mean time to exacerbation in this population).

Figure 3.5 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 3.4. A recommended stopping boundary for safety with spending function, $\alpha_{JT}(\gamma) = 0.2\gamma^{1.5}$, is superimposed in Figure 3.5. This boundary is a special case of a Jennison and Turnbull (*Jennison and Turnbull*, 2000) 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 3.6. The CL stops at the 4th interim analysis (2 years into the study) with 59 additional acute exacerbations and 4 additional deaths observed

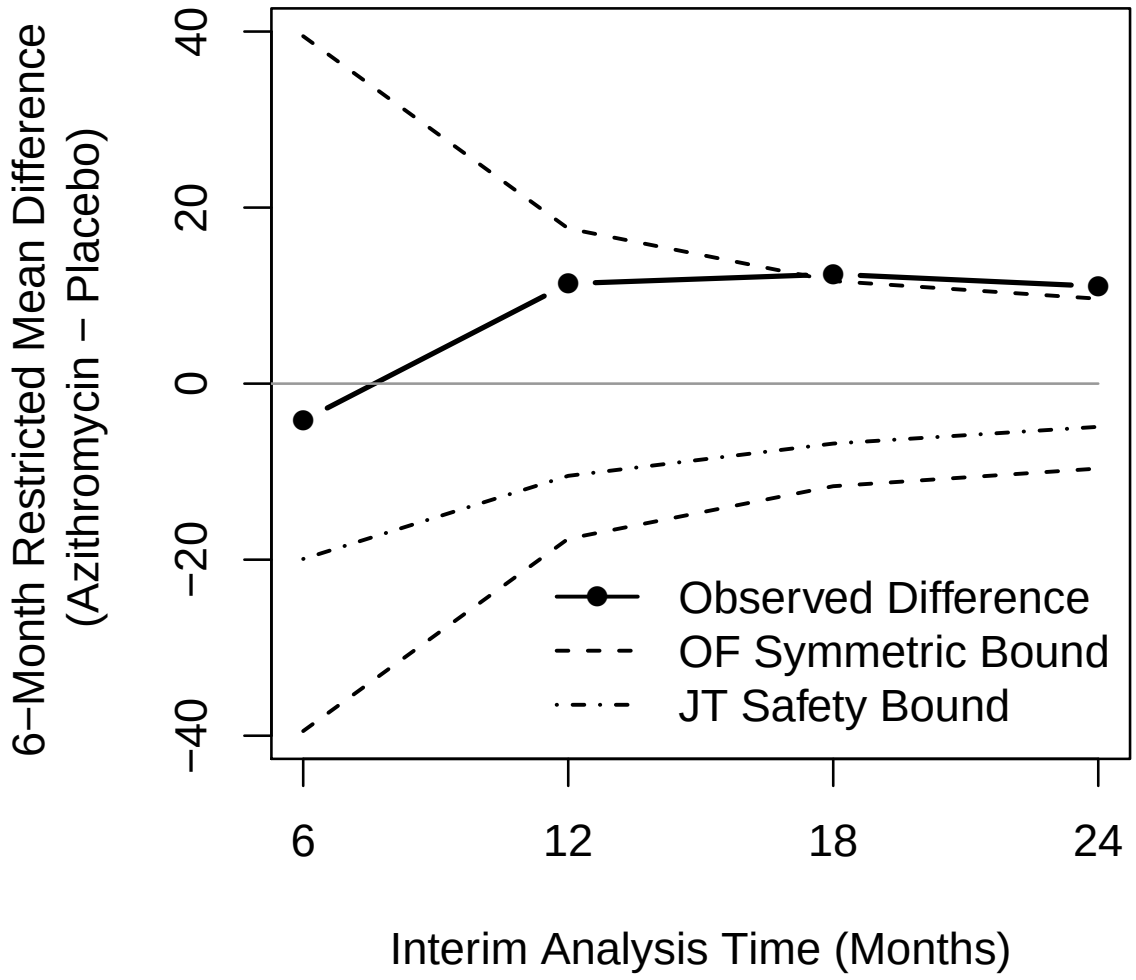


Figure 3.5: Additional Days Free of Acute Exacerbation or Death per 6-months of Follow-up When Using Azithromycin versus Placebo, Based on the TM Statistic

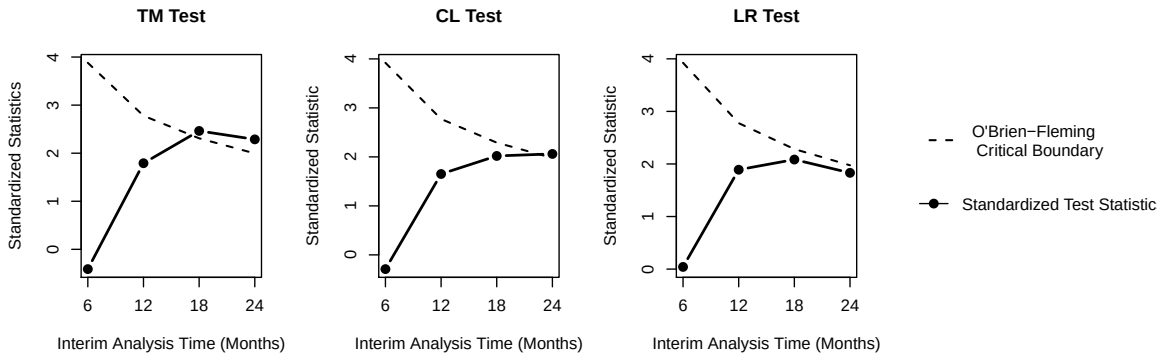


Figure 3.6: Standardized Test Statistics and Critical Boundaries for NACT Example. Lower (symmetric) O'Brien-Fleming Boundary and JT Boundary Are Not Displayed.

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.

3.7 Discussion

In this chapter, 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) 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.

CHAPTER IV

Commentary on Tayob and Murray (2014) with a Useful Update Pertaining to Study Design

The two-sample tests described in *Tayob and Murray* (2014) combine information from recurrent and terminal events in order to detect treatment differences. Instead of following the standard recurrent events paradigm that uses information on gap times between events, their work repurposes the data into a regularly spaced longitudinal form that avoids the usual dependent censoring issues that often plague gap time analyses. The endpoints are based on τ-length follow-up windows that start at evenly spaced times $\{t_1, \ldots, t_b\}$; b is chosen so that the final τ-length follow-up interval starting at t_b does not exceed the study period. In each of these follow-up windows the observed endpoint is the time to first event (recurrent event or death) or τ if no event occurs during that window. These endpoints are subject to the usual independent right censoring that occurs in the clinical trial setting. The analysis proceeds by comparing either the overall τ-restricted mean survival estimated from the follow-up windows from the two treatment groups, or alternatively, the area under the τ-restricted mean residual lifetime function.

Tayob and Murray (2014) recommended using $t_k = (k-1)\tau/2$, for $k = 1, \ldots, b$, as the starting points of the incorporated follow-up windows. This recommendation arose from closed-form variance calculations and accompanying simulations in the

special case of a single time-to-event. For our purposes it is convenient to describe these starting points as being initiated every a units starting at time 0 and ending at time t_b, with Tayob and Murray's recommendation equivalent to setting $a = \tau/2$.

In the recurrent events setting, if a is large relative to the mean of the recurrent event time, there is potential for a few of the recurrent events to be left out of the analysis, potentially reducing power. Intuitively, smaller values of a will create more follow-up windows that capture more recurrent events, but at the cost of computational efficiency. Computation burden increases substantially as the number of follow-up windows incorporated into the Tayob and Murray statistic increases. For example, in a dataset with 100 patients per group observed over 48 months with control versus treatment mean recurrent event times of 3 versus 4 months, computation of the Tayob and Murray statistic took an hour when incorporating follow-up windows starting every 10 days versus 1.7 minutes when spacing follow-up windows 1.5 months apart; times based on running R version 3.4.1 on a MacBook Pro with a macOS High Sierra operating system, a 2.9 GHz Intel Core i5 processor and 8 GB of Memory.

In this chapter we give improved guidance on the choice of a. Our recommendation is framed in terms of the average proportion, p, of recurrent events captured in at least one follow-up window for individuals followed s time units. This measure, p, is a compromise between the special case of a single time-to-event considered by *Tayob and Murray* (2014) and the impractical opposite extreme where window start times are taken along a continuum from time 0. Our calculations assume independent exponential times between recurrent events, with rate λ.

Let $\text{pdf}_{\text{Exp}(\lambda)}(g)$ be the exponential(λ) probability density function (pdf) with rate λ and $g > 0$. Let $\text{pdf}_{\text{Gamma}(\alpha,\lambda)}(r)$ be the gamma pdf with shape α, rate λ and $r > 0$, with corresponding cumulative distribution function (cdf) denoted as $\text{cdf}_{\text{Gamma}(\alpha,\lambda)}(r)$. After some calculation shown in Appendix C, we obtain an expression for p that can

be numerically evaluated in terms of a, λ, and s, or inverted to solve for a as a function of p, λ and s. That is,

$$
\begin{aligned}
p =& 1- \\
& \sum_{k=2}^{\infty} \frac{1}{k} \sum_{j=2}^{k} \sum_{w=1}^{b} \Bigg[\int_{0}^{\min(aw,s)} \mathrm{pdf}_{\mathrm{Gamma}(j,\lambda)}(r) \Big\{ \mathrm{cdf}_{\mathrm{Gamma}(k-j,\lambda)}(s-r) \\
& - \mathrm{cdf}_{\mathrm{Gamma}(k-j+1,\lambda)}(s-r) \Big\} dr \\
& - \int_{0}^{(w-1)a \min(aw,s)-r} \int_{0} \mathrm{pdf}_{\mathrm{Gamma}(j-1,\lambda)}(r) \mathrm{pdf}_{\mathrm{Exp}(\lambda)}(g) \Big\{ \mathrm{cdf}_{\mathrm{Gamma}(k-j,\lambda)}(s-r-g) \\
& - \mathrm{cdf}_{\mathrm{Gamma}(k-j+1,\lambda)}(s-r-g) \Big\} dg dr \Bigg]
\end{aligned}
$$

As one might expect, larger values of p are monotonically related to smaller values of a in this expression.

As an example of how p influences our choice of a and therefore study power, suppose we are designing a 48-month study where $n = 100$ per group. Thirty percent of these patients are recruited at the start of the study and observed for the full 48 months; the remaining 70% are uniformly accrued over the first 24 months. We wish to base the analysis on the Tayob and Murray statistic using $\tau = 12$. Table 4.1 below summarizes operational study design results for 3 different values of $p = 0.7, 0.8$ and 0.9 and 4 different recurrent event rates in the control group. In each case, the mean recurrent event time in the control group is 25% shorter than the mean event time in the treatment group. Values of a in each tabulated scenario are numerically calculated based on $s = 48$, λ corresponding to the control group recurrent event rate and p. Power shown in Table 4.1 is simulated from 1000 Monte Carlo iterations for both independent and dependent recurrent events with the correlation between recurrent event times, ρ, equal to $0, 0.25$ or 0.5. Type I error rates are 0.05 in all scenarios.

If we first focus on the recommended values of a shown in Table 4.1, we see a wide

range of values [0.7 to 12 months] with more closely spaced follow-up window start times recommended when the control group experiences shorter mean times-to-event, and additionally when we wish to capture a higher proportion of these events in the analysis plan. This is as opposed to using $a = \tau/2 = 6$ months as recommended in the original Tayob and Murray manuscript for all scenarios. Our simulations confirm that there are diminishing returns in power gain as more follow-up windows are included. The 'sweet spot' for clinical trial design seems to be $p = 0.8$, or 80% of events captured in the analysis, although for particularly expensive clinical trials even modest gains in power may be worth the extra computational burden. The simulations also suggest a convenient rule of thumb of choosing a equal to half of the mean recurrent event time in the control group, which gave p between 0.8 and 0.9 in all scenarios shown.

Similar to all other clustered data structures, our longitudinal restricted mean event times contain the most statistical information when event times within an individual are independent. Hence it is not surprising that power is highest in each scenario when $\rho = 0$. In clustered data analyses, ρ is typically described as an intra-class correlation coefficient, and our simulated power results reflect similar results to those seen in these settings. Namely, higher correlation between events severely impacts available power for analysis. To the extent that our updated recommendations for a can recover some of the power lost by using the original recommendation of Tayob and Murray, we feel this chapter may be quite helpful in conserving clinical trial resources. It should be noted that even with the original recommendation of $a = \tau/2$, Tayob and Murray saw gains in power compared to methods recommended in *Ghosh and Lin* (2000) and *Lin et al.* (2000) when times between recurrent events were correlated.

Our closed-form calculation for p ignores the possibility of a terminal event, which is allowed by the Tayob and Murray statistic. When terminal events are present, our calculation provides a lower bound on the estimated proportion of observed events

captured in at least one follow-up window. The intuitive explanation is that follow-up windows with a terminating event have an observed follow-up duration less than the value of s assumed in our equation for p. As can be seen in Figure 4.1, smaller values of s trend, but not monotonically, towards larger values of p. In additional simulations introducing terminal events, we've observed the proportion of captured events increase by as much as 5% in scenarios with a high terminal event hazard relative to the recurrent event rate. When terminal events occur much less frequently than recurrent events, our calculations continue to give very similar results to those seen in simulation [data not shown].

There are not straightforward extensions of the closed-form formulas to other distributions. We've performed additional simulations with event rates following the Weibull distribution that show for similar mean gap-times, our closed form formula underestimates the proportion of captured events when the shape parameter is greater than 1 and overestimates this proportion when the shape parameter is less than one. Hence, as is typically the case, simulations are required when event rates are not anticipated to follow an exponential distribution. Our work provides a framework (proportion of captured events) for programmers to use in designing such simulations to understand operating characteristics in non-standard settings.

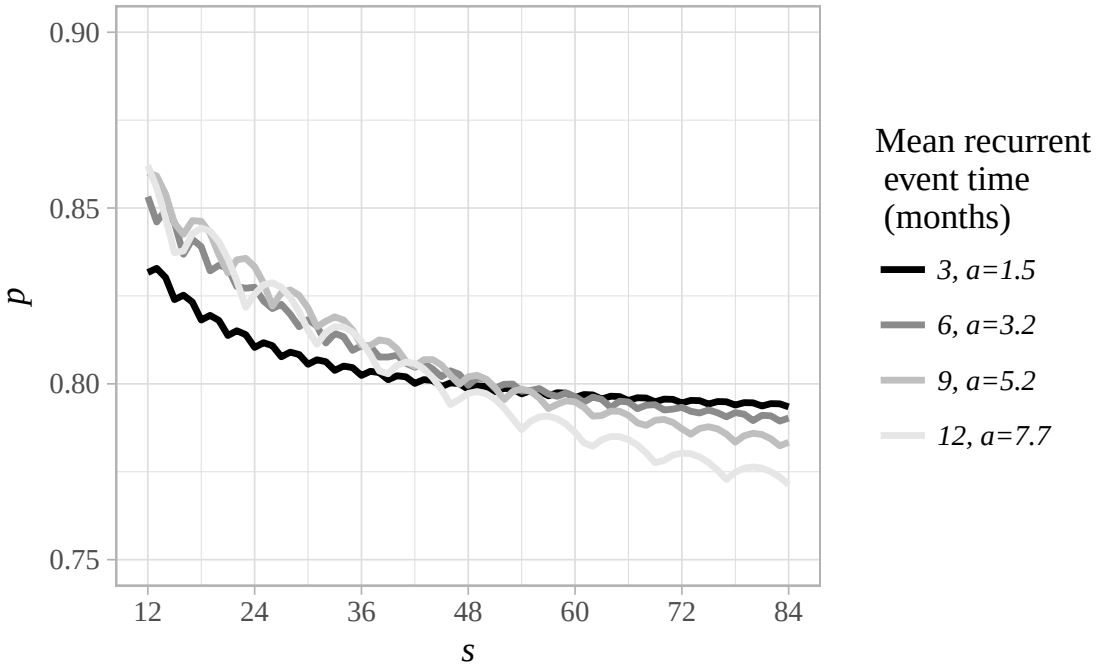


Figure 4.1: Proportion (p) of Events Captured by Follow-up Times (s) in Months. (Curves vary across mean recurrent event times assumed in columns of Table 1. For each curve a fixed value of a is assumed that corresponds to the recommendation given in Table 1 for achieving $p = 0.8$ over an $s = 48$ month follow-up period. Values of p trend higher, but not monotonically, as s decreases. The non-monotonic pattern is caused by cumulative increases in the number of follow-up windows (b) as s increases.)

Proportion of Events Captured (p)		Control Group Mean Recurrent Event Time in Months			
		3	6	9	12
	a	2.4	5.3	8.8	12
	Power				
0.7	$\rho = 0$	0.969	0.842	0.690	0.548
	$\rho = 0.25$	0.858	0.736	0.634	0.497
	$\rho = 0.5$	0.700	0.588	0.527	0.416
	a	1.5	3.2	5.2	7.7
	Power				
0.8	$\rho = 0$	0.975	0.868	0.748	0.639
	$\rho = 0.25$	0.887	0.755	0.670	0.547
	$\rho = 0.5$	0.703	0.602	0.544	0.449
	a	0.7	1.5	2.4	3.4
	Power				
0.9	$\rho = 0$	0.975	0.881	0.767	0.670
	$\rho = 0.25$	0.887	0.773	0.683	0.594
	$\rho = 0.5$	0.710	0.627	0.558	0.505

Table 4.1: Calculated a Values and Estimated Power

CHAPTER V

Regression Analysis of Recurrent-Event-Free Time from Multiple Follow-up Windows

5.1 Introduction

Recurrent events are frequently seen in participants of clinical trials and observational studies of chronic diseases. For instance, patients in the Azithromycin in Chronic Obstructive Pulmonary Disease (COPD) Trial (*Albert et al.*, 2011) were followed for recurrent acute pulmonary exacerbations. Other settings with recurrent events include recurrent ischemic cardiovascular events after acute coronary syndrome (*Schwartz et al.*, 2018), recurrent clostridium difficile infection (*Wilcox et al.*, 2017) and even repetitive head injuries in high-contact sports (*DeKosky et al.*, 2010). Poisson and negative binomial count models have been used to analyze recurrent event data per time at risk (*Frome et al.*, 1973; *Lawless*, 1987; *Lambert*, 1992; *Greene*, 1994). These approaches to not take advantage of the timing of events, however, and may therefore not provide the most powerful analysis (*Ozga et al.*, 2018).

The most commonly used multivariable regression analysis methods for recurrent events data are extensions of the Cox proportional hazards model to the recurrent event setting. The extension proposed by *Andersen and Gill* (1982) analyzes the time between recurrent events, called gap times, assuming independence between these gap

times within an individual. *Prentice et al.* (1981) considered an extension of the Cox model that allowed stratification of the baseline hazard to depend on time-dependent features including previous recurrent event time information; both gap time models and models of time from beginning of follow-up are considered. *Wei et al.* (1989) proposed a multivariate proportional hazards model, where the multivariate outcomes are based on separate recurrent events modeled from the beginning of follow-up. An arbitrary covariance structure is allowed between the different event-times, fit with a robust sandwich variance estimate. *Pepe and Cai* (1993) described several manners of modeling recurrent event rates based on the number of previous recurrent events, advocating for a Markov approach that models each recurrent event conditional on information from the immediately preceding event. *Lawless and Nadeau* (1995) and *Lin et al.* (2000) developed models for the cumulative mean number of events, assuming proportionality on the cumulative means over time. A number of authors introduced random effects or frailties to parameterize the dependence between recurrent event times (*Aalen and Husebye*, 1991; *Hougaard*, 1995; *Rondeau et al.*, 2007; *Mazroui et al.*, 2013; *Rogers et al.*, 2016).

In pursuing any new modeling framework for recurrent events, three issues are paramount to address (1) the potential correlation between times between recurrent events, (2) the potentially censored nature of the data and (3) the interpretability of results. In addressing each of these issues in this chapter, we take an entirely different approach to modeling recurrent event data that provides a natural way to handle correlation between event times and is highly interpretable. In short, we transform the recurrent event data structure into a very tractible censored longitudinal data structure. The longitudinal outcomes are τ-restricted times-to-first-event as captured in follow-up windows that are reinitiated at regularly-spaced intervals. Instead of modeling the rate or cumulative number of recurrent events, our model estimates time free from recurrence over a τ-length follow-up period.

In Section 5.2, we describe notation required to repurpose traditional recurrent events data into a series of censored longitudinal endpoints. Section 5.3 describes differences between this data structure and that of the multivariate distribution of gap times between recurrent events. In section 5.4, we develop a model framework that can be fit using generalized estimating equation methods, along with two methods for handling the censored nature of the data: a pseudo-observation approach (Section 5.4.1) and a multiple imputation approach (Section 5.4.2). Section 5.5 describes finite sample properties of our methodology in scenarios where times between recurrent events are independent (Section 5.5.1) and correlated (Section 5.5.2). We then reanalyze data from the Azithromycin in COPD Trial using our methodology in Section 5.6. Discussion follows in Section 5.7.

5.2 Notation

For the most part, notation in this chapter is similar to that used in Chapter III. Because group sequential analysis are not being performed, the parameter, s, for analysis time is removed. We also drop use of the group subscript, g, since predictors of this nature will be absorbed into a covariate vector in this chapter.

Suppose $i = 1, \ldots, N$ independent patients are followed for recurrent events. Without loss of generality, we assume each patient's follow-up period starts from a baseline time of 0; hereafter, we refer to baseline and time 0 interchangeably. For each individual patient, i, let $T_{ij}, j = 1, \ldots, J_i$ be the time from baseline to the j^{th} recurrent event, so that $0 < T_{i1} < T_{i2} < \cdots T_{iJ_i}$. Let C_i be the censoring time from baseline for patient i, where C_i is independent of T_{ij}, for $j = 1, \ldots, J_i$. Correlation between recurrent event times in an individual i (or lack thereof) is typically formulated in terms of gap times between events, $\left\{G_{i1} = T_{i1}, G_{i2} = T_{i2} - T_{i1}, \ldots, G_{iJ_i} = T_{iJ_i} - T_{iJ_{i-1}}\right\}$. We allow an arbitrary dependence structure between gap time random variables for patient i, with independent gap times as a special case. Traditional observed recurrent

event data for patient i is recorded in data pairs $\{X_{ij} = min(T_{ij}, C_i), \delta_{ij} = I(T_{ij} \leq C_i)\}$, $j = 1, \ldots, \tilde{J}_i$, where $\tilde{J}_i \leq J_i$; in most cases the $\tilde{J}_i^{th}$ data pair corresponds to a censored event time.

In this chapter, we construct a streamlined censored longitudinal data structure from the recurrent event times. That is, each longitudinally measured outcome contributed by patient i is a censored time-to-first-event in a follow-up window starting at time t, where $t \in \{t_1, \ldots, t_b\}$ with $t_1 = 0$ and $t_k = t_{k-1} + a, k = 2, \ldots, b$. As only one time-to-first-event in each follow-up window is measured, we incorporate at most b outcomes from each individual, regardless of how many recurrent events they experience. Hence, for a fixed overall study duration, the choice of spacing, $a = t_k - t_{k-1}, k = 2, \ldots, b$, between initiation of each subsequent follow-up window increases the proportion of recurrent events captured by the censored longitudinal data structure. It is theoretically possible to create a censored longitudinal dataset with follow-up windows initiated every day ($a = 1$), although the computational burden of working with this extended dataset becomes cumbersome. *Xia and Murray* (2018) showed that in the case of exponentially distributed times between events with common intensity λ, using $a = 1/(2\lambda)$ captures approximately 80% of the recurrent events in at least one of the constructed follow-up windows over a fixed follow-up period.

For patient i and follow-up window starting at t, we index the first recurrent event occurring after time t with the subscript $\eta_i(t) = min\{j = 1, \ldots, J_i : T_{ij} \geq t\}$ so that $T_i(t) = T_{i\eta_i(t)} - t$ is the time-to-first-recurrent-event measured from t, sometimes called the residual event-free time from t. We collect individual i's newly formatted longitudinal outcomes, $\{T_i(t_1), T_i(t_2), \ldots, T_i(t_b)\}$, into a vector, $\mathcal{T}_i$, for $i = 1, \ldots, N$.

The observed data counterpart to $\eta_i(t)$ is $\tilde{\eta}_i(t) = min\{j = 1, \ldots, \tilde{J}_i : X_{ij} \geq t\}$. For each follow-up window starting at time t where $C_i > t$, patient i contributes the observed data triplet $\left\{\tilde{\eta}_i(t), X_i(t) = X_{i\tilde{\eta}_i(t)} - t, \delta_i(t) = \delta_{i\tilde{\eta}_i(t)}\right\}$. For follow-up windows

starting at t where $C_i \leq t$, we use the convention that $\tilde{\eta}_i(t) = X_i(t) = \delta_i(t) = 0$.

Figure 5.1 displays how censored longitudinal data is created from traditional recurrent event data using a participant from the Azithromycin in COPD Trial. During 353 days of follow-up for this patient, $\tilde{J}_i = 4$ traditional recurrent event data pairs emerge: $(X_{i1} = 53 \text{ days}, \delta_{i1} = 1)$, $(X_{i2} = 111 \text{ days}, \delta_{i2} = 1)$, $(X_{i3} = 170 \text{ days}, \delta_{i3} = 1)$ and $(X_{i4} = 353 \text{ days}, \delta_{i4} = 0)$, where the first three data pairs denote acute exacerbation (AE) event times and the last data pair reflects a censored event.

Two examples of converting these data into a censored longitudinal data structure are given, one based on follow-up windows starting at $a = 120$ day intervals and one constructed with $a = 60$ day intervals. As with all longitudinal data structures, the additional data triplets included using $a = 60$ day intervals as opposed to $a = 120$ day intervals afford capturing more time-to-first events supplied by the recurrent event times.

Data triplets based on follow-up windows starting at $t = \{0, 120, 240\}$ days become

$$\{\tilde{\eta}_i(0) = 1, X_i(0) = X_{i1} - 0 = 53, \delta_i(0) = 1\},$$

$$\{\tilde{\eta}_i(120) = 3, X_i(120) = X_{i3} - 120 = 170 - 120 = 50, \delta_i(120) = 1\}$$

and

$$\{\tilde{\eta}_i(240) = 4, X_i(240) = X_{i4} - 240 = 353 - 240 = 113, \delta_i(240) = 0\}.$$

Data triplets based on follow-up windows initiated every $a = 60$ days include the above data triplets plus those starting at days $t = \{60, 180, 300\}$ days:

$$\{\tilde{\eta}_i(60) = 2, X_i(60) = X_{i2} - 60 = 111 - 60 = 51, \delta_i(60) = 1\},$$

$$\{\tilde{\eta}_i(180) = 4, X_i(180) = X_{i4} - 180 = 353 - 180 = 173, \delta_i(180) = 0\}$$

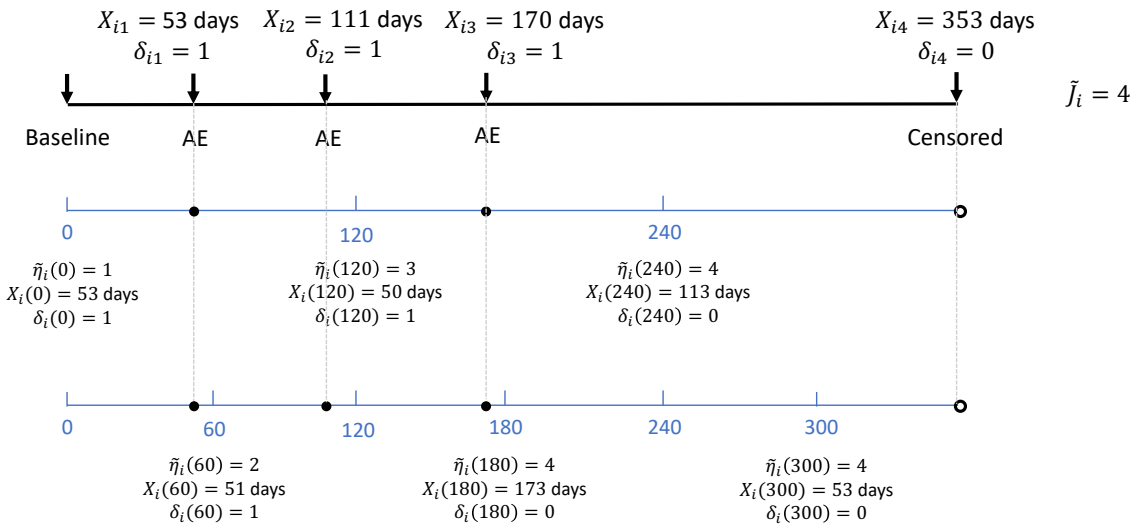


Figure 5.1: Creating Censored Longitudinal Data From Recurrent Event Data for an Example Participant of the Azithromycin in COPD Trial. (AE: Acute Exacerbation)

and

$$\{\tilde{\eta}_i(300) = 4, X_i(300) = X_{i4} - 300 = 353 - 300 = 53, \delta_i(300) = 0\}.$$

5.3 Times-to-first-event from $t \in \{t_1, \ldots, t_b\}$ versus Gap Times

The marginal distribution of the time-to-first recurrent event after t, $T_i(t)$, is a quite different creature from the marginal distribution of a gap time between recurrent events, $G_{ij}, j = 1, \ldots, J_i,$. On a practical note, the random variable, $T_i(t)$, better reflects the recurrent event time one might seek advice about at a regularly scheduled clinic visit at time t or at entry into a clinical trial at t. These patient interactions rarely coincide exactly with a recurrent event, so that a gap time random variable measured from an individual's previous event is not the most appropriate random variable for these settings. On a statistical note, when gap times within an individual are correlated there is a well-known dependent censoring bias that must be addressed

in any analysis of the gap time data (*Lin et al.*, 1999). This dependent censoring issue is circumvented by our censored longitudinal data structure since times-to-first-event are measured from pre-specified times $\{t_1, \ldots, t_b\}$ rather than a correlated time-to-event.

For these different random variables to coincide with one another, the distributions of $T_i(t)$ and G_{ij} must both be entirely memoryless. This is formally demonstrated in the following, where we show that the only case where the marginal distribution of a gap time, G_{ij}, coincides with that of a time-to-first-recurrent-event, $T_i(t)$, is the special case with independent and identically distributed exponential gap times $\{G_{i1}, \ldots, G_{iJ_i}\}$ for each patient, $i = 1, \ldots, N$. For settings with gap times that are not exponentially distributed, or for settings with correlated, but otherwise identically distributed, exponential gap times, the marginal distribution of $T_i(t)$ shifts from a memoryless distribution to a distribution very much influenced by the series of recurrent events with positive probabilistic support for occurring in the follow-up period after t.

Consider the event-free probability function for $T_i(t)$ that is written in terms of gap time random variables as follows:

$$Pr\{T_i(t) > u\} = Pr\{T_{i1} > t + u\} + \lim_{J_i \to \infty} \sum_{j=2}^{J_i} Pr\{T_{ij-1} \leq t, T_{ij} > t + u\}$$

$$= Pr\left\{G_{i1} > t + u\right\} + \lim_{J_i \to \infty} \sum_{j=2}^{J_i} Pr\left\{\sum_{l=1}^{j-1} G_{il} \leq t, \sum_{l=1}^{j-1} G_{il} + G_{ij} > t + u\right\}. \qquad (5.1)$$

For independently and identically distributed exponential gap times with intensity λ, Appendix D.1 shows that these terms reduce to $exp(-\lambda u)$, so that $T_i(t)$ also has an exponential distribution with intensity λ.

However, when the gap times are correlated, the term

$$Pr\left\{\sum_{l=1}^{j-1} G_{il} \leq t, \sum_{l=1}^{j-1} G_{il} + G_{ij} > t + u\right\} \tag{5.2}$$

from the previous equation does not reduce to a simple expression. Term (5.2) is the probability that an individual's j^{th} recurrent event will be the first to occur in the follow-up window starting at t, but that it has not yet occurred as of time $t + u$.

To better appreciate the influence of (5.2) on the expression in (5.1), we consider special cases with correlated and independent exponential(λ_i) distributed times between recurrent events. Figure 5.2 displays term (5.2) as a function of time, u, for different combinations of recurrent event index, j, and follow-up window start time, t, with $\lambda_i = 1/3$. The solid blue and dashed red lines show the cases with independent and correlated event times, respectively. For the independence case, the curves have a closed-form shown in Appendix D.1 to be $(\lambda_i t)^{j-1} e^{-\lambda_i(t+u)}/\Gamma(j)$. For the correlated case, we first simulated correlated exponential event times using a Gaussian copula approach described in further detail in Section 5.5; the approximate correlation between recurrent event times was 0.8. We then empirically estimated and plotted term (5.2) from a large number (N=10,000) of simulated individuals.

In nearly every panel of Figure 5.2, term (5.2) is smaller when times between recurrent events are correlated. As j increases relative to t the depicted probability curves get lower, and the curves generated from the two different correlation structures also get closer together. Overall these curves indicate that $Pr\{T_i(t) > u\}$ tends to be much smaller than the exponential(1/3) survival curve that equation (5.1) reduces to in the case with independent event times.

Of course, we were immediately curious to know whether the distribution of the time-to-first-event from t, in this special case with correlated exponential(1/3) gap times, stabilizes. The intuition behind this thought was that the mixture distribution

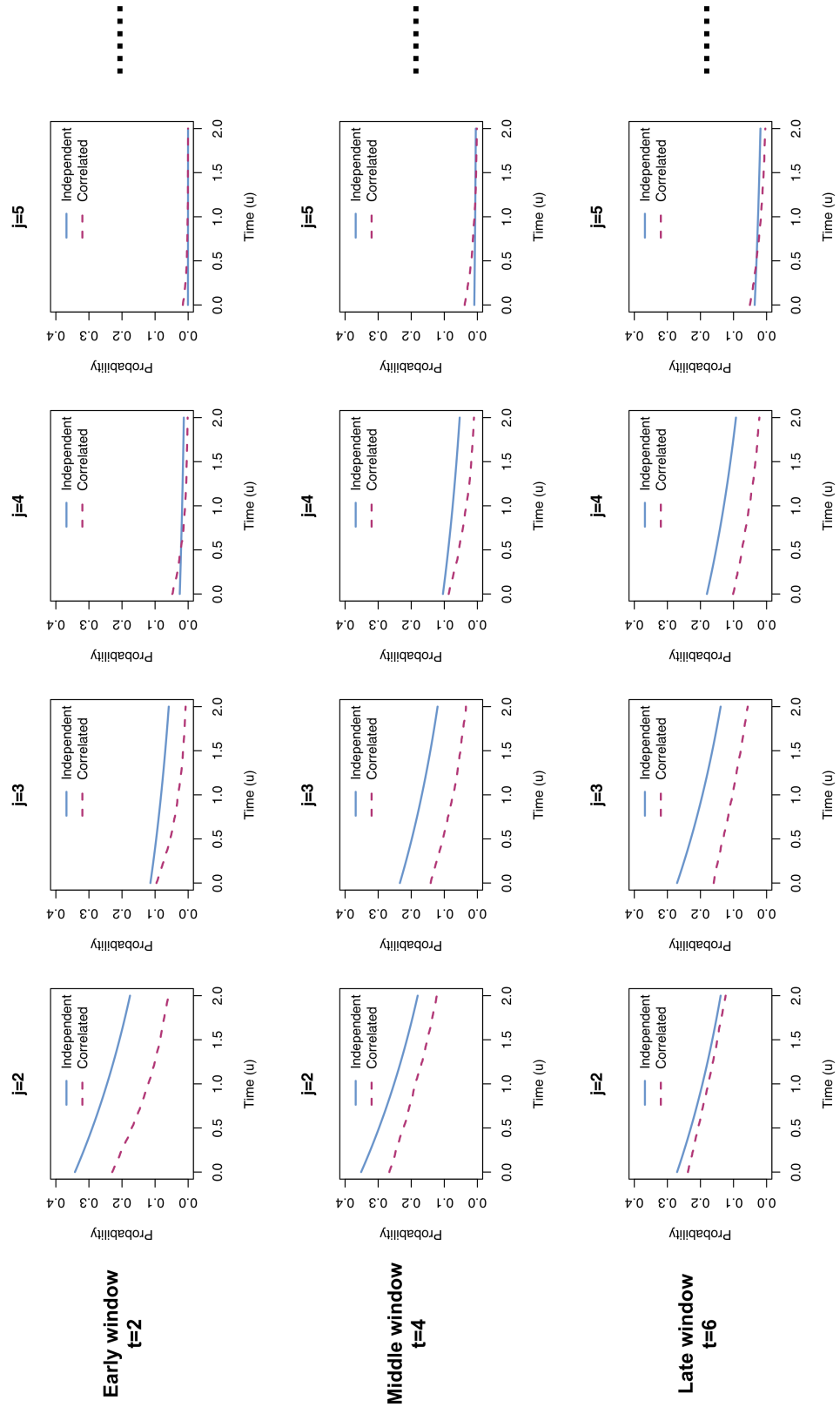


Figure 5.2: $Pr\{T_{ij-1} \leq t, T_{ij-1} + G_{ij} > t + u\}$ over u for Specific t and j.

of gap time histories preceding t and likely to influence the distribution of $T_i(t)$ might stabilize. As seen in Figure 5.3, where (again using the large simulated dataset of 10,000 individuals) $Pr\{T_i(t) > u\}$ is plotted for increasing values of t, this does seem to be the case. The distribution of $T_i(t)$ seems to stabilize for values of $t \geq 3$, or $1/\lambda_i$. Stabilization of the distribution of $T_i(t)$, for $t > 1/\lambda_i$ was further explored for different values of λ_i and found to be a reliable pattern. This feature will be utilized later in Section 5.5.2, when we simulate a stable time-to-first-event distribution given covariates in the setting with correlated exponential gap times.

5.4 Multivariable Regression Model of τ-Restricted Times-to-first-recurrent-event Measured Across Multiple Overlapping Follow-up Periods

To study the association between patient covariates, Z, and τ-restricted times-to-first-recurrent event across follow-up windows starting at times $\{t_1, \ldots, t_b\}$, we consider the following model:

$$E(log[min\{\tau, \mathcal{T}\}]|Z) = \beta^T Z \tag{5.3}$$

Two features of our data need to be addressed for successful estimation of model (5.3): (1) the censored nature of the vector of newly formatted longitudinal outcomes, $\mathcal{T}_i$, from patient i and (2) the correlated nature of these longitudinal outcomes. We develop two approaches that address the censoring aspect of the data, a pseudo-observation approach in section 5.4.1 and a multiple imputation approach in section 5.4.2. Each of these approaches converts the censored longitudinal outcomes into a format appropriate for complete data methods.

Once this feature of the data is addressed, we tackle the correlated nature of the longitudinal outcomes $\mathcal{T}_i$ from each patient, $i = 1, \ldots, N$ using existing methods, such

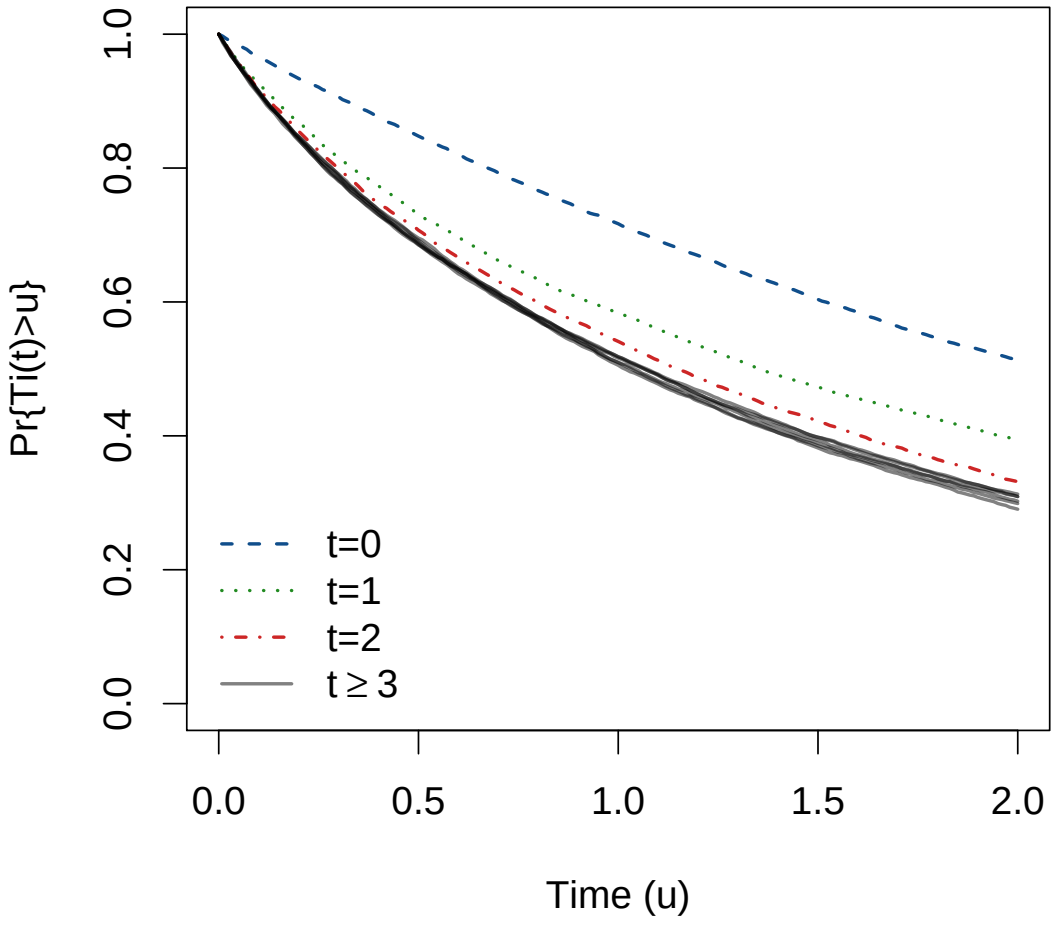


Figure 5.3: $Pr\{T_i(t) > u\}$ by Follow-up Window Start Times, t.

as generalized estimating equations (GEE). Because our recurrent events data have been restructured into times-to-first event from regularly spaced follow-up periods and because we consider $\tau-$restricted times-to-first-events in these periods, the correlation structure of the outcomes can be modeled via well organized correlation matrices.

Two underlying layers of correlation are at work: the natural correlation between recurrent events within an individual and the possibility that the same event is captured as the first-time-to-event in more than one follow-up period. In the most general case, we assume an $b \times b$ unstructured correlation matrix with components:

$$\begin{bmatrix} 1 & corr\{T_i(t_1), T_i(t_2)\} & corr\{T_i(t_1), T_i(t_3)\} & \dots & corr\{T_i(t_1), T_i(t_b)\} \\ corr\{T_i(t_2), T_i(t_1)\} & 1 & corr\{T_i(t_2), T_i(t_3)\} & \dots & corr\{T_i(t_2), T_i(t_b)\} \\ \dots & \dots & \dots & \dots & \dots \\ corr\{T_i(t_b), T_i(t_1)\} & corr\{T_i(t_b), T_i(t_2)\} & corr\{T_i(t_b), T_i(t_3)\} & \dots & 1 \end{bmatrix}$$

However, for settings with fairly stable time-to-first-event distributions over time, we consider a (banded) Toeplitz correlation structure that allows for correlation to decrease as the degree of overlap between τ-restricted follow-up periods decreases. The degree of overlap in τ-restricted follow-up windows can be determined from a and τ and follows a regular pattern. For instance with $a = \tau/3$ and $b = 4$ windows starting at times $t = \{0, \tau/3, 2\tau/3, \tau\}$, the Toeplitz correlation matrix is:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

where ρ_1 is the correlation between times-to-first event in adjacent τ-restricted follow-up windows that overlap by $2\tau/3$ follow-up units, ρ_2 is the correlation between times-to-first event in windows that start $2a$ units apart from one another and overlap by

$\tau/3$ units. Finally, τ-restricted follow-up windows starting $a = 3$ units apart from one another do not overlap and are assumed to have correlation ρ_3.

The Toeplitz correlation structure requires fewer parameters than the unstructured matrix. For very large b, the Toeplitz correlation structure may be more feasible to implement that an entirely unstructured variance matrix. GEE also provides model results based on robust sandwich variance estimation, which provides protection against misspecification of the working correlation matrix. The general recommendation when working with large datasets is to use the sandwich estimator, regardless of the working correlation matrix assumed by the model. We follow this recommendation throughout the remainder of the chapter.

5.4.1 Pseudo-Observation (PO) Approach For Censored Recurrent Events

For a single time-to-event, *Andersen et al.* (2004) introduced the idea of using pseudo-observations (POs) in lieu of censored times-to-event when estimating regression parameters for the restricted mean model. This method has been successfully applied in a variety of settings where a single event time is of interest (*Klein and Andersen*, 2005; *Andersen and Klein*, 2007; *Andrei and Murray*, 2007; *Graw et al.*, 2009; *Xiang and Murray*, 2012; *Tayob and Murray*, 2017). The appeal of this method is its ease of use. That is, once appropriate pseudo-observations are estimated for each patient, they can be used as if they are uncensored counterparts to the original censored data in standard regression models. In this section, we describe how to create pseudo-observations that correspond to our censored longitudinal data structure. In particular, for each follow-up window starting at t, we define pseudo observations for the random variables $log[min\{\tau, T_i(t)\}], i = 1, \ldots, N$, using a method similar to that described in *Xiang and Murray* (2012).

The general intuition behind pseudo-observation approaches for modeling censored survival data is similar to that of the jackknife method (*Quenouille*, 1949, 1956; *Tukey*,

1958). One first defines a consistent nonparametric estimate, $\hat{\theta}$, of the marginal quantity of interest, θ. In our setting, for each $t \in \{t_1, \ldots, t_b\}$, we define $\theta(t) = E[log\{min(\tau, T(t))\}]$ with consistent nonparametric estimator

$$\hat{\theta}(t) = -\int_0^{\tau} log(u) d\hat{P}(T_i(t) > u) + log(\tau)\hat{P}(T_i(t) > \tau),$$

where Kaplan-Meier estimation is used for $\hat{P}(T_i(t) > u)$.

The form of an appropriate PO for any setting arises from framing θ both as a marginal mean and a weighted average of θ_Z, the conditional mean given covariates, Z. Most readers will recognize this relationship when formally depicted as

$$\theta = \int \theta_Z dF_Z(z),$$

where $dF_Z(z)$ reflects Riemann-Stieltjes integration across the distribution of Z. When the empirical (discrete) distribution of Z is used in framing the relationship above, $dF_Z(z) = 1/N$ and the right hand side of the expression becomes

$$\frac{1}{N}\sum_{i=1}^{N} \theta_{Z_i}.$$

One can algebraically isolate θ_{Z_i} (individual i's mean given Z_i) from the expression above via

$$\theta_{Z_i} = N\left\{\frac{1}{N}\sum_{j=1}^{N}\theta_{Z_j}\right\} - (N-1)\left\{\frac{1}{N-1}\sum_{j=1, j\neq i}^{N}\theta_{Z_j}\right\}.$$

Marginal means corresponding to the terms in curly brackets can be consistently estimated using nonparametric estimates $\hat{\theta}$ and $\hat{\theta}^{(-i)}$, respectively, where $\hat{\theta}^{(-i)}$ is the "leave-one-out" estimator of θ, i.e., estimated without individual i. So, taking advantage of large sample properties of $\hat{\theta}$ and $\hat{\theta}^{(-i)}$, a natural pseudo-observation for

individual i to use in modeling θ_Z is

$$N\hat{\theta} - (N-1)\hat{\theta}^{(-i)},$$

a fully observed random variable that asymptotically shares a conditional mean, θ_{Z_i}, with patient i.

In our setting, $\theta_Z = E(log[min\{\tau, \mathcal{T}\}]|Z)$. For each t, we define pseudo-observations

$$PO_i(t) = N\hat{\theta}(t) - (N-1)\hat{\theta}^{(-i)}(t), i = 1, \ldots, N,$$

where

$$\hat{\theta}^{(-i)}(t) = -\int_0^{\tau} log(u) d\hat{P}^{(-i)}(T_i(t) > u) + log(\tau)\hat{P}^{(-i)}(T_i(t) > \tau),$$

where leave-one-out Kaplan-Meier estimation is used for $\hat{P}^{(-i)}(T_i(t) > u)$, i.e., excluding patient i. We denote the vector of pseudo-observations contributed by individual i as $PO_i = \{PO_i(t_1), PO_i(t_2), \ldots, PO_i(t_b)\}$. Parameter estimates for model (5.3) can be estimated using the longitudinally created PO data via

$$E[PO|Z] = \beta^T Z. \tag{5.4}$$

Hereafter, we refer to estimates from equation (5.4) as estimates using the proposed PO approach.

5.4.2 Multiple Imputation Approach for Censored Recurrent Events

Another approach for producing a complete dataset when a single time-to-event is subject to censoring is multiple imputation (MI). This approach has been developed by many authors (*Faucett et al.*, 2002; *Taylor et al.*, 2002; *Hsu et al.*, 2006; *Liu et al.*, 2011; *Xiang et al.*, 2014; *Tayob and Murray*, 2017). For our longitudinal

data structure, we propose multiply imputing outcomes for observed data pairs with $\{X_i(t) > 0, \delta_i(t) = 0\}, t \in \{t_1, \ldots, t_b\}$.

For our longitudinal data structure, the i^{th} individual requires imputation for times-to-first-event in the set, $\mathcal{S}_i$, of follow-up windows starting at times $\{t \in \{t_1, \ldots, t_b\} : X_i(t) > 0, \delta_i(t) = 0\}$. If $\mathcal{S}_i$ consists of more than one follow-up window, it suffices to impute the time-to-first-event corresponding to the window starting at follow-up time $t^{sup}(\mathcal{S}_i) = max\{\text{follow-up window start time } t \text{ for windows} \in \mathcal{S}_i\}$, which then determines imputes for all times-to-first-event in the set of follow-up windows, $\mathcal{S}_i$, that require imputation (See Appendix D.2 for further details). For better short-hand terminology, we call the imputed event time corresponding to follow-up window start time $t^{sup}(\mathcal{S}_i)$ the 'sup impute', denoted as $\tilde{T}_i\{t^{sup}(\mathcal{S}_i)\}$, and the follow-up window starting at time $t^{sup}(\mathcal{S}_i)$ the 'sup window'. Then imputed event times for follow-up windows in $\mathcal{S}_i$ with start times $t^* < t^{sup}(\mathcal{S}_i)$ become $\tilde{T}_i\{t^*\} = \tilde{T}_i\{t^{sup}(\mathcal{S}_i)\} + t^{sup}(\mathcal{S}_i) - t^*$.

The gestalt of the imputation strategy is to base the sup impute in the sup window on model (5.4) using individual i's covariates, Z_i. Random error for the sup impute is sampled nonparametrically from a set of residuals contributed by individuals in a risk set, $\mathcal{R}_i$, similar to individual i. Further details are described below.

The first step of the imputation procedure is to obtain parameter estimates, $\hat{\beta}^{PO}$, from model (5.4). For individual i requiring a sup impute, $\tilde{T}_i\{t^{sup}(\mathcal{S}_i)\}$, in the sup window, we then define a risk set, $\mathcal{R}_i$, of candidate individuals $l = 1, \ldots, N_i$ satisfying two constraints: (1) $X_l\{t^{sup}(\mathcal{S}_i)\} > X_i\{t^{sup}(\mathcal{S}_i)\}$, that is, candidate l is still at risk for their first event time in the sup window as of the time individual i is censored and (2) $|\hat{\beta}^{PO^T} Z_i - \hat{\beta}^{PO^T} Z_l| \leq \epsilon$, where ϵ is a user-defined parameter that controls how similar individual l's linear predictor is to individual i's linear predictor. Our algorithm used $\epsilon = 0.01$. In cases where ϵ resulted in a risk set with $N_i < 5$, our algorithm added 0.001 to ϵ until $N_i \geq 5$.

The next step of the imputation procedure is use candidate individuals, $l =$

$1, \ldots, N_i, \in \mathcal{R}_i$ to estimate the survival function for $T\{t^{sup}(\mathcal{S}_i)\}$ given membership in $\mathcal{R}_i$. Nonparametric Kaplan-Meier estimation is used for this purpose, resulting in estimate, $\hat{S}_{T\{t^{sup}(\mathcal{S}_i)\}}(v|\mathcal{R}_i)$. Then an inverse transform imputation algorithm (*Taylor et al.*, 2002; *Hsu et al.*, 2006; *Liu et al.*, 2011; *Xiang et al.*, 2014; *Tayob and Murray*, 2017) is used to select an impute following the distribution of $T\{t^{sup}(\mathcal{S}_i)\}$ given membership in $\mathcal{R}_i$ based on $\hat{S}_{T\{t^{sup}(\mathcal{S}_i)\}}(v|\mathcal{R}_i)$. In particular, the inverse transform imputation method first generates a uniform(0,1) random variable, u. If $\hat{S}_{T\{t^{sup}(\mathcal{S}_i)\}}(v|\mathcal{R}_i) > u$ for all observed event times v, we impute $\tilde{T}_i\{t^{sup}(\mathcal{S}_i)\} = \tau$. Otherwise, we find the smallest value v where $\hat{S}_{T\{t^{sup}(\mathcal{S}_i)\}}(v|\mathcal{R}_i) \leq u$ and identify the observed event time, $T_l\{t^{sup}(\mathcal{S}_i)\}$, that corresponds to v.

The inverse transform impute for patient i's time-to-first-event in the sup window would be $T_l\{t^{sup}(\mathcal{S}_i)\}$. However, our proposed imputation algorithm goes one step further, by defining residual $\varepsilon_l = log(min[\tau, T_l\{t^{sup}(\mathcal{S}_i)\}]) - \hat{\beta}^{PO^T} Z_l$ and then defining our final impute $\tilde{T}_i\{t^{sup}(\mathcal{S}_i)\} = exp[\hat{\beta}^{PO^T} Z_i + \varepsilon_l]$. This extra step allows for variability of the impute to be contributed by individial l, while further targeting the impute using individual i's covariate structure. If $\tilde{T}_i\{t^{sup}(\mathcal{S}_i)\} < X_i\{t^{sup}(\mathcal{S}_i)\}$, we sample another uniform(0,1), u, and repeat the process. This ensures that the impute occurs beyond the last observed time participant i was at risk for a recurrent event in the sup window.

We repeat the imputation procedure until we obtain M completed datasets and then analyze the M imputed datasets with methods guided by *Little and Rubin* (1986). For dataset m, we fit model (5.3) using GEE as described at the beginning of this section and obtain parameter estimates, $\hat{\beta}_m^{MI}$, and standard error estiamtes, $\hat{SE}(\hat{\beta}_m^{MI})$, for $m = 1, \ldots, M$. Then the final estimate of β from the multiple imputation procedure becomes $\hat{\beta}^{MI} = \sum_{m=1}^{M} \hat{\beta}_m^{MI}/M$ with corresponding standard error

estimate,

$$\hat{SE}(\hat{\beta}^{MI}) = \sqrt{\sum_{m=1}^{M} \hat{SE}(\hat{\beta}_m^{MI})^2/M + (1 + M^{-1}) \times \sum_{m=1}^{M} (\hat{\beta}_m^{MI} - \hat{\beta}^{MI})^2/(M-1)}.$$

5.5 Simulation

We now evaluate the finite sample performance of the proposed PO and MI methods for fitting equation (5.3) with simulated recurrent event data from $N = 500$ individuals over 5 years of follow-up. All simulation results are based on 10,000 iterations. Details of how recurrent events times are simulated are described in sections 5.5.1 and 5.5.2, where independent and dependent recurrent event distributions are considered, respectively.

In scenarios where censoring is present, the independent censoring random variable is $C_i = 5 \times I\{V_i > 5\} + V_i \times I\{V_i \leq 5\}, i = 1, \ldots, 500$, where V_i has an exponential distribution with hazard 1/14. This censoring mechanism corresponds to approximately 70% of participants having 5 years of follow-up, and 30% of participants being subject to censoring prior to 5 years. In addition to summarizing Model (5.3) results where censoring is handled via our customized PO and MI approaches, we report results in the case where outcomes are fully observed through 5 years for comparison.

In each simulation scenario, we build our longitudinal data structure with follow-up windows starting every $a = 1$ years apart at times $t = 0, 1, 2$ and 3 years with $\tau = 2$ years. These choices coincide with recommendations from *Xia and Murray* (2018) based on the recurrent event distributions used in simulation.

With four follow-up windows generating four longitudinal outcomes, we require a 4×4 working correlation structure to use with GEE software. We consider both (1) unstructured and (2) Toeplitz structures, with robust sandwich estimates ultimately

used in all inference. The Toeplitz structure takes the form

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_2 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Since only adjacent τ-restricted follow-up windows have overlap in our setting with $t = 0, 1, 2, 3$ and $\tau = 2$, this structure requires only two parameters, as opposed to six parameters used with the unstructured working correlation matrix.

5.5.1 Independent Times Between Recurrent Events

We first describe the scenario where times between recurrent events are independent. Recall from Section 5.3 and Appendix D.1, that when gap times between events for individual, i, are independent and identically distributed (i.i.d.) exponential(λ_i) random variables, then $T_i(t)$ is also marginally distributed as an exponential(λ_i) random variable for each $t \in \{t_1, \ldots, t_b\}$. Hence, if we generate times-to-first-event from i.i.d exponential gap times, the mean structure for $T_i(t), t \in \{t_1, \ldots, t_b\}$, will follow the same mean structure as the simulated gap times. This offers some computational convenience for generating outcomes that follow Model (5.3).

For the i^{th} individual, we allow λ_i to depend on two covariates, i.e., $Z_i = \{B_i, U_i\}$, where B_i a Bernoulli(0.5) random variable and U_i is a uniform(0,1) random variable. We generate mild correlation between B_i and U_i using a Gaussian copula approach (*Li*, 1999a). That is, we first generate bivariate normal(0, 1) pairs ($Q_{1i} = q_{1i}, Q_{2i} = q_{2i}$) with correlation 0.3. We then define $B_i = I(Q_{1i} \geq 0)$ and $U_i = P(Q_{2i} \leq q_{2i})$; the uniform(0, 1) distribution of U_i follows from the inverse transform theorem. Finally,

we generate times-to-first-event for patient i, $\mathcal{T}_i$, that satisfy model

$$E(log[min\{\tau, \mathcal{T}\}]|Z) = -0.7 + 0.5B_i + 0.5U_i.$$

This is accomplished by first simulating i.i.d. exponential(λ_i) gap times with λ_i taken as the numerical solution to

$$\int_{-\infty}^{log\tau^-} y\lambda_i e^y e^{-\lambda_i e^y} dy + e^{-\lambda_i \tau} \times log\tau = \beta_0 + \beta_1 B_i + \beta_2 U_i$$

and then converting the resulting recurrent event times into times-to-first event as described in Section 5.2.

Simulation results for the case with independent times between recurrent events are shown in Table 5.1. For each method and each coefficient, we present simulation averages for (1) $\hat{\beta}$, (2) bias $\hat{\beta} - \beta$ and estimated robust standard errors (SEs) assuming (3) unstructured or (4) Toeplitz working correlation matrices. We also report (5) the empirical standard deviation (ESD) of $\hat{\beta}$ across the 10,000 iterations and empirical coverage probabilities (CP) for the true coefficient using robust standard errors and either (6) unstructured or (7) Toeplitz working covariance matrices.

Both PO and MI approaches yield approximately unbiased estimates, with absolute bias <0.003. SE results are suitably close to ESD results to ensure that variability is being estimated well across all methods. As expected, standard errors are slightly smaller when GEE is fit in the uncensored case compared to censored cases, since more statistical information is available in the uncensored setting. However, there is not a clear winner between the PO and MI methods for handling the censored longitudinal data analysis. SEs attributed to the MI method are negligibly smaller than those using the PO method; all coverage probabilities are suitably close to 0.95. Both proposed PO and MI analysis methods perform well in our setting with 30%

Coef.	**Method**	$\hat{\beta}$	**Bias**	**ESD**	**SE** Unstr.	**SE** Toepl.	**CP** Unstr.	**CP** Toepl.
	Uncen.	-0.700	<0.001	0.057	0.057	0.057	0.945	0.944
$\beta_0 = -0.7$	PO	-0.702	-0.002	0.061	0.061	0.061	0.948	0.948
	MI	-0.698	0.002	0.060	0.060	0.060	0.947	0.948
	Uncen.	0.503	0.003	0.054	0.054	0.054	0.945	0.946
$\beta_1 = 0.5$	PO	0.503	0.003	0.058	0.057	0.057	0.942	0.946
	MI	0.502	0.002	0.058	0.057	0.057	0.942	0.944
	Uncen.	0.497	-0.003	0.089	0.092	0.092	0.958	0.959
$\beta_2 = 0.5$	PO	0.498	-0.002	0.096	0.098	0.099	0.951	0.951
	MI	0.497	-0.003	0.095	0.098	0.098	0.948	0.948

Table 5.1: Simulated Finite Sample Performance for $N = 500$ Individuals with Independently Generated Times Between Recurrent Events. Results Are Based on 10,000 Iterates.

(Coef.: True value of the coefficient;
For Methods, Uncen.: standard GEE approach applied to uncensored version of the data, PO: pseudo observation approach, MI: multiple imputation approach;
For remaining column headings, $\hat{\beta}$: average coefficient estimate; Bias: average $\hat{\beta} - \beta$; ESD: empirical standard deviation of $\hat{\beta}$; SE Unstr.: the average estimated robust standard error using an unstructured working correlation matrix; SE Toepl.: the average estimated robust standard error using a Toeplitz working correlation matrix; CP Unstr.: empirical coverage probability for true coefficient based on 95% confidence interval using robust standard error with an unstructured working correlation matrix; CP Toepl.: empirical coverage probability for true coefficient based on 95% confidence interval using robust standard error with an Toeplitz working correlation matrix.)

of patients censored. In practice, the PO method is particularly easy to program compared to the MI method and runs a bit more quickly, since the PO method is nested within the MI method. We suspect the PO method will be implemented more in practice as a result.

5.5.2 Simulating Distribution of Times-to-First-Event Based on Correlated Times Between Recurrent Events and Comparison of Proposed Methods

Simulating a multivariate time-to-first-event distribution is more complex when times between events are correlated random variables. Recall from Section 5.3 that positive correlation between exponential(λ_i) gap times causes the corresponding distribution for $\mathcal{T}$ to change; that $P\{T_i(t) > u\}$ tends to be smaller than an exponential(λ_i) survival function and stabilizes after approximately $t > 1/\lambda_i$ follow-up units of gap-time history has passed. The intuition behind this phenomenon, described in Section 5.3, also suggests the approach for successfully simulating the desired stable multivariable distribution for $\mathcal{T}$ to be used in this section. That is, upon simulating correlated exponential(λ_i) gap times for individual i, we discard at least the first $1/\lambda_i$ follow-up units of generated information, starting $t_1 = 0$ for individual i after this 'burn-in' period has passed.

To verify that our model works correctly for finite sample sizes when exponentially distributed times between event are correlated, we need (1) to generate data that follows model (5.3) for this setting and (2) have a way to verify that estimated parameters appropriately represent the data. To address (2), we assume a categorical predictor, $Z = \{0, 1, 2\}$, so that $E(log[min\{\tau, T_i(t)\}]|Z)$ can be consistently estimated from a large dataset (N=10,000) within each level of Z via an empirical mean. From this model-free process, we can determine values, $\tilde{\beta}$, of regression parameters that should be estimated if model (5.3) is working correctly. In particular, we assume that individuals with $Z_i = 0, 1$ or 2 have a history of exponential gap times with $\lambda_i = 1/2, 1/3$ or $1/5$, respectively, where correlation between any two gap times from individual i is approximately 0.8. Stabilization of the resulting multivariate time-to-first-event process is done by defining $t_1 = 0$ after a burn-in period of 5 follow-up units has passed for each individual.

A Gaussian copula approach (*Li*, 1999a) is used to generate correlated exponential gap times in this section and section 5.3. This approach first simulates mean zero multivariate normal random variables $\{Q_{i1}, Q_{i2}, \ldots, Q_{i500}\}$ with variance one and 0.8 correlation between Q_{ij} and $Q_{ij\prime}$, for $j \neq j\prime$; 500 was chosen to ensure that individuals would have at least 10 years of gap time history (5 year burn-in period, followed by 5 years of potential follow-up, subject to the previously described censoring mechanism). We then transform the multivariate normal random variables to multivariate uniform(0,1) and then multivariate exponential random variables via repeated applications of the inverse transform theorem. The exponential random variables that result from this process become the correlated gap times $\{G_{i1}, G_{i2}, \ldots, G_{i500}\}$, which are then converted into times-to-first event as described in Section 5.2.

Figure 5.4 shows the empirical average of $log[min\{2, T_i(t)\}]$ for each value of $Z = \{0, 1, 2\}$ based on $N = 10,000$ individuals with correlated exponential gap time histories as generated using the copula approach described above. As expected based on the stabilization of the survival curves seen in Figure 5.3, the empirical average of $log[min\{\tau, T_i(t)\}]$ seems to stabilize successfully after the 5 year burn-in period. Results from $t = 5$ to $t = 8$ in this figure are averaged to provide nonparametric large sample estimates, $\tilde{\beta}$, of parameters in model 5.3:

$$E(log[min\{\tau, \mathcal{T}\}]|Z) = -0.677 + 0.306I(Z = 1) + 0.637I(Z = 2).$$

Finite sample properties of our PO and MI methods shown in Table 5.2 are based on N=500 individuals simulated to have an equal chance of following the simulated time-to-first-event longitudinal data structure governed by covariate values, $Z = 0, 1$ or 2. The case where these subjects are not subject to censoring is also presented for comparison. Results are laid out in a similar manner to that seen in Table 5.1, except that bias is defined in relation to the nonparametric large sample estimate, $\tilde{\beta}$ rather

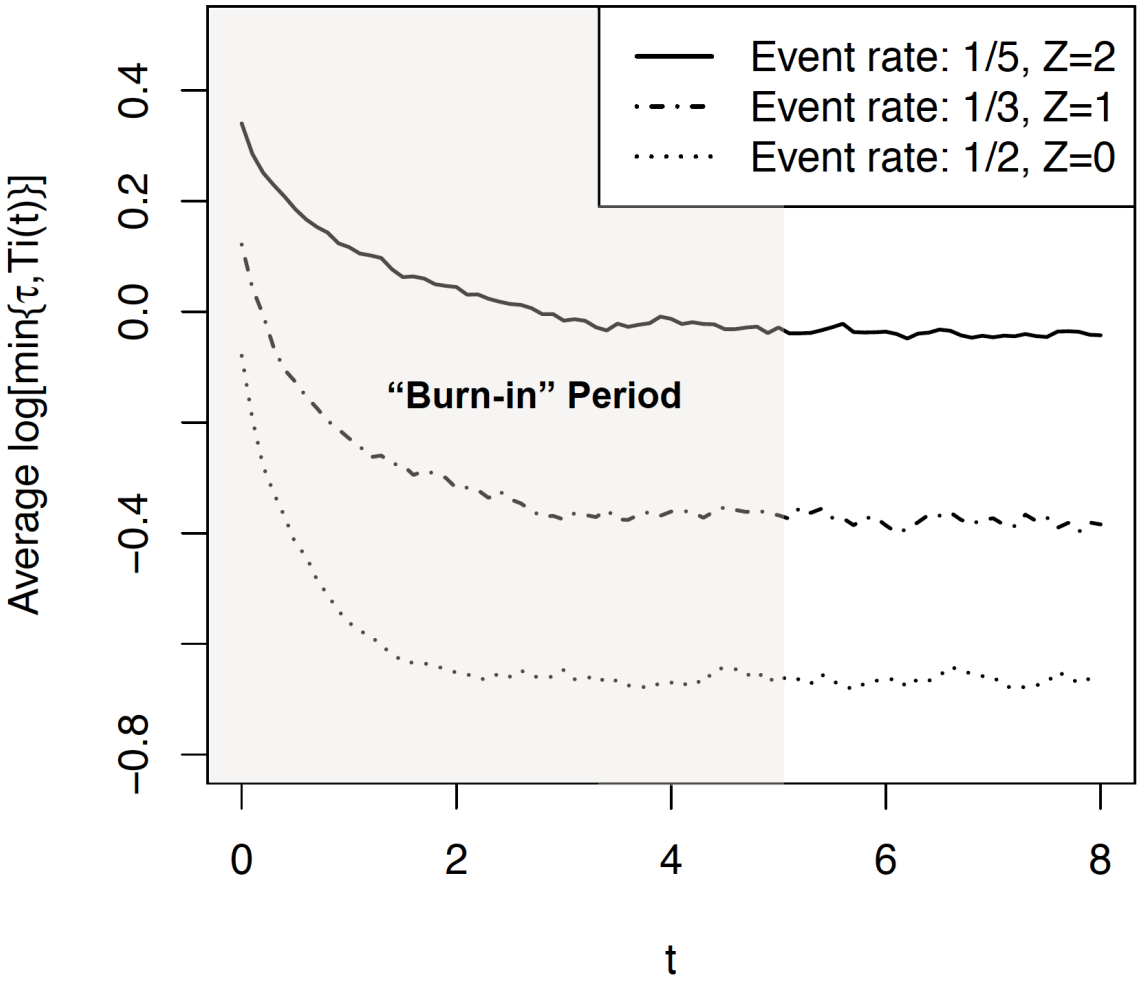


Figure 5.4: Empirical Average of $log[min\{2, T_i(t)\}]$, Based on N=10,000 Individuals with Correlated Exponential Gap Time Histories per Curve; Correlation is Approximately 0.8; $t \in \{0, \ldots, 8\}$; a= 0.1 Units Apart. Curves Seem to Stabilize after Shaded Burn-in Period of 5 Years.

Coef.	Method	$\hat{\beta}$	Bias	ESD	SE Unstr.	SE Toepl.	CP Unstr.	CP Toepl.
	Uncen.	-0.671	0.006	0.076	0.075	0.075	0.940	0.940
$\tilde{\beta}_0$=-0.677	PO	-0.669	0.008	0.078	0.078	0.078	0.943	0.944
	MI	-0.670	0.007	0.078	0.077	0.077	0.943	0.945
	Uncen.	0.302	-0.004	0.102	0.102	0.103	0.948	0.950
$\tilde{\beta}_1$=0.306	PO	0.300	-0.006	0.105	0.106	0.106	0.952	0.953
	MI	0.302	-0.004	0.105	0.106	0.106	0.951	0.950
	Uncen.	0.622	-0.015	0.099	0.097	0.097	0.945	0.946
$\tilde{\beta}_2$=0.637	PO	0.617	-0.020	0.101	0.100	0.101	0.944	0.946
	MI	0.621	-0.016	0.101	0.100	0.100	0.944	0.946

Table 5.2: Simulated Finite Sample Performance for $N = 500$ Individuals with Correlated Times Between Recurrent Events. Results Are Based on 10,000 Iterates.

(Coef.: True value of the coefficient;
For Methods, Uncen.: standard GEE approach applied to uncensored version of the data, PO: pseudo observation approach, MI: multiple imputation approach;
For remaining column headings, $\hat{\beta}$: average coefficient estimate; Bias: average $\hat{\beta} - \tilde{\beta}$; ESD: empirical standard deviation of $\hat{\beta}$; SE Unstr.: the average estimated robust standard error using an unstructured working correlation matrix; SE Toepl.: the average estimated robust standard error using a Toeplitz working correlation matrix; CP Unstr.: empirical coverage probability for true coefficient based on 95% confidence interval using robust standard error with an unstructured working correlation matrix; CP Toepl.: empirical coverage probability for true coefficient based on 95% confidence interval using robust standard error with an Toeplitz working correlation matrix.)

than a true β, since a closed form value for the true β is unavailable.

Results are generally comforting. Again the MI method has very slightly smaller SE estimates when compared to the PO method. All coverage probabilities are close to the desired 95%. Interesting, parameter estimate bias is smallest for the groups where the burn-in period of 5 years was longer than their respective $1/\lambda_i$ values would have suggested ($\hat{\beta}_0$ and $\hat{\beta}_1$). These results verify that both the PO and MI methods estimate parameters with good finite sample properties in this case, but also verify that we have successfully simulated this complex data structure.

5.6 Example

In this section, we use the proposed methods to analyze results from the Azithromycin in COPD Trial. This study followed 1112 patients with a history of acute exacerbations (AEs) for recurrent AEs after randomization to either placebo or 250mg azithromycin daily. This trial ended with favorable results for the azithromycin arm (*Albert et al.*, 2011), based on an analysis of the time to first acute exacerbation using the logrank test. Multivariable Cox proportional hazard analysis modeling time-to-first-exacerbation confirmed azithromycin benefit after adjustment for forced expiratory volume in one second (FEV_1), age, gender, smoking status and study sites.

In our analysis, we estimate parameters in Model (5.3) for $\tau = 6$ months and a longitudinal data structure, $\mathcal{T}$, measuring times-to-first-recurrent-event in follow-up windows starting at times $t = 0, 2, 4, 6$ months. Our selections of τ=6 and $a = 2$ months are based on the 6-month historic mean time-to-exacerbation in this patient population and recommendations from *Xia and Murray* (2018) that approximately 90% of recurrent events should be captured when spacing windows apart by one-third of a historic mean.

We present results from a univariate analysis that evaluates azithromycin versus placebo, a forest plot analysis of treatment effect in subgroups of interest, and a multivariable analysis of treatment effect that adjusts for age, gender, FEV1, smoking status and study site. We tested for and found no statistically significant interactions between follow-up window start times and treatment, indicating relatively stable patterns of treatment effect over time ($p > 0.41$). The Toeplitz working correlation structure gave a slightly lower QIC value compared to the unstructured working correlation structure in our multivariable model and was used in all models of the azithromycin data. All confidence intervals and p-values are based on robust sandwich estimation of variability.

Forest plots of univariate treatment effects, overall and by subgroup, are shown in

Figure 5.5 for the PO (left panel) and MI (right panel) methods. Tabulated versions of these results are located in Supplemental Table D.1 in Appendix D.3. Treatment effects are displayed on the scale of $e^{\hat{\beta}}$ and can be interpreted as multiplicative increases (or decreases) on the time to first exacerbation over the next 6 months of follow-up. Overall, azithromycin is estimated to extend the time to the first exacerbation over a 6 month period by approximately 14% using either the PO or the MI method (95% CI approximately 5%-24% longer, p=0.002 for PO method and p=0.001 for MI method). Stated as an absolute difference, there was an estimated 0.43 month increase in time-to-first-exacerbation for the azithromycin group compared to the placebo group over a 6-month period (I.e., $e^{\hat{\beta}_0+\hat{\beta}_1} - e^{\hat{\beta}_0} \approx 0.43$ using either the PO or MI method).

Across the various subgroup analyses shown in Figure 5.5 and Table D.1, the treatment benefit was most pronounced in COPD patients with better preserved lung function, that is, FEV1 % of predicted > 50 % (approximately 29% longer time to first exacerbation in the next 6 months, 95% CI 10%-51% longer using PO method and 11%-50% longer using MI method). In general, point estimates shown for the PO and MI methods in Table D.1 are very close to one another and 95% CI results for the methods are also close, but with slightly narrower CI widths using the MI approach. P-values for the MI method are also slightly smaller using the MI versus the PO method.

As seen in Table 5.3, the azythromycin group maintains its estimated treatment benefit when adjusted for confounders in a multivariable model using either the PO or MI method (approximately 15% longer time to first exacerbation in the next 6 months, 95% CI 6%-25% longer, p=0.001, using PO method and 6%-24% longer, p < 0.001, using MI method. Interactions between treatment and FEV_1, age, gender, smoking status other than study sites were tested and no significant interactions were found.

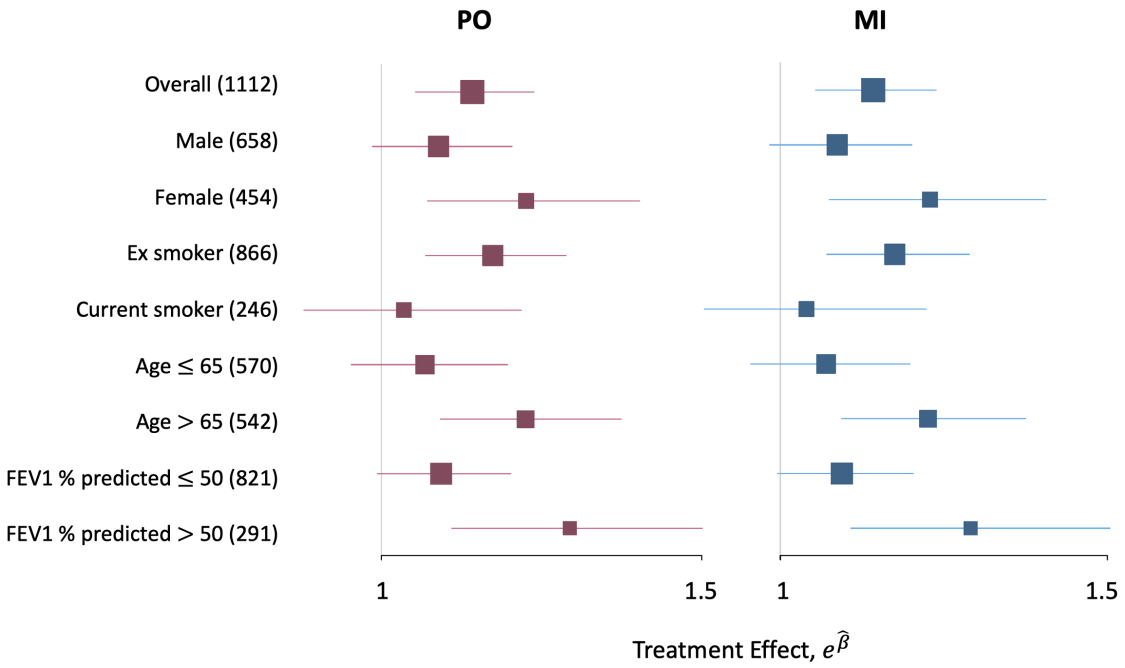


Figure 5.5: Forest Plot of Univariate Treatment Effects, Overall and by Subgroups of Interest.

	PO				MI			
	$e^{\hat{\beta}}$	95% CI		P	$e^{\hat{\beta}}$	95% CI		P
Intercept	60.34	41.17	88.43	<0.001	61.28	42.33	88.72	<0.001
Azithromycin (vs. Placebo)	1.153	1.063	1.250	0.001	1.150	1.063	1.244	<0.001
FEV_1 (per 10% Predicted)	1.041	1.013	1.068	0.003	1.039	1.013	1.066	0.003
Age (per 10 Years)	1.052	0.999	1.108	0.055	1.052	1.001	1.107	0.046
Male (vs. Female)	1.181	1.084	1.287	<0.001	1.171	1.078	1.272	<0.001
Current Smoker (vs. Ex)	1.074	0.967	1.193	0.184	1.071	0.967	1.186	0.188

Table 5.3: Multivariable Results Using PO and MI Methods. Displayed Estimates Are Additionally Adjusted for Center [Data Not Shown].
(CI: confidence interval; PO: pseudo-observation; MI: multiple imputation.)

5.7 Discussion

In this chapter, we take a fresh look at the manner in which recurrent event data is analyzed. By first restructuring the data into a censored longitudinal form, and then transforming the data via PO or MI models into a complete data format, we are able to take advantage of existing software from longitudinal data analysis literature. Our model estimates time free from recurrence over a τ-length follow-up period. In our opinion, this model gives a clear manner of assessing clinical and statistical significance of associations simultaneously. As with most longitudinal data, our method allows for either time-independent or dependent predictors.

We develop two methods for handling censoring that allow standard GEE methods to be applied to our censored longitudinal data structure: a PO approach and an MI approach. Both the methods have attractive performance in simulation, even with high correlation underlying the multivariate gap time distribution.

Experts in multiple imputation theory often prefer these methods to include a draw from the parameter space, going as far as to say that imputation must include this step to be proper. The argument is bolstered by noticeably improved coverage probabilities when this step is included in some cases. In our own work with inverse probability transform imputation methods, we have not observed a sufficient improvement in coverage probabilities to justify the extra computing time needed to perform this extra step. Although purists will likely agree to disagree, we feel comfortable recommending our imputation algorithm, as is, given the very solid coverage probability results seen in simulation.

An additional contribution of our research is a better understanding of how times-to-first-event fit within the context of recurrent events data. In most practical clinical settings, follow-up begins at a clinic visit that does not coincide with a recurrent event. When patients are scheduled for their next visit, likewise, this is unlikely to occur at a recurrent event time. Hence in practice, the information most relevant to a

patient is what to expect between clinic visits, and this is closer to a time-to-first-event analysis than a gap time analysis. By constructing a longitudinal data structure built from times-to-first-event, and evaluating this analysis in the presence of dependent gap times, we have introduced some needed intuition about how these distributions behave. The notion that the mixture of gap time histories prior to the moment a patient walks into a clinic for advice influences the time-to-first-event with correlated gap times, but not independent gap times, seems obvious in hindsight. However, simulating a stable time-to-first-event distribution required a better understanding of this process, and led to our suggestion of a burn-in period for simulating this data. We are hopeful that other researchers will benefit from this simulation framework alone.

CHAPTER VI

Conclusion

The objective of this dissertation is to develop study design and analysis methods for testing or predicting the τ-restricted survival/event-free time from a repurposed censored longitudinal data framework. We emphasize the advantage of using the restricted time as a more interpretable outcome, with less assumptions, and hope to contribute to methods for analyzing restricted survival/event-free times as alternative tools to hazard-oriented methods. To this end, we estimate the restricted time-to-first-event based on a paradigm shift from traditionally recorded censored time-to-event and recurrent events data. The philosophy hinges on the idea that traditional time data can be transformed into censored short-term outcomes measured longitudinally over potentially overlapping follow-up periods of length τ. Further tests and analysis of the τ-restricted survival/event-free time can benefit from shifting our thoughts on traditional time data structure towards the longitudinal outcomes.

In Chapter II-V, we consider three practical settings that can benefit statistically from repurposing traditional data into this censored longitudinal data structure. The first setting is clinical trial design and analysis where the outcome of interest is a single time-to-event compared between treatments. Chapter II develops a new two-sample test based on the restructured censored longitudinal data paradigm along with corresponding methods for group sequentially monitoring across several planned

interim analyses. Simulations suggest that shifting towards a longitudinal view of censored survival outcomes has practical advantages. Our proposed test procedure 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. When treatment differences emerge only after a certain period of time or in settings where there is potential for cure, we find our test has a notably improved performance over its competitors.

The second setting is also based on group sequential monitoring of clinical trial data, where instead of a single time-to-event endpoint the primary endpoints are recurrent in nature and can be subject to a terminating event. Chapter III develops group sequential methods for monitoring the *Tayob and Murray* (2014) statistic in this case. Our method is appropriate and robust for events that are correlated within an individual or for completely independent event times. Treatment effects observed across analysis times are simple to interpret. Besides, the assumption of proportionality between groups of the cumulative mean number of events over time is not required. Focusing on the time from the pre-specified windows' start instead of the previous event avoids dependent censoring issues. Chapter IV also provides a useful guide to help construct the data frame with more intelligence for the second setting. The recommendation is framed to seek a balance between the average proportion of recurrent events captured in at least one window and computational efficiency. We suggest a convenient rule of thumb of choosing spaced time that ensures at least 80% events are captured to achieve a decent power.

In our third and final setting, we develop multivariable models of recurrent events based on the censored longitudinal data framework in Chapter V. Instead of analyzing recurrent events based on intensity function or gap times, we propose to focus on the time free from recurrence over a prespecified follow-up period whenever an individual is at risk. Approaches of pseudo-observations and multiple imputations are utilized

to account for censoring and result a complete data set for further analysis with GEE model. We generalize and integrate the existing methods in a new way to analyze the censored longitudinal data transformed from recurrent events. Thus we provide a fresh perspective of assessing patients' progression status based on recurrent events. We also give a guide on how to implement and present our methods by the example data from the Azithromycin in COPD Trial.

This dissertation plus the researches from *Tayob and Murray* (2014, 2016, 2017) give a road map of the paradigm shifting from the traditionally recorded time data into a censored longitudinal data framework. To sum up, methods benefit from the data reconstruction include: (1) the improved estimation of τ-restricted mean survival time; (2) nonparametric two-sample tests of single time analysis and group sequentially monitoring for two types of data: the standard censored time-to-event data and recurrent events data subject to terminal events and censoring; and (3) multivariable methods for τ-restricted outcomes drawn from single time-to-event data as well as recurrent events data.

APPENDICES

APPENDIX A

Supplementary Materials for Chapter II

A.1 Derivation of the Asymptotic Joint Distribution of the Proposed Test Statistic at Interim Analysis Times

In this section, we derive the asymptotic joint distribution of the proposed test statistics, $\mathscr{T}(s_1), \ldots, \mathscr{T}(s_K)$ at interim analysis times, $s_1, \ldots, s_K$. The overall strategy is to first show that the vector of test statistics, $\{\mathscr{T}(s_1), \ldots, \mathscr{T}(s_K)\}$, is asymptotically equivalent in distribution to the more tractable vector of random variables, $\{\mathscr{T}^*(s_1), \ldots, \mathscr{T}^*(s_K)\}$, where elements of this latter vector are based on sums of independent and identically distributed quantities. From there, a standard application of the multivariate central limit theorem gives the desired result.

Our test statistic at analysis time s,

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

can be rewritten as

$$\mathscr{T}(s) = \sqrt{\frac{n_2(s)}{n_1(s)+n_2(s)}}\sqrt{n_1(s)}\hat{\mu}_1(s,\tau) - \sqrt{\frac{n_1(s)}{n_1(s)+n_2(s)}}\sqrt{n_2(s)}\hat{\mu}_2(s,\tau), \quad \text{(A.1)}$$

where $n_g(s)/\{n_1(s)+n_2(s)\} \xrightarrow{p} \pi_g(s)$. Suppose at analysis time s, combining information across b follow-up windows of length τ, we record M events $\{0 \equiv T_0 < T_1 < ... < T_M < T_{M+1} \equiv \tau\}$. Then, by Taylor series expansion,

$$\sqrt{n_g(s)}\hat{\mu}_g(s,\tau) = \sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)exp\left\{-\sum_{j=0}^{m}\frac{dN_g(s,T_j)}{Y_g(s,T_j)}\right\}$$

is asymptotically equivalent in distribution to the following terms:

$$\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\,exp\left\{-\sum_{j=0}^{m}\lambda_g^W(s,T_j)dT_j\right\} \tag{A.2}$$

$$+\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\left[\sum_{j=0}^{m}-exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\frac{dN_g(s,T_j)}{Y_g(s,T_j)}\right] \tag{A.3}$$

$$-\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\left[\sum_{j=0}^{m}-exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\lambda_g^W(s,T_j)dT_j\right] \tag{A.4}$$

$$\begin{aligned}&+\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\frac{1}{2!}exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\\&\left[\sum_{j=0}^{m}\left\{\frac{dN_g(s,T_j)}{Y_g(s,T_j)}-\lambda_g^W(s,T_j)dT_j\right\}\right]^2\end{aligned} \tag{A.5}$$

$$+\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)[\text{higher order terms}] \tag{A.6}$$

Using arguments similar to those in *Tayob and Murray* (2016) Appendix B, terms (A.5) and (A.6) converge to zero in probability. When there is no treatment effect (i.e., the null hypothesis is true), terms (A.2) and (A.4) for group $g=1$ will cancel the corresponding terms for group $g=2$ in the test statistic $\mathscr{T}(s)$. Hence, the asymptotic distribution of $\mathscr{T}(s)$ is based on the behavior of term (A.3) for groups $g=1,2$. Term

(A.3) can be further rewritten as

$$\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\Big[\sum_{j=0}^{m}-exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\frac{dN_g(s,T_j)}{Y_g(s,T_j)}\Big]$$

$$=-\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)exp\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\}\sum_{j=0}^{m}\frac{dN_g(s,T_j)}{Y_g(s,T_j)},$$

which is asymptotically equivalent in distribution (via Taylor series) to

$$-\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\times$$

$$\left\{\sum_{j=0}^{m}\frac{EdN_g(s,T_j)}{EY_g(s,T_j)}\right. \tag{A.7}$$

$$+\sum_{j=0}^{m}\left[\frac{1}{EY_g(s,T_j)}[dN_g(s,T_j)-EdN_g(s,T_j)]-\frac{EdN_g(s,T_j)}{EY_g(s,T_j)^2}[Y_g(s,T_j)-EY_g(s,T_j)]\right] \tag{A.8}$$

$$\left.+[\text{higher order terms}]\right\}. \tag{A.9}$$

Using arguments similar to those in *Tayob and Murray* (2016) Appendix B once again, the higher order terms in (A.9) converge to zero in probability. In addition when the null hypothesis is true, term (A.7) for group $g=1$ will cancel with its counterpart term for $g=2$ in the test statistic $\mathscr{T}(s)$. Hence, the asymptotic distribution of $\mathscr{T}(s)$ is based on the behavior of term (A.8) for groups $g=1,2$ which upon noting that $EdN_g(s,T_j)/EY_g(s,T_j)=\lambda_g^W(s,T_j)$ and $EY_g(s,T_j)=\sum_{l=1}^{b}Pr(X_{gi}(s,t_l)\geq T_j)$ can be algebraically rearranged as:

$$-\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\sum_{j=0}^{m}\frac{dN_g(s,T_j)-Y_g(s,T_j)\lambda_g^W(s,T_j)}{\sum_{l=1}^{b}Pr(X_{gi}(s,t_l)\geq T_j)}$$

or returning to more standard stochastic integral notation as:

$$-\sqrt{n_g(s)}\int_0^{\tau} exp\left\{-\int_0^{u_2}\lambda_g^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_g(s,u_1)-Y_g(s,u_1)\lambda_g^W(s,u_1)}{\sum_{l=1}^{b}Pr(X_{gi}(s,t_l)\geq u_1)}du_2. \tag{A.10}$$

Summarizing calculations from equation (A.1) to equation (A.10),

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

is asymptotically equivalent in distribution to

$$\sqrt{\pi_1(s)}\sqrt{n_2(s)}\int_0^{\tau} exp\left\{-\int_0^{u_2}\lambda_2^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_2(s,u_1)-Y_2(s,u_1)\lambda_2^W(s,u_1)}{\sum_{l=1}^{b}Pr(X_{2i}(s,t_l)\geq u_1)}du_2$$

$$-\sqrt{\pi_2(s)}\sqrt{n_1(s)}\int_0^{\tau} exp\left\{-\int_0^{u_2}\lambda_1^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_1(s,u_1)-Y_1(s,u_1)\lambda_1^W(s,u_1)}{\sum_{l=1}^{b}Pr(X_{1i}(s,t_l)\geq u_1)}du_2.$$

From here, we note that the remaining terms above can be written in terms of independent and identically distributed random variables that lend themselves to standard limiting distribution results via the multivariate central limit theorem. Recall that

$$N_g(s,u)=\sum_{i=1}^{n_g(s)}N_{gi}(s,u)=\sum_{i=1}^{n_g(s)}\sum_{j=1}^{b}N_{gi}(s,t_j,u)$$

and

$$Y_g(s,u)=\sum_{i=1}^{n_g(s)}Y_{gi}(s,u)=\sum_{i=1}^{n_g(s)}\sum_{j=1}^{b}Y_{gi}(s,t_j,u).$$

Define:

$$Z_{ij}\{\hat{\mu}_g(s,\tau)\}=\int_0^{\tau} exp\left\{-\int_0^{u_2}\lambda_g^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_{gi}(s,t_j,u_1)-Y_{gi}(s,t_j,u_1)\lambda_g^W(s,u_1)du_1}{\sum_{l=1}^{b}Pr(X_{gi}(s,t_l)\geq u_1)}du_2$$

and

$$Z_i\{\hat{\mu}_g(s,\tau)\} = \sum_{j=1}^{b} Z_{ij}\{\hat{\mu}_g(s,\tau)\}.$$

Note that $Z_i\{\hat{\mu}_g(s,\tau)\}$ only depends on patient i and is independent and identically distributed for $i = 1, \ldots, n_g(s)$. Using this notation, the above asymptotically equivalent representation of the distribution of $\mathscr{T}(s)$ becomes

$$\mathscr{T}^*(s) = \sqrt{\pi_1(s)}\sqrt{n_2(s)}\frac{\sum_{i=1}^{n_2(s)} Z_i\{\hat{\mu}_2(s,\tau)\}}{n_2(s)} - \sqrt{\pi_2(s)}\sqrt{n_1(s)}\frac{\sum_{i=1}^{n_1(s)} Z_i\{\hat{\mu}_1(s,\tau)\}}{n_1(s)}. \tag{A.11}$$

Application of the multivariate central limit theorem to the vector of test statistics $\{\mathscr{T}^*(s_1), \ldots, \mathscr{T}^*(s_K)\}$ calculated at calendar times, $s_1, s_2, \ldots, s_K$ (K finite), gives a limiting multivariate normal distribution as $n_g(s_1) \to \infty, g = 1, 2$, with asymptotic covariance matrix estimated empirically as described in Appendix A.2. A closed-form version of the asymptotic covariance is described in Appendix A.3.

For convenience, we explicitly describe the special case where only a single analysis is performed. When the null hypothesis is true, the asymptotic limiting distribution of $\mathscr{T}(s)$ is Normal with mean 0 and variance $\pi_2(s)\sigma_1^2(s)+\pi_1(s)\sigma_2^2(s)$, where $\sigma_g^2(s), g = 1, 2$ is the variance of $Z_i(\hat{\mu}_g(s,\tau))$ and can be estimated using the sampling variability of $Z_i\{\hat{\mu}_g(s,\tau)\}$, that is, $\hat{\sigma}_g^2(s) = \sum_{i=1}^{n_g(s)}[z_i\{\hat{\mu}_g(s,\tau)\} - \bar{z}\{\hat{\mu}_g(s,\tau)\}]^2/[n_g(s)-1]$, where

$$z_i\{\hat{\mu}_g(s,\tau)\} = \sum_{j=1}^{b} z_{ij}\{\hat{\mu}_g(s,\tau)\}; \qquad \bar{z}\{\hat{\mu}_g(s,\tau)\} = \sum_{i=1}^{n_g(s)} z_i\{\hat{\mu}_g(s,\tau)\}/n_g(s)$$

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

$$\int_0^{\tau} exp\left\{-\int_0^{u_2} \frac{dN_g(s,u_1)}{Y_g(s,u_1)}\right\}\left\{\int_0^{u_2} \frac{dN_{gi}(s,t_j,u_1) - Y_{gi}(s,t_j,u_1)\frac{dN_g(s,u_1)}{Y_g(s,u_1)}}{Y_g(s,u_1)/n_g(s)}\right\} du_2. \tag{A.12}$$

For finite sample sizes, we use a standardized version of the test statistic,

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

which has an approximate Normal(0,1) distribution, with critical values of $\pm$ 1.96 conferring an overall type I error of 5% when a single analysis is performed.

A.2 Empirical Covariance Matrix for $\left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_K)\right\}$

In this appendix, we describe how to estimate the empirical version of the $K \times K$ asymptotic covariance matrix, Σ, corresponding to standardized test statistics, $\left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_K)\right\}$. By design, diagonal elements of this matrix are equal to one, so that this covariance matrix is also a correlation matrix. Off-diagonal elements, $\sigma_{k_1k_2} = \sigma_{k_2k_1}, k_1 < k_2$, can be estimated based on the more updated dataset at analysis s_{k_2}.

In Appendix A.1, we show that $\{\mathscr{T}(s_1), \ldots, \mathscr{T}(s_K)\}$ is asymptotically equivalent in distribution to $\{\mathscr{T}^*(s_1), \ldots, \mathscr{T}^*(s_K)\}$. Similarly, for the standardized version of each test statistic, $\tilde{\mathscr{T}}(s_k), s_k = s_1, \ldots, s_K$, we work with the corresponding asymptotically equivalent in distribution standardized form, $\mathscr{T}^*(s_k)/\sqrt{\pi_2(s_k)\sigma_1^2(s_k) + \pi_1(s_k)\sigma_2^2(s_k)}$. Hence, off-diagonal elements $\sigma_{k_1k_2} = \sigma_{k_2k_1}, k_1 < k_2$, of the covariance matrix, Σ, can be estimated by

$$\hat{\sigma}_{k_1k_2} = \frac{\hat{Cov}\left\{\mathscr{T}^*(s_{k_1}), \mathscr{T}^*(s_{k_2})\right\}}{\sqrt{\hat{\pi}_2(s_{k_1})\tilde{\sigma}_1^2(s_{k_1}) + \hat{\pi}_1(s_{k_1})\tilde{\sigma}_2^2(s_{k_1})}\sqrt{\hat{\pi}_2(s_{k_2})\hat{\sigma}_1^2(s_{k_2}) + \hat{\pi}_1(s_{k_2})\hat{\sigma}_2^2(s_{k_2})}} \quad \text{(A.13)}$$

We define each component of $\hat{\sigma}_{k_1k_2}$ in more detail below.

Estimated terms that use the most up-to-date information at analysis time s_{k_2} have already been described for $\hat{\sigma}_g^2(s_{k_2}), g = 1, 2$, in Appendix A.1, captured by terms in equation (A.12). Appendix A.1 also defines $\hat{\pi}_g(s_k) = n_g(s_k)/\{n_1(s_k) + n_2(s_k)\}$,

for $g = 1, 2$ and $s_k = s_{k_1}, s_{k_2}$. Estimates of $\sigma_g^2(s_{k_1}), g = 1, 2$ used in the covariance estimate are modified to take advantage of additional information available at s_{k_2} for estimating terms that do not depend on analysis time. In particular, since both $dN_{gi}(s_{k_1}, t_j, u_1)/Y_{gi}(s_{k_1}, t_j, u_1)$ and $dN_{gi}(s_{k_2}, t_j, u_1)/Y_{gi}(s_{k_2}, t_j, u_1)$ estimate $\lambda_{gi}(t_j, u_1)du_1$, and the latter term uses more data, we replace $dN_{gi}(s_{k_1}, t_j, u_1)$ in equation (1) with $Y_{gi}(s_{k_1}, t_j, u_1) \times\ dN_{gi}(s_{k_2}, t_j, u_1)/Y_{gi}(s_{k_2}, t_j, u_1)$. Similarly in equation (A.12), we replace $Y_g(s_{k_1}, u_1)/n_g(s_{k_1})$, which is an estimate of $\sum_{l=1}^{b} Pr(T_{gi}(s_{k_1}, t_l) \geq u_1) Pr(C_{gi}(s_{k_1}, t_l) \geq u_1)$, with $\left[\sum_{i=1}^{n_g(s_{k_2})} I\{T_{gi} \geq u_1 + t_l\}/n_g(s_{k_2})\right]$ $\times \left[\sum_{i=1}^{n_g(s_{k_1})} I\{C_{gi}(s_{k_1}) \geq u_1 + t_l\}/n_g(s_{k_1})\right]$. Here, terms involving the event time are estimated using updated data, while terms involving the censoring distribution remain relevant to analysis time s_{k_1}. Putting these modifications together gives us

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

as an updated version of $z_{ij}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ for use in covariance terms. And mimicking Appendix A.1 notation, $\tilde{\sigma}_g^2(s_{k_1})$ used in equation (A.13) is calculated by replacing $z_{ij}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ with $\tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1}, \tau)\}$ terms in corresponding formulas for $\hat{\sigma}_g^2(s_{k_1})$ from Appendix A.1.

The only remaining undefined term from equation (A.13) is $\hat{Cov}\left\{\mathscr{T}^*(s_{k_1}), \mathscr{T}^*(s_{k_2})\right\}$, which is described in the following. From equation (A.11),

$$Cov\left\{\mathscr{T}^*(s_{k_1}), \mathscr{T}^*(s_{k_2})\right\}$$

$$= \sum_{g=1}^{2} Cov\left[\sqrt{\pi_{3-g}(s_{k_1})} \sqrt{n_g(s_{k_1})} \frac{\sum_{i=1}^{n_g(s_{k_1})} Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}}{n_g(s_{k_1})}, \right.$$

$$\sqrt{\pi_{3-g}(s_{k_2})}\sqrt{n_g(s_{k_2})}\frac{\sum_{i=1}^{n_g(s_{k_2})} Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}}{n_g(s_{k_2})}\Bigg].$$

Without loss of generality, assume $k_1 \leq k_2$ so that $n_g(s_{k_1}) \leq n_g(s_{k_2})$ and there are $n_g(s_{k_1})$ patients contributing (correlated) data from both analysis times. Then the previous expression reduces to

$$= \sum_{g=1}^{2} \sqrt{\pi_{3-g}(s_{k_1})}\sqrt{\pi_{3-g}(s_{k_2})}\frac{n_g(s_{k_1})}{\sqrt{n_g(s_{k_1})n_g(s_{k_2})}} Cov\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right],$$

which is asymptotically equivalent to

$$= \sum_{g=1}^{2} \sqrt{\pi_{3-g}(s_{k_1})\pi_{3-g}(s_{k_2})\psi_g(s_{k_1},s_{k_2})} Cov\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right]$$

where $\psi_g(s_{k_1}, s_{k_2})$ is the limiting proportion of patients entered at s_{k_1} of those eventually entered by s_{k_2} of group g, that is estimated by $n_g(s_{k_1})/n_g(s_{k_2})$.

In practice, $Cov\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right]$ can be estimated based on the empirical covariance of sample realizations of $Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}$ and $Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}$, that is,

$$\hat{Cov}\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right] =$$

$$\sum_{i=1}^{n_g(s_{k_1})} \frac{[\tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\} - \bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1},\tau)\}][z_i\{\hat{\mu}_g(s_{k_2},\tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2},\tau)\}]}{n_g(s_{k_1}) - 1}, \text{ where}$$

$\tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\} = \sum_{j=1}^{b} \tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$, $\bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1},\tau)\} = \sum_{i=1}^{n_g(s_{k_1})} \tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\}/n_g(s_{k_1})$. Putting each described component into equation (A.13), we have the version of $\hat{\sigma}_{k_1k_2}$ listed in Chapter 2.4.

A.3 Closed Form Covariance Matrix for $\left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_K)\right\}$

At times it is convenient to have an asymptotic closed form version of the covariance matrix for $\left\{\tilde{\mathscr{T}}(s_1), \ldots, \tilde{\mathscr{T}}(s_K)\right\}$, for instance in assessing whether an inde-

pendent increments variability structure is present. Working from results in the last paragraph of Appendix A.2, instead of estimating $Cov\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right]$ with the empirical covariance, in this section we derive its asymptotic closed-form formula. Consider $Z_i\{\hat{\mu}_g(s_k,\tau)\} = \sum_{j=1}^{b} Z_{ij}\{\hat{\mu}_g(s_k,\tau)\}$ at analysis times $s_k = s_{k_1}$ and s_{k_2} and recall that group g patients are independent and identically distributed. Then

$$Cov\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right] = \sum_{j=1}^{b}\sum_{j\prime=1}^{b} Cov\left[Z_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_{ij\prime}\{\hat{\mu}_g(s_{k_2},\tau)\}\right].$$

For notational simplicity, we submerge the group indicator g as we work with the summand term above. That is,

$$\begin{aligned}
&Cov\left[Z_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_{ij\prime}\{\hat{\mu}_g(s_{k_2},\tau)\}\right] \\
=&\sum_{j=1}^{b}\sum_{j\prime=1}^{b}\int_0^{\tau}\int_0^{\tau} exp\{-\int_0^{u_2}\lambda^W(s_{k_1},u_1)du_1\}exp\{-\int_0^{v_2}\lambda^W(s_{k_2},v_1)dv_1\} \\
&\times\int_0^{u_2}\int_0^{v_2}\frac{1}{\sum_l Pr(X_i(s_{k_1},t_l)\geq u_1)\sum_{l\prime} Pr(X_i(s_{k_2},t_{l\prime})\geq v_1)} \\
&\times Cov\Big\{dN_i(s_{k_1},t_j,u_1) - Y_i(s_{k_1},t_j,u_1)\lambda^W(s_{k_1},u_1)du_1, dN_i(s_{k_2},t_{j\prime},v_1) \\
&\qquad - Y_i(s_{k_2},t_{j\prime},v_1)\lambda^W(s_{k_2},v_1)dv_1\Big\}du_2dv_2.
\end{aligned}$$

Focusing on this last term:

$$\begin{aligned}
&Cov\Big\{dN_i(s_{k_1},t_j,u_1) - Y_i(s_{k_1},t_j,u_1)\lambda^W(s_{k_1},u_1)du_1, \\
&\qquad dN_i(s_{k_2},t_{j\prime},v_1) - Y_i(s_{k_2},t_{j\prime},v_1)\lambda^W(s_{k_2},v_1)dv_1\Big\} \\
&\qquad = E[dN_i(s_{k_1},t_j,u_1)dN_i(s_{k_2},t_{j\prime},v_1)] \qquad \text{(A.14)} \\
&\qquad -\lambda^W(s_{k_1},u_1)E[Y_i(s_{k_1},t_j,u_1)dN_i(s_{k_2},t_{j\prime},v_1)]du_1 \qquad \text{(A.15)}
\end{aligned}$$

$$-\lambda^W(s_{k_2}, v_1)E[Y_i(s_{k_2}, t_{j\prime}, v_1)dN_i(s_{k_1}, t_j, u_1)]dv_1 \tag{A.16}$$

$$+\lambda^W(s_{k_1}, u_1)\lambda^W(s_{k_2}, v_1)E[Y_i(s_{k_1}, t_j, u_1)Y_i(s_{k_2}, t_{j\prime}, v_1)]du_1dv_1 \tag{A.17}$$

$$-E[dN_i(s_{k_1}, t_j, u_1) - Y_i(s_{k_1}, t_j, u_1)\lambda^W(s_{k_1}, u_1)du_1] \tag{A.18}$$

$$\times E[dN_i(s_{k_2}, t_{j\prime}, v_1) - Y_i(s_{k_2}, t_{j\prime}, v_1)\lambda^W(s_{k_2}, v_1)dv_1]. \tag{A.19}$$

Term (A.14) becomes:

$$\begin{aligned}
E[dN_i(s_{k_1}, t_j, u_1)dN_i(s_{k_2}, t_{j\prime}, v_1)] &= \lim_{\Delta u_1, \Delta v_1 \to 0} Pr\{u_1 \le X_i(s_{k_1}, t_j) < u_1 + \Delta u_1, \\
&\qquad \delta_i(s_{k_1}, t_j) = 1, \\
&\qquad v_1 \le X_i(s_{k_2}, t_{j\prime}) < v_1 + \Delta v_1, \\
&\qquad \delta_i(s_{k_2}, t_{j\prime}) = 1\} \\
&= \lim_{\Delta u_1 \to 0} Pr\{u_1 \le X_i(s_{k_1}, t_j) < u_1 + \Delta u_1, \delta_i(s_{k_1}, t_j) = 1\} \\
&\qquad \times I\{u_1 + t_j = v_1 + t_{j\prime}\} \\
&= \lambda(s_{k_1}, t_j, u_1)Pr\{X_i(s_{k_1}, t_j) \ge u_1\} \\
&\qquad \times I\{u_1 + t_j = v_1 + t_{j\prime}\}du_1.
\end{aligned}$$

Term (A.15) becomes

$$\begin{aligned}
E[Y_i(s_{k_1}, t_j, u_1)dN_i(s_{k_2}, t_{j\prime}, v_1)] &= \lim_{\Delta v_1 \to 0} Pr\{X_i(s_{k_1}, t_j) \ge u_1, v_1 \le X_i(s_{k_2}, t_{j\prime}) < v_1 + \Delta v_1, \\
&\qquad \delta_i(s_{k_2}, t_{j\prime}) = 1\} \\
&= \lambda(s_{k_2}, t_{j\prime}, v_1)Pr\{X_i(s_{k_1}, t_j) \ge u_1, X_i(s_{k_2}, t_{j\prime}) \ge v_1\} \\
&\qquad [I\{u_1 + t_j \le v_1 + t_{j\prime}\} + I\{u_1 = 0, t_j > v_1 + t_{j\prime}\}]dv_1,
\end{aligned}$$

where the expectation is only none-zero when $u_1 + t_j \le v_1 + t_{j\prime}$. The term $I\{u_1 = 0, t_j > v_1 + t_{j\prime}\}$ comes from the case when the failure occurs before calendar time t_j,

namely $t_j > v_1 + t_{j\prime}$, by definition $X_i(s_{k_1}, t_j) = 0$. Therefore the expectation is also non-zero when $u_1 = 0$.

Term (A.16) becomes

$$\begin{aligned} E[Y_i(s_{k_2}, t_{j\prime}, v_1) dN_i(s_{k_1}, t_j, u_1)] = \lim_{\Delta u_1 \to 0} & Pr\{X_i(s_{k_2}, t_{j\prime}) \geq v_1, u_1 \leq X_i(s_{k_1}, t_j) < u_1 + \Delta u_1, \\ & \delta_i(s_{k_1}, t_j) = 1\} \\ = & \lambda(s_{k_1}, t_j, u_1) Pr\{X_i(s_{k_1}, t_j) \geq u_1, X_i(s_{k_2}, t_{j\prime}) \geq v_1\} \\ & [I\{u_1 + t_j \geq v_1 + t_{j\prime}\} + I\{v_1 = 0, u_1 + t_j < t_{j\prime}\}] du_1, \end{aligned}$$

where the expectation is only none-zero when $u_1 + t_j \geq v_1 + t_{j\prime}$. The term $I\{v_1 = 0, u_1 + t_j < t_{j\prime}\}$ comes from the case when the failure occurs before calendar time $t_{j\prime}$, namely $u_1 + t_j < t_{j\prime}$, by definition $X_i(s_{k_2}, t_{j\prime}) = 0$. Therefore the expectation is also non-zero when $v_1 = 0$.

Term (A.17) becomes

$$E[Y_i(s_{k_1}, t_j, u_1) Y_i(s_{k_2}, t_{j\prime}, v_1)] = Pr\{X_i(s_{k_1}, t_j) \geq u_1, X_i(s_{k_2}, t_{j\prime}) \geq v_1\}.$$

Term (A.18) becomes

$$\begin{aligned} E[dN_i(s_{k_1}, t_j, u_1) - Y_i(s_{k_1}, t_j, u_1) \lambda^W(s_{k_1}, u_1) du_1] = & [\lambda(s_{k_1}, t_j, u_1) - \lambda^W(s_{k_1}, u_1)] \\ & \times Pr\{X_i(s_{k_1}, t_j) \geq u_1\} du_1. \end{aligned}$$

And term (A.19) becomes

$$\begin{aligned} E[dN_i(s_{k_2}, t_{j\prime}, v_1) - Y_i(s_{k_2}, t_{j\prime}, v_1) \lambda^W(s_{k_2}, v_1) dv_1] = & [\lambda(s_{k_2}, t_{j\prime}, v_1) - \lambda^W(s_{k_2}, v_1)] \\ & \times Pr\{X_i(s_{k_2}, t_{j\prime}) \geq v_1\} dv_1. \end{aligned}$$

Substituting appropriate terms we now have

$$
\begin{aligned}
&Cov\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right] \\
&= \sum_{j=1}^{b}\sum_{j\prime=1}^{b} \int_0^{\tau}\int_0^{\tau} exp\{-\int_0^{u_2} \lambda_g^W(s_{k_1},u_1)du_1\} exp\{-\int_0^{v_2} \lambda_g^W(s_{k_2},v_1)dv_1\} \\
&\quad \times \int_0^{u_2}\int_0^{v_2} \frac{1}{\sum_l Pr(X_{gi}(s_{k_1},t_l)\geq u_1)\sum_{l\prime} Pr(X_{gi}(s_{k_2},t_{l\prime})\geq v_1)} \\
&\quad \times \Big\{ \lambda_g(t_j,u_1) Pr\{X_{gi}(s_{k_1},t_j)\geq u_1\} I\{u_1+t_j = v_1+t_{j\prime}\}du_1 \\
&\quad - \Big[\lambda_g^W(s_{k_1},u_1)\lambda(t_{j\prime},v_1)[I\{u_1+t_j\leq v_1+t_{j\prime}\} + I\{u_1=0, t_j > v_1+t_{j\prime}\}] \\
&\qquad + \lambda_g^W(s_{k_2},v_1)\lambda(t_j,u_1)[I\{u_1+t_j\geq v_1+t_{j\prime}\} + I\{v_1=0, u_1+t_j < t_{j\prime}\}] \\
&\qquad - \lambda_g^W(s_{k_1},u_1)\lambda^W(s_{k_2},v_1)\Big] Pr\{X_{gi}(s_{k_1},t_j)\geq u_1, X_{gi}(s_{k_2},t_{j\prime})\geq v_1\}du_1dv_1 \\
&\quad - \{\lambda_g(t_j,u_1) - \lambda_g^W(s_{k_1},u_1)\}\{\lambda_g(t_{j\prime},v_1) - \lambda_g^W(s_{k_2},v_1)\} \\
&\qquad \times Pr\{X_{gi}(s_{k_1},t_j)\geq u_1\} Pr\{X_{gi}(s_{k_2},t_{j\prime})\geq v_1\}du_1dv_1 \Big\} du_2dv_2
\end{aligned}
$$

Unfortunately, this covariance does not simplify to an independent increments structure except in special cases such as an exponentially distributed event time. The independent increments structure emerges in this special case upon noting that $\lambda_g^W(s,u) = \lambda(t,u) = \lambda$ for all s,t and u. However, given the advantages of avoiding parametric assumptions, there is no practical computation savings that can be made from knowledge of this special case.

We've also used this asymptotic closed form variance as a method to double-check that R code for our empirically calculated covariance is on target. For example, a covariance matrix estimated from 500 individuals' data should be relatively close to the asymptotic closed form. Assuming an $Exp(0.5)$ event time with 2 years of uniform accrual, and analyses using $\tau = 1$ conducted at 1, 2, 3, 4 and 5 years in calendar time,

the closed form covariance matrix calculation gives:

$$\begin{bmatrix} 0.175 & 0.085 & 0.042 & 0.039 & 0.035 \\ 0.085 & 0.091 & 0.052 & 0.041 & 0.036 \\ 0.042 & 0.052 & 0.054 & 0.042 & 0.038 \\ 0.039 & 0.041 & 0.042 & 0.043 & 0.038 \\ 0.035 & 0.036 & 0.038 & 0.038 & 0.039 \end{bmatrix},$$

whereas the corresponding empirical covariance estimate from the 500 individuals was

$$\begin{bmatrix} 0.149 & 0.082 & 0.050 & 0.040 & 0.034 \\ 0.082 & 0.104 & 0.056 & 0.044 & 0.039 \\ 0.050 & 0.056 & 0.059 & 0.045 & 0.041 \\ 0.040 & 0.044 & 0.045 & 0.046 & 0.042 \\ 0.034 & 0.039 & 0.041 & 0.042 & 0.043 \end{bmatrix}$$

with difference matrix

$$\begin{bmatrix} -0.026 & -0.003 & 0.008 & 0.001 & -0.001 \\ -0.003 & 0.013 & 0.004 & 0.003 & 0.003 \\ 0.008 & 0.004 & 0.005 & 0.003 & 0.003 \\ 0.001 & 0.003 & 0.003 & 0.003 & 0.004 \\ -0.001 & 0.003 & 0.003 & 0.004 & 0.004 \end{bmatrix}.$$

Repeating this exercise for different simulated datasets and sample sizes is a comforting coding check.

A.4 Supplementary Simulation Results of Section 2.5

In this section we show supplemental simulation results for our proposed method using the same simulation scenarios 1-9 described in the Chapter 2.5. In Tables A.1 and A.2, we (1) examine the performance of our method for alternative choices of $\tau = 0.25, 0.50$ and 0.75 years, (2) show results for the Peto and Peto (WLR-PP) test that places more weight on hazards at the beginning of the study and (3) show results for the Fleming-Harrington (WLR-FH) $(0.5, 0.5)$ test that places more weight on hazards at the end of the study. Table A.1 shows stopping rates based on OF efficacy, JT safety, Pocock safety and OF safety bounds. Table A.2 shows the average study time (AST) in years, the average sample number (ASN) and the average number of events (ANE).

All test statistic boundaries meet their targets within simulation error under Scenario 1, the null hypothesis (Table A.1, Scenario 1).

For the most part, stopping rates do not seem to vary much based on the selection for τ. The only possible exception is in Scenario 4, the delayed treatment effect scenario, where power is slightly smaller for smaller values of τ. The WLR-FH test does well in this setting, with slightly less power than the proposed test using $\tau = 1$ year and slightly more power than the proposed test with smaller values of τ. The WLR-PP test has much lower power than all other methods in this setting. The WLR-PP test also performs poorly in Scenario 8, the Scenario with mixed cure distribution alternatives under consideration.

Note that these extra simulations for $\tau = 0.25, 0.5$ and 0.75 are not intended to be an exhaustive look at how to choose τ since we believe most applications will have a natural choice. But these additional simulations verify that the method performs well for a broader selection of short-term window lengths.

Scenario	Test Statistic	OF Efficacy	JT Safety	P Safety	OF Safety
1	Proposed $\tau = 0.75$	0.024	0.192	0.025	0.024
	Proposed $\tau = 0.5$	0.023	0.197	0.024	0.023
	Proposed $\tau = 0.25$	0.024	0.193	0.024	0.024
	WLR-PP	0.023	0.195	0.025	0.026
	WLR-FH $(0.5, 0.5)$	0.023	0.196	0.026	0.026
2	Proposed $\tau = 0.75$	0.813	0	0	0
	Proposed $\tau = 0.5$	0.803-0.804	0.002	0	0
	Proposed $\tau = 0.25$	0.806-0.807	0.002	0	0
	WLR-PP	0.75	0	0	0
	WLR-FH $(0.5, 0.5)$	0.807	0	0	0
3	Proposed $\tau = 0.75$	0	0.977	0.79	0.847
	Proposed $\tau = 0.5$	0	0.979	0.773	0.829
	Proposed $\tau = 0.25$	0	0.973	0.778	0.839
	WLR-PP	0	0.967	0.724	0.76
	WLR-FH $(0.5, 0.5)$	0	0.971	0.773	0.815
4	Proposed $\tau = 0.75$	0.813-0.825	0.024	0.007	0
	Proposed $\tau = 0.5$	0.817-0.824	0.019	0.005	0
	Proposed $\tau = 0.25$	0.803-0.811	0.025	0.007	0
	WLR-PP	0.367	0.029	0.008	0
	WLR-FH $(0.5, 0.5)$	0.823-0.834	0.031	0.008	0
5	Proposed $\tau = 0.75$	0	0.970	0.742	0.817
	Proposed $\tau = 0.5$	0	0.967	0.730	0.819
	Proposed $\tau = 0.25$	0	0.965	0.718	0.809
	WLR-PP	0	0.743	0.325	0.381
	WLR-FH $(0.5, 0.5)$	0	0.970	0.778	0.848
6	Proposed $\tau = 0.75$	0.764	0	0	0
	Proposed $\tau = 0.5$	0.767	0.001	0	0
	Proposed $\tau = 0.25$	0.761	0	0	0
	WLR-PP	0.753	0.001	0	0
	WLR-FH $(0.5, 0.5)$	0.784	0	0	0
7	Proposed $\tau = 0.75$	0	0.960	0.707	0.744
	Proposed $\tau = 0.5$	0	0.961	0.701	0.748
	Proposed $\tau = 0.25$	0	0.959	0.696	0.743
	WLR-PP	0	0.956	0.696	0.736
	WLR-FH $(0.5, 0.5)$	0	0.963	0.725	0.768
8	Proposed $\tau = 0.75$	0.883	0	0	0
	Proposed $\tau = 0.5$	0.876	0	0	0
	Proposed $\tau = 0.25$	0.879	0	0	0
	WLR-PP	0.777	0	0	0
	WLR-FH $(0.5, 0.5)$	0.854	0	0	0
9	Proposed $\tau = 0.75$	0	0.991	0.849	0.890
	Proposed $\tau = 0.5$	0	0.989	0.840	0.881
	Proposed $\tau = 0.25$	0	0.988	0.843	0.884
	WLR-PP	0	0.959	0.735	0.774
	WLR-FH $(0.5, 0.5)$	0	0.982	0.807	0.852

Table A.1: Rates of Stopping for Efficacy or for Safety

		AST			ASN			ANE		
Scenario	**Test Statistic**	**JT**	**P**	**OF**	**JT**	**P**	**OF**	**JT**	**P**	**OF**
1	Proposed $\tau = 0.75$	4.7	4.9	5.0	195	199	200	156	163	164
	Proposed $\tau = 0.5$	4.7	4.9	5.0	195	199	200	156	163	165
	Proposed $\tau = 0.25$	4.7	4.9	5.0	195	199	200	156	164	165
	WLR-PP	4.7	4.9	5.0	195	199	200	156	163	165
	WLR-FH $(0.5, 0.5)$	4.7	4.9	5.0	195	199	200	156	163	164
2	Proposed $\tau = 0.75$	3.8	3.8	3.8	186	186	186	144	144	144
	Proposed $\tau = 0.5$	3.8	3.8	3.8	185	185	185	144	144	144
	Proposed $\tau = 0.25$	3.8	3.8	3.8	186	186	186	145	145	145
	WLR-PP	3.8	3.8	3.8	186	186	186	145	145	145
	WLR-FH $(0.5, 0.5)$	3.7	3.7	3.7	184	184	184	142	142	142
3	Proposed $\tau = 0.75$	2.1	3.0	3.7	151	169	185	93	120	142
	Proposed $\tau = 0.5$	2.1	3.1	3.7	151	170	185	93	123	143
	Proposed $\tau = 0.25$	2.1	3.1	3.7	152	170	185	94	123	144
	WLR-PP	2.1	3.1	3.8	152	170	185	94	123	144
	WLR-FH $(0.5, 0.5)$	2.1	3.0	3.6	152	169	184	94	121	141
4	Proposed $\tau = 0.75$	3.9	3.9	4.0	189	190	190	135	137	138
	Proposed $\tau = 0.5$	3.9	3.9	4.0	189	190	190	136	137	138
	Proposed $\tau = 0.25$	3.9	4.0	4.0	190	191	191	137	139	140
	WLR-PP	4.5	4.6	4.7	195	196	197	152	154	155
	WLR-FH $(0.5, 0.5)$	3.7	3.8	3.8	186	188	188	132	134	135
5	Proposed $\tau = 0.75$	2.9	3.9	4.0	171	186	191	110	135	140
	Proposed $\tau = 0.5$	3.0	3.9	4.0	171	187	191	111	135	140
	Proposed $\tau = 0.25$	3.0	3.9	4.1	172	187	192	112	136	141
	WLR-PP	3.7	4.5	4.6	182	194	197	129	151	154
	WLR-FH $(0.5, 0.5)$	2.8	3.7	3.9	168	184	189	107	132	137
6	Proposed $\tau = 0.75$	3.7	3.7	3.7	184	184	184	143	143	143
	Proposed $\tau = 0.5$	3.7	3.7	3.7	184	184	184	144	144	144
	Proposed $\tau = 0.25$	3.7	3.7	3.7	185	185	185	145	145	145
	WLR-PP	3.7	3.7	3.7	184	184	184	145	145	145
	WLR-FH $(0.5, 0.5)$	3.6	3.6	3.6	183	183	183	142	142	142
7	Proposed $\tau = 0.75$	2.2	3.1	3.7	152	170	184	95	124	144
	Proposed $\tau = 0.5$	2.2	3.2	3.8	153	170	185	96	125	145
	Proposed $\tau = 0.25$	2.2	3.2	3.8	152	171	185	95	126	146
	WLR-PP	2.2	3.2	3.8	152	171	185	95	126	145
	WLR-FH $(0.5, 0.5)$	2.1	3.1	3.7	152	170	184	94	124	144
8	Proposed $\tau = 0.75$	3.5	3.5	3.5	181	181	181	129	129	129
	Proposed $\tau = 0.5$	3.5	3.5	3.5	181	181	181	129	129	129
	Proposed $\tau = 0.25$	3.5	3.5	3.5	182	182	182	130	130	130
	WLR-PP	3.7	3.7	3.7	184	184	184	133	133	133
	WLR-FH $(0.5, 0.5)$	3.5	3.5	3.5	181	181	181	129	129	129
9	Proposed $\tau = 0.75$	2.1	3.0	3.5	151	169	183	91	114	131
	Proposed $\tau = 0.5$	2.1	3.0	3.6	152	170	183	91	114	131
	Proposed $\tau = 0.25$	2.1	3.0	3.5	151	170	183	91	115	131
	WLR-PP	2.2	3.2	3.8	153	171	185	92	116	134
	WLR-FH $(0.5, 0.5)$	2.1	3.0	3.6	152	170	183	91	115	131

Table A.2: AST in Years, ASN and ANE in Scenarios 1 - 9

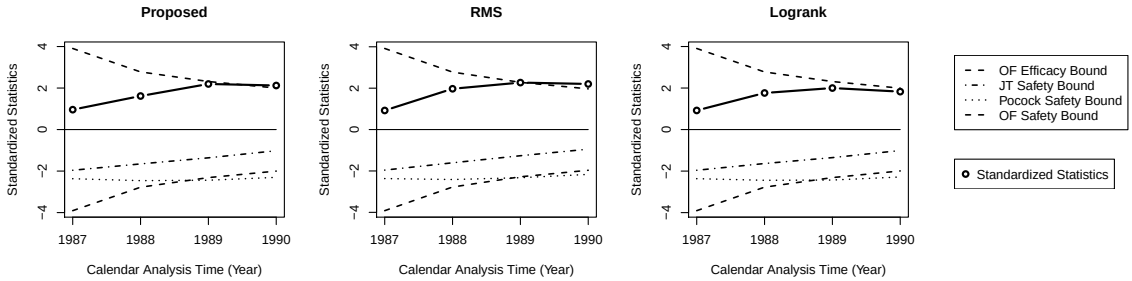


Figure A.1: Standardized Test Statistics and Stopping Boundaries (RMS: Restricted Mean Survival; OF:O'Brien and Fleming; JT: Jennison and Turnbull)

A.5 Supplementary Example Results of Section 2.6

Figure A.1 shows group sequential OF efficacy boundaries as well as OF, Pocock and JT safety boundaries for the proposed test statistic (left panel), the RMS statistic (middle panel) and the logrank statistic (right panel). All test statistics are standardized to ease comparisons between panels of the figure. Boundaries and test statistics shown in Figure A.1 are enumerated for clarity in Table A.3. Although historically during that period of clinical trial design symmetric stopping boundaries were typically used, a more modern safety boundary would make sense in this setting, particularly since it was not known for certain that the low-dose was sufficient to protect against mortality in the same way the high dose had up to that time. Observed values of the test statistics in each panel of Figure A.1 are superimposed as dots with bold connecting lines. None of the test statistics approached the safety boundaries at any of the interim analyses. As shown in Table A.3, the standardized proposed test statistics and the standardized RMS test statistics crossed the OF efficacy boundary at year 1990. The logrank test did not cross the OF efficacy boundary at any interim analysis time.

	Proposed				**RMS**				**Logrank**			
	1987	1988	1989	1990	1987	1988	1989	1990	1987	1988	1989	1990
Test Statistics	0.96	1.62	2.20	2.12	0.92	1.97	2.27	2.20	0.92	1.76	2.00	1.83
OF Efficacy	3.91	2.78	2.31	2.00	3.92	2.78	2.28	1.97	3.91	2.77	2.31	1.99
JT Safety	-1.96	-1.65	-1.36	-1.02	-1.96	-1.61	-1.26	-0.95	-1.96	-1.64	-1.35	-1.01
Pocock Safety	-2.37	-2.46	-2.44	-2.30	-2.37	-2.41	-2.33	-2.17	-2.37	-2.44	-2.43	-2.28
OF Safety	-3.91	-2.78	-2.31	-2.00	-3.92	-2.78	-2.28	-1.97	-3.91	-2.77	-2.31	-1.99

Table A.3: Test Statistics and Efficacy or Safety Boundaries

APPENDIX B

Supplementary Materials for Chapter III

B.1 Asymptotic Multivariate Distribution of $\tilde{\mathscr{T}}_k$

In this section, we prove that the multivariate distribution of

$$\tilde{\mathscr{T}}_k = \left\{ \tilde{\mathscr{T}}_1(s_1), \ldots, \tilde{\mathscr{T}}_k(s_k) \right\}$$

is a mean zero Normal distribution with covariance matrix Σ_k for $k = 1, \ldots, K$ as described in Chapter 3.4.

We start from our unstandardized test statistic at analysis time s:

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

which can be written as

$$\mathscr{T}(s) = \sqrt{\frac{n_2(s)}{n_1(s) + n_2(s)}} \sqrt{n_1(s)} \hat{\mu}_1(s,\tau) - \sqrt{\frac{n_1(s)}{n_1(s) + n_2(s)}} \sqrt{n_2(s)} \hat{\mu}_2(s,\tau), \quad \text{(B.1)}$$

where $n_g(s)/\{n_1(s) + n_2(s)\} \xrightarrow{p} \pi_g(s)$. Suppose at analysis time s, combining information of the time to first event captured in all b follow-up windows of length τ, we

record M unique event times $\{0 \equiv T_0 < T_1 < ... < T_M < T_{M+1} \equiv \tau\}$. Then, by Taylor series expansion,

$$\sqrt{n_g(s)}\hat{\mu}_g(s,\tau) = \sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)exp\left\{-\sum_{j=0}^{m}\frac{dN_g(s,T_j)}{Y_g(s,T_j)}\right\}$$

is asymptotically equivalent in distribution to:

$$\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\,exp\left\{-\sum_{j=0}^{m}\lambda_g^W(s,T_j)dT_j\right\} \tag{B.2}$$

$$+\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\left[\sum_{j=0}^{m}-exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\frac{dN_g(s,T_j)}{Y_g(s,T_j)}\right] \tag{B.3}$$

$$-\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\left[\sum_{j=0}^{m}-exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\lambda_g^W(s,T_j)dT_j\right] \tag{B.4}$$

$$\begin{aligned}&+\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\frac{1}{2!}exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\\&\times\left[\sum_{j=0}^{m}\left\{\frac{dN_g(s,T_j)}{Y_g(s,T_j)}-\lambda_g^W(s,T_j)dT_j\right\}\right]^2\end{aligned} \tag{B.5}$$

$$+\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)[\text{higher order terms}] \tag{B.6}$$

Terms (B.5) and (B.6) converge to zero in probability using similar arguments to those shown in *Tayob and Murray* (2014) Appendix A. When terms (B.2) and (B.4) are combined into the test statistic, $\mathscr{T}(s)$, under the null hypothesis, they cancel with terms from the other treatment group. Hence, the asymptotic behavior of $\mathscr{T}(s)$ is based on term (B.3) for groups $g = 1, 2$, which can be further rewritten as

$$\sqrt{n_g(s)}\sum_{m=0}^{M}(T_{m+1}-T_m)\left[\sum_{j=0}^{m}-exp\left\{-\sum_{j'=0}^{m}\lambda_g^W(s,T_{j'})dT_{j'}\right\}\frac{dN_g(s,T_j)}{Y_g(s,T_j)}\right]$$

$$= -\sqrt{n_g(s)} \sum_{m=0}^{M} (T_{m+1} - T_m) exp\{-\sum_{j'=0}^{m} \lambda_g^W(s, T_{j'}) dT_{j'}\} \sum_{j=0}^{m} \frac{dN_g(s, T_j)}{Y_g(s, T_j)},$$

By Taylor series expansion, this term is asymptotically equivalent in distribution to

$$-\sqrt{n_g(s)} \sum_{m=0}^{M} (T_{m+1} - T_m) exp\left\{-\sum_{j'=0}^{m} \lambda_g^W(s, T_{j'}) dT_{j'}\right\} \times$$

$$\left\{\sum_{j=0}^{m} \frac{E dN_g(s, T_j)}{E Y_g(s, T_j)} \right. \tag{B.7}$$

$$+\sum_{j=0}^{m} \left[\frac{1}{E Y_g(s, T_j)} [dN_g(s, T_j) - E dN_g(s, T_j)] - \frac{E dN_g(s, T_j)}{E Y_g(s, T_j)^2} [Y_g(s, T_j) - E Y_g(s, T_j)]\right] \tag{B.8}$$

$$\left. +[\text{higher order terms}]\right\}. \tag{B.9}$$

Using arguments similar to those given in *Tayob and Murray* (2014), the higher order terms in (B.9) converge to zero in probability. When term (B.7) appears in $\mathscr{T}(s)$, it cancels with its corresponding term from the other treatment group under the null hypothesis. Hence, the asymptotic behavior of $\mathscr{T}(s)$ is based on term (B.8) which upon noting that $E dN_g(s, T_j)/E Y_g(s, T_j) = \lambda_g^W(s, T_j)$ and $E Y_g(s, T_j) = \sum_{l=1}^{b} Pr(X_{gi}(s, t_l) \geq T_j)$ can be algebraically rearranged as:

$$-\sqrt{n_g(s)} \sum_{m=0}^{M} (T_{m+1} - T_m) exp\left\{-\sum_{j'=0}^{m} \lambda_g^W(s, T_{j'}) dT_{j'}\right\} \sum_{j=0}^{m} \frac{dN_g(s, T_j) - Y_g(s, T_j) \lambda_g^W(s, T_j)}{\sum_{l=1}^{b} Pr(X_{gi}(s, t_l) \geq T_j)}$$

or in more standard stochastic integral notation as:

$$-\sqrt{n_g(s)} \int_0^{\tau} exp\left\{-\int_0^{u_2} \lambda_g^W(s, u_1) du_1\right\} \int_0^{u_2} \frac{dN_g(s, u_1) - Y_g(s, u_1) \lambda_g^W(s, u_1)}{\sum_{l=1}^{b} Pr(X_{gi}(s, t_l) \geq u_1)} du_2. \tag{B.10}$$

Summarizing the above remarks,

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

is asymptotically equivalent in distribution to

$$\sqrt{\pi_1(s)}\sqrt{n_2(s)}\int_0^\tau exp\left\{-\int_0^{u_2}\lambda_2^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_2(s,u_1)-Y_2(s,u_1)\lambda_2^W(s,u_1)}{\sum_{l=1}^b Pr(X_{2i}(s,t_l)\geq u_1)}du_2$$

$$-\sqrt{\pi_2(s)}\sqrt{n_1(s)}\int_0^\tau exp\left\{-\int_0^{u_2}\lambda_1^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_1(s,u_1)-Y_1(s,u_1)\lambda_1^W(s,u_1)}{\sum_{l=1}^b Pr(X_{1i}(s,t_l)\geq u_1)}du_2. \tag{B.11}$$

Recall that

$$N_g(s,u) = \sum_{i=1}^{n_g(s)} N_{gi}(s,u) = \sum_{i=1}^{n_g(s)}\sum_{j=1}^{b} N_{gi}(s,t_j,u)$$

and

$$Y_g(s,u) = \sum_{i=1}^{n_g(s)} Y_{gi}(s,u) = \sum_{i=1}^{n_g(s)}\sum_{j=1}^{b} Y_{gi}(s,t_j,u).$$

We define:

$$Z_{ij}\{\hat{\mu}_g(s,\tau)\} =$$

$$\int_0^\tau exp\left\{-\int_0^{u_2}\lambda_g^W(s,u_1)du_1\right\}\int_0^{u_2}\frac{dN_{gi}(s,t_j,u_1)-Y_{gi}(s,t_j,u_1)\lambda_g^W(s,u_1)du_1}{\sum_{l=1}^b Pr\{X_{gi}(s,t_l)\geq u_1\}}du_2$$

and

$$Z_i\{\hat{\mu}_g(s,\tau)\} = \sum_{j=1}^{b} Z_{ij}\{\hat{\mu}_g(s,\tau)\}.$$

We can write equation (B.11) as

$$\mathscr{T}^*(s) = \sqrt{\pi_1(s)}\sqrt{n_2(s)}\frac{\sum_{i=1}^{n_2(s)} Z_i\{\hat{\mu}_2(s,\tau)\}}{n_2(s)} - \sqrt{\pi_2(s)}\sqrt{n_1(s)}\frac{\sum_{i=1}^{n_1(s)} Z_i\{\hat{\mu}_1(s,\tau)\}}{n_1(s)}. \tag{B.12}$$

Note that $Z_i\{\hat{\mu}_g(s,\tau)\}$ only depends on patient i and is independent and iden-

tically distributed for $i = 1, \ldots, n_g(s)$. As a result, the multivariate central limit theorem can be used to determine the asymptotic joint distribution of $\{\mathscr{T}^*(s_1), \ldots, \mathscr{T}^*(s_k)\}$, $k = 1, \ldots, K$, when each statistic is formulated as in equation (B.12). As a result, the covariance matrix of $\{\mathscr{T}^*(s_1), \ldots, \mathscr{T}^*(s_k)\}$ with component $Cov\{\mathscr{T}^*(s_{k_1}), \mathscr{T}^*(s_{k_2})\}$, can be estimated using empirical covariances of $Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}$ and $Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}$, for $g = 1, 2$, where appropriate, as follows.

First, without loss of generality, assume $s_{k_1} \leq s_{k_2}$ so that $n_g(s_{k_1}) \leq n_g(s_{k_2})$ with $n_g(s_{k_1})$ patients contributing (correlated) data from both analysis times. Then

$$
\begin{aligned}
&Cov\{\mathscr{T}^*(s_{k_1}), \mathscr{T}^*(s_{k_2})\} \\
&= \sum_{g=1}^{2} Cov\left[\sqrt{\pi_{3-g}(s_{k_1})}\sqrt{n_g(s_{k_1})}\frac{\sum_{i=1}^{n_g(s_{k_1})} Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}}{n_g(s_{k_1})},\right. \\
&\qquad\qquad \left.\sqrt{\pi_{3-g}(s_{k_2})}\sqrt{n_g(s_{k_2})}\frac{\sum_{i=1}^{n_g(s_{k_2})} Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}}{n_g(s_{k_2})}\right] \\
&= \sum_{g=1}^{2} \sqrt{\pi_{3-g}(s_{k_1})}\sqrt{\pi_{3-g}(s_{k_2})}\frac{n_g(s_{k_1})}{\sqrt{n_g(s_{k_1})n_g(s_{k_2})}} Cov\left[Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}, Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}\right],
\end{aligned}
$$

which is asymptotically equivalent to

$$
= \sum_{g=1}^{2} \sqrt{\pi_{3-g}(s_{k_1})\pi_{3-g}(s_{k_2})\psi_g(s_{k_1}, s_{k_2})} Cov\left[Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}, Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}\right],
$$

where for group $g = 1, 2$, $\psi_g(s_{k_1}, s_{k_2})$ is the limiting proportion of patients entered at s_{k_1} of those eventually entered by s_{k_2}, that is estimated by $n_g(s_{k_1})/n_g(s_{k_2})$. Therefore, we can estimate $Cov\left[Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}, Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}\right]$ with the empirical covariance of sample realizations of $Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}$ and $Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}$, that is,

$$
\widehat{Cov}\left[Z_i\{\hat{\mu}_g(s_{k_1}, \tau)\}, Z_i\{\hat{\mu}_g(s_{k_2}, \tau)\}\right] =
$$

$$
\sum_{i=1}^{n_g(s_{k_1})} \frac{[z_i\{\hat{\mu}_g(s_{k_1}, \tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_1}, \tau)\}][z_i\{\hat{\mu}_g(s_{k_2}, \tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2}, \tau)\}]}{n_g(s_{k_1}) - 1}.
$$

where $z_i\{\hat{\mu}_g(s,\tau)\}$ and $\bar{z}\{\hat{\mu}_g(s,\tau)\}$ are defined in terms of $z_{ij}\{\hat{\mu}_g(s,\tau)\}$ in Chapter 3.3. However, this estimation can be improved upon by updating $z_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$ with quantities that that do not depend on analysis time and thus can be estimated better using the full data at the later analysis time s_{k_2}. In particular, since both $dN_{gi}(s_{k_1},t_j,u_1)/Y_{gi}(s_{k_1},t_j,u_1)$ and $dN_{gi}(s_{k_2},t_j,u_1)/Y_{gi}(s_{k_2},t_j,u_1)$ estimate $\lambda_{gi}(t_j,u_1)du_1$, and the latter term uses more data, we replace $dN_{gi}(s_{k_1},t_j,u_1)$ with

$$Y_{gi}(s_{k_1},t_j,u_1)\frac{dN_{gi}(s_{k_2},t_j,u_1)}{Y_{gi}(s_{k_2},t_j,u_1)}.$$

Similarly, we replace $Y_g(s_{k_1},u_1)/n_g(s_{k_1})$, which is an estimate of $\sum_{l=1}^{b} Pr\{T_{gi}(s_{k_1},t_l)\geq u_1\}Pr\{C_{gi}(s_{k_1},t_l)\geq u_1\}$, with

$$\left[\sum_{i=1}^{n_g(s_{k_2})}\frac{I\{T_{gi}\geq u_1+t_l\}}{n_g(s_{k_2})}\right]\left[\sum_{i=1}^{n_g(s_{k_1})}\frac{I\{C_{gi}(s_{k_1})\geq u_1+t_l\}}{n_g(s_{k_1})}\right].$$

Here, terms involving the event time are estimated using updated data, while terms involving the censoring distribution remain relevant to analysis time s_{k_1}. Putting these modifications together gives us

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

as an updated version of $z_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$ for use in covariance terms. So that we replace the $z_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$ terms in $z_i\{\hat{\mu}_g(s_{k_1},\tau)\}$ and $\bar{z}\{\hat{\mu}_g(s_{k_1},\tau)\}$ with $\tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$ to

obtain $\tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\}$ and $\bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1},\tau)\}$. And we update the empirical covariance as

$$\widehat{Cov}\left[Z_i\{\hat{\mu}_g(s_{k_1},\tau)\}, Z_i\{\hat{\mu}_g(s_{k_2},\tau)\}\right] =$$

$$\sum_{i=1}^{n_g(s_{k_1})} \frac{[\tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\} - \bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1},\tau)\}][z_i\{\hat{\mu}_g(s_{k_2},\tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2},\tau)\}]}{n_g(s_{k_1}) - 1}.$$

For the standardized version of test statistic, $\tilde{\mathscr{T}}(s_k), s_k = s_1, \ldots, s_K$, we work with the corresponding standardized form of the more tractible random variable that is asymptotically equivalent in distribution, namely,

$$\frac{\mathscr{T}^*(s_k)}{\sqrt{\pi_2(s_k)\sigma_1^2(s_k) + \pi_1(s_k)\sigma_2^2(s_k)}}.$$

which also gives $\tilde{\mathscr{T}}_k$ an asymptotic mean zero multivariate Normal distribution with covariance matrix Σ_k. Because the test statistic is standardized to have variance 1.0. We only need to estimate the off-diagonal elements $\sigma_{k_1k_2}$ via

$$\hat{\sigma}_{k_1k_2} = \frac{\widehat{Cov}\left\{\mathscr{T}^*(s_{k_1}), \mathscr{T}^*(s_{k_2})\right\}}{\sqrt{\hat{\pi}_2(s_{k_1})\tilde{\sigma}_1^2(s_{k_1}) + \hat{\pi}_1(s_{k_1})\hat{\sigma}_2^2(s_{k_1})}\sqrt{\hat{\pi}_2(s_{k_2})\tilde{\sigma}_1^2(s_{k_2}) + \hat{\pi}_1(s_{k_2})\hat{\sigma}_2^2(s_{k_2})}}. \quad \text{(B.13)}$$

Chapter 3.3 gives estimates $\hat{\pi}_g(s_k)$ for $s_k = s_{k_1}, s_{k_2}$ and $\hat{\sigma}_g^2(s_{k_2})$ using the most up-to-date information. Estimate $\tilde{\sigma}_g^2(s_{k_1})$ in equation (B.13) for $g = 1, 2$ is modified by replacing $z_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$ with $\tilde{z}_{ij}\{\hat{\mu}_g(s_{k_1},\tau)\}$. Therefore, we have

$$\begin{aligned}\hat{\sigma}_{k_1k_2} =& \{\hat{\pi}_2(s_{k_1})\tilde{\sigma}_1^2(s_{k_1}) + \hat{\pi}_1(s_{k_1})\tilde{\sigma}_2^2(s_{k_1})\}^{-\frac{1}{2}}\{\hat{\pi}_2(s_{k_2})\hat{\sigma}_1^2(s_{k_2}) + \hat{\pi}_1(s_{k_2})\hat{\sigma}_2^2(s_{k_2})\}^{-\frac{1}{2}} \\ &\times \sum_{g=1}^{2}\sqrt{\hat{\pi}_{3-g}(s_{k_1})\hat{\pi}_{3-g}(s_{k_2})\hat{\psi}_g(s_{k_1},s_{k_2})}\Bigg(\sum_{i=1}^{n_g(s_{k_1})}\{n_g(s_{k_1})-1\}^{-1} \\ &\times [\tilde{z}_i\{\hat{\mu}_g(s_{k_1},\tau)\} - \bar{\tilde{z}}\{\hat{\mu}_g(s_{k_1},\tau)\}]\,[z_i\{\hat{\mu}_g(s_{k_2},\tau)\} - \bar{z}\{\hat{\mu}_g(s_{k_2},\tau)\}]\Bigg).\end{aligned} \quad \text{(B.14)}$$

B.2 Simulated Cumulative Power in the Special Case with Independent Recurrent and Terminal Event Distributions

Figure B.1 shows simulated power for the group sequentially monitored CL, TM and LR statistics when all events within each individual are statistically independent, but otherwise have marginal distributions as given in Chapter 3.5. The CL statistic (triangles) had the highest in this special case, followed closely by the TM statistic (circles) and distantly by the LR method(+).

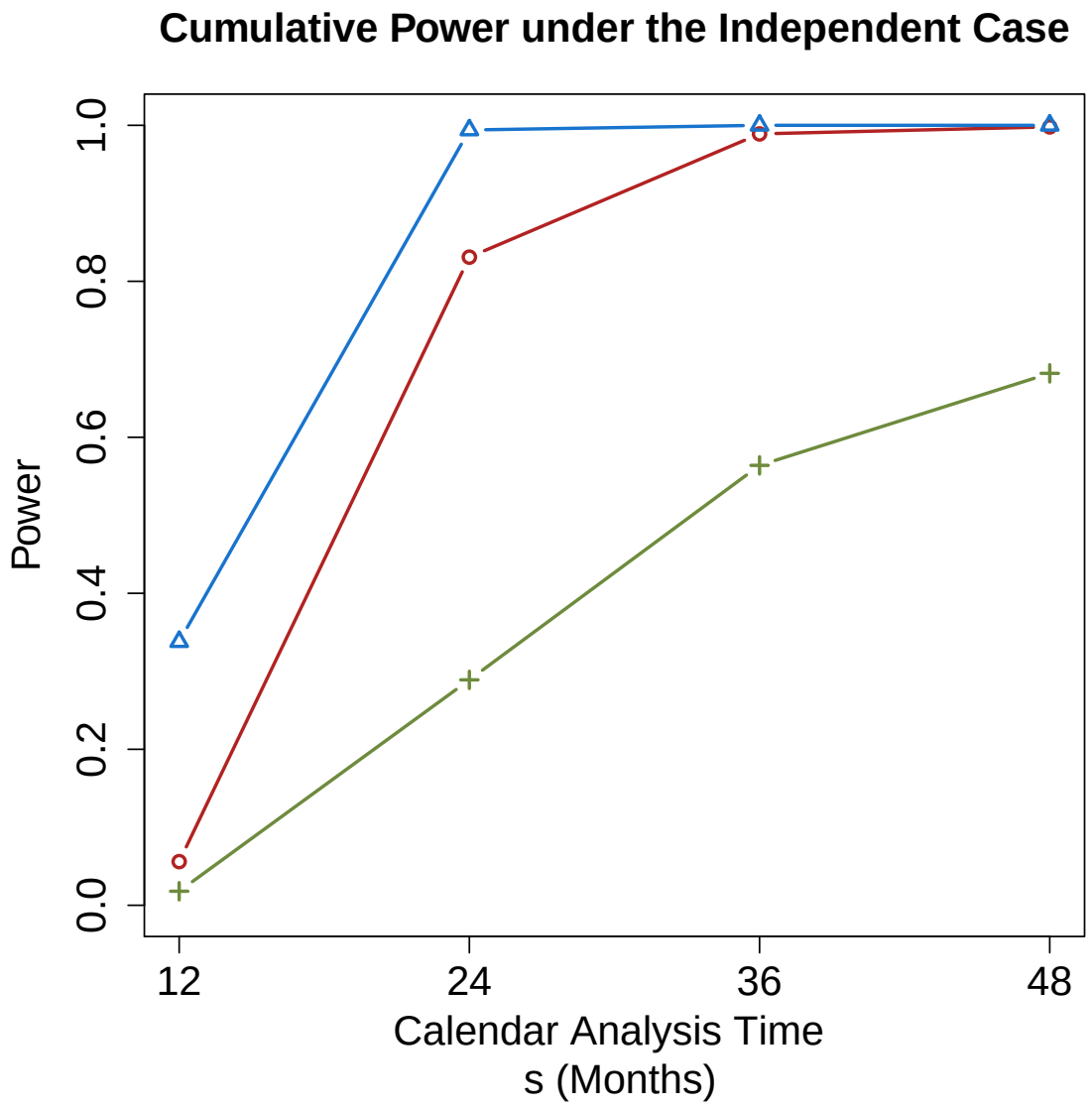


Figure B.1: Simulated Cumulative Power in the Special Case with Independent Recurrent and Terminal Event Distributions

APPENDIX C

Supplementary Materials for Chapter IV

Derivations are conducted for a typical individual in the study with submerged subscripts for this individual. Let gap times, G_j, between the $j-1^{st}$ and the j^{th} recurrent events independently follow an exponential distribution with rate λ. Then the total number of observed recurrent events K, occurring during s follow-up units, follows a Poisson(λs) distribution. Define $R_j = \sum_{i=1}^{j} G_i$ as the time from the initiation of follow-up to the j^{th} recurrent event; R_j follows a Gamma(j, λ) distribution with shape j and rate λ. Recall that the Tayob and Murray statistic only includes the first event in each τ-length follow-up window in the analysis, with follow-up windows $k = 1, \ldots, b$ starting at follow-up times $t_k = (k-1)a \leq s - \tau$. A recurrent event may be included in more than follow-up window, particularly if a is small relative to the mean time between events, $\frac{1}{\lambda}$.

Conditional on K, it will be convenient to define a random variable, W_j, that denotes the last follow-up window to overlap the j^{th} observed recurrent event, $j = 1, \ldots, K$. Inclusion of this j^{th} recurrent event in the analysis (at least once), is satisfied if it is the first such event to occur in window W_j, or equivalently $W_j > W_{j-1}$. Otherwise, if $W_j = W_{j-1}$, the $j-1^{st}$ event precedes the j^{th} event in any follow-up windows these two events have in common and the j^{th} recurrent event is left out of

the analysis. By definition, since the follow-up windows are spaced a units apart, $W_j = \text{ceiling}\left[\frac{R_j}{a}\right]$, so that probability calculations for W_j can be framed in terms of functions of well understood random variables.

Let M_j be an indicator variable denoting that an individual's j^{th} observed event is left out of the analysis, with $Pr\{M_j = 1\} = Pr\{W_j = W_{j-1}\}$ for $j \geq 2$. Then conditional on the K observed events for the individual, $M = \sum_{j=2}^{K} M_j$ is the number of observed events left out of the analysis, with $M = 0$ when $K \leq 1$. Hence an expression for p, the average proportion of recurrent events captured in at least one follow-up window for an individual followed s time units, can be framed as

$$p = E\left(\frac{K-M}{K}\right) = E\left[E\left(\frac{K-M}{K}\middle| K\right)\right] = 1 - E\left[\frac{E(M|K)}{K}\right].$$

For $K \leq 1$, $E(M|K) = 0$. For $K \geq 2$,

$$E(M|K) = \sum_{j=2}^{K} E\left(M_j|K\right) = \sum_{j=2}^{K}\sum_{w=1}^{b} Pr\{W_j = W_{j-1} = w|K\}.$$

Let $\text{pmf}_{\text{Poisson}(\lambda s)}(k)$ be the Poisson(λs) probability mass function (pmf) for $k = 0, 1, \cdots, \infty$. Since K has a Poisson(λs) distribution, the previous expression for p becomes

$$p = 1 - \sum_{k=0}^{\infty}\left[\frac{\sum_{j=2}^{k}\sum_{w=1}^{b} Pr\{W_j = W_{j-1} = w|K = k\}}{k} I\{k \geq 2\}\right] \text{pmf}_{\text{Poisson}(\lambda s)}(k)$$

The remainder of the appendix details calculations of $Pr\{W_j = W_{j-1} = w|K = k\}$ for $k \geq 2$ in the expression for p. That is,

$$Pr\{W_j = W_{j-1} = w|K = k\} = Pr\{(w-1)a < R_{j-1} < R_j \leq \min(aw, s)|R_k \leq s, R_{k+1} > s\}$$

$$= \frac{Pr\left\{(w-1)a < R_{j-1} < R_j \leq \min(aw, s), R_k \leq s, R_{k+1} > s\right\}}{Pr\{K = k\}}$$

Random variables in the numerator can be rewritten in terms of components that are independent, since in general $R_k = R_{j-1} + \sum_{\ell=j}^{k} G_\ell$ for any $j \leq k$. For simplicity let $R^{(j,k)} = \sum_{\ell=j}^{k} G_\ell$, for $j \leq k$, which has a Gamma$(k-j+1, \lambda)$ distribution. Then the numerator becomes

$$Pr\{R_{j-1} > (w-1)a, R_{j-1} + G_j \leq \min(aw, s), R_{j-1} + G_j + R^{(j+1,k)} \leq s, R_{j-1} + G_j + R^{(j+1,k)} + G_{k+1} > s\}$$

$$= Pr\{R_{j-1} > (w-1)a, R_{j-1} + G_j \leq \min(aw, s), R_{j-1} + G_j + R^{(j+1,k)} \leq s\} \quad \text{(C.1)}$$

$$- Pr\{R_{j-1} > (w-1)a, R_{j-1} + G_j \leq \min(aw, s), R_{j-1} + G_j + R^{(j+1,k)} + G_{k+1} \leq s\} \quad \text{(C.2)}$$

Equation (C.1) becomes

$$= Pr\{R_{j-1} + G_j \leq \min(aw, s), R_{j-1} + G_j + R^{(j+1,k)} \leq s\}$$
$$- Pr\{R_{j-1} \leq (w-1)a, R_{j-1} + G_j \leq \min(aw, s), R_{j-1} + G_j + R^{(j+1,k)} \leq s\}$$
$$= Pr\{R_j \leq \min(aw, s), R_j + R^{(j+1,k)} \leq s\}$$
$$- Pr\{R_{j-1} \leq (w-1)a, R_{j-1} + G_j \leq \min(aw, s), R_{j-1} + G_j + R^{(j+1,k)} \leq s\}$$
$$= \int_0^{\min(aw,s)} \text{pdf}_{\text{Gamma}(j,\lambda)}(r) \cdot \text{cdf}_{\text{Gamma}(k-j,\lambda)}(s-r)dr$$
$$- \int_0^{(w-1)a} \int_0^{\min(aw,s)-r} \text{pdf}_{\text{Gamma}(j-1,\lambda)}(r) \cdot \text{pdf}_{\text{Exp}(\lambda)}(g) \cdot \text{cdf}_{\text{Gamma}(k-j,\lambda)}(s-r-g)dgdr$$

Similarly equation (C.2) becomes

$$= \int_0^{\min(aw,s)} \text{pdf}_{\text{Gamma}(j,\lambda)}(r) \cdot \text{cdf}_{\text{Gamma}(k-j+1,\lambda)}(s-r)dr$$
$$- \int_0^{(w-1)a} \int_0^{\min(aw,s)-r} \text{pdf}_{\text{Gamma}(j-1,\lambda)}(r) \cdot \text{pdf}_{\text{Exp}(\lambda)}(g) \cdot \text{cdf}_{\text{Gamma}(k-j+1,\lambda)}(s-r-g)dgdr$$

So that combining (C.1) and (C.2) we finally have

$$p = 1 - \sum_{k=2}^{\infty} \frac{1}{k} \sum_{j=2}^{k} \sum_{w=1}^{b}$$
$$\Bigg[\int_0^{\min(aw,s)} \text{pdf}_{\text{Gamma}(j,\lambda)}(r) \Big\{ \text{cdf}_{\text{Gamma}(k-j,\lambda)}(s-r) - \text{cdf}_{\text{Gamma}(k-j+1,\lambda)}(s-r) \Big\} dr$$
$$- \int_0^{(w-1)a} \int_0^{\min(aw,s)-r} \text{pdf}_{\text{Gamma}(j-1,\lambda)}(r) \cdot \text{pdf}_{\text{Exp}(\lambda)}(g)$$
$$\Big\{ \text{cdf}_{\text{Gamma}(k-j,\lambda)}(s-r-g) - \text{cdf}_{\text{Gamma}(k-j+1,\lambda)}(s-r-g) \Big\} dgdr \Bigg],$$

as given in Chapter IV.

APPENDIX D

Supplementary Materials for Chapter V

D.1 The Derivation of the Marginal Distribution of $T_i(t)$ With Independent Recurrent Event Times

In this appendix section we describe that for an individual i, when $G_{ij}, j = 1, \ldots, J_i$ are independently and identically distributed as exponential with intensity λ_i the marginal distribution of $T_i(t)$ for a fixed t is also distributed as exponential with hazard λ_i. This result is trivial for the case where $t = 0$, since

$T_i(t=0) \equiv G_{i1} \sim Exp(\lambda_i)$. For the case when $t > 0$,

$$
\begin{aligned}
&Pr\{T_i(t) > u\} \\
=&Pr\{T_{i1} > t+u\} + Pr\{T_{i1} \leq t, T_{i2} > t+u\} + \cdots + Pr\{T_{iJ_i-1} \leq t, T_{iJ_i} > t+u\} \\
=&Pr\{T_{i1} > t+u\} + \sum_{j=2}^{J_i} Pr\{T_{ij-1} \leq t, T_{ij} > t+u\} \\
=&Pr\{G_{i1} > t+u\} + \sum_{j=2}^{J_i} Pr\{T_{ij-1} \leq t, T_{ij-1} + G_{ij} > t+u\} \\
=&\int_{t+u}^{\infty} \lambda_i e^{-\lambda_i y} dy + \sum_{j=2}^{J_i} \int_0^t \int_{t+u-p}^{\infty} f(T_{ij-1} = p, G_{ij} = q) dq dp
\end{aligned}
\tag{D.1}
$$

where $T_{ij-1} \perp G_{ij}$ by assumption. We know that $T_{ij-1} = \sum_{k=1}^{k=j-1} G_{ik}$. When $G_{ik} \overset{iid}{\sim} Exp(\lambda_i)$, $T_{ij-1} \sim Gamma(j-1, \lambda_i)$, where $j-1$ is the shape parameter and λ_i is the rate parameter. Then,

$$
\begin{aligned}
&Pr\{T_i(t) > u\} \\
=&e^{-\lambda_i(t+u)} + \sum_{j=2}^{J_i} \int_0^t \int_{t+u-p}^{\infty} \text{pdf}_{Gamma(j-1,\lambda_i)}(T_{ij-1} = p) \text{pdf}_{Exp(\lambda_i)}(G_{ij} = q) dp dq
\end{aligned}
$$

where

$$
\text{pdf}_{Gamma(j-1,\lambda_i)}(T_{ij-1} = p) = \frac{\lambda_i^{j-1}}{\Gamma(j-1)} p^{j-2} e^{-\lambda_i p},
$$

$$
\text{pdf}_{Exp(\lambda_i)}(G_{ij} = q) = \lambda_i e^{-\lambda_i q}.
$$

So,

$$
\begin{aligned}
& Pr\{T_i(t) > u\} \\
=& e^{-\lambda_i(t+u)} + \sum_{j=2}^{J_i} \int_0^t \frac{\lambda_i^{j-1}}{\Gamma(j-1)} p^{j-2} e^{-\lambda_i p} \int_{t+u-p}^{\infty} \lambda_i e^{-\lambda_i q} dp dq \\
=& e^{-\lambda_i(t+u)} + \sum_{j=2}^{J_i} \int_0^t \frac{\lambda_i^{j-1}}{\Gamma(j-1)} p^{j-2} e^{-\lambda_i(t+u)} dp \\
=& e^{-\lambda_i(t+u)} + \sum_{j=2}^{J_i} \frac{(\lambda_i t)^{j-1}}{\Gamma(j)} e^{-\lambda_i(t+u)} \\
=& \sum_{j=1}^{J_i} \frac{(\lambda_i t)^{j-1}}{\Gamma(j)} e^{-\lambda_i t} \times e^{-\lambda_i u}
\end{aligned}
$$

When $J_i \to \infty$, by Taylor series,

$$\sum_{j-1=0}^{\infty} \frac{(\lambda_i t)^{j-1}}{(j-1)!} = e^{\lambda_i t}.$$

Therefore, $Pr\{T_i(t) > u\} = e^{-\lambda_i u}$, namely, $T_i(t) \sim Exp(\lambda_i)$.

D.2 Example Showing the Imputation of Event Times

Figure D.1 shows the same example participant from the Azithromycin in COPD Trial shown in Figure 5.1. For this example participant, $\mathcal{S}_i = \{180, 240, 300\}$ so that the sup window starts at $t^{sup}(\mathcal{S}_i) = 300$. The sup impute becomes $\tilde{T}_i\{t^{sup}(\mathcal{S}_i)\} = \tilde{T}_i\{300\} = 65$. For the window starting at 240 days, the imputed time-to-first-event becomes $\tilde{T}_i\{240\} = \tilde{T}_i\{300\} + 300 - 240 = 125$, which is greater than the censored time-to-first-event that was observed for this window, $X_i(240) = 113$. Similarly for the window starting at 180 days, $\tilde{T}_i\{180\} = \tilde{T}_i\{300\} + 300 - 180 = 185$, which is greater than the censored time-to-first-event that was observed for this window, $X_i(180) = 173$.

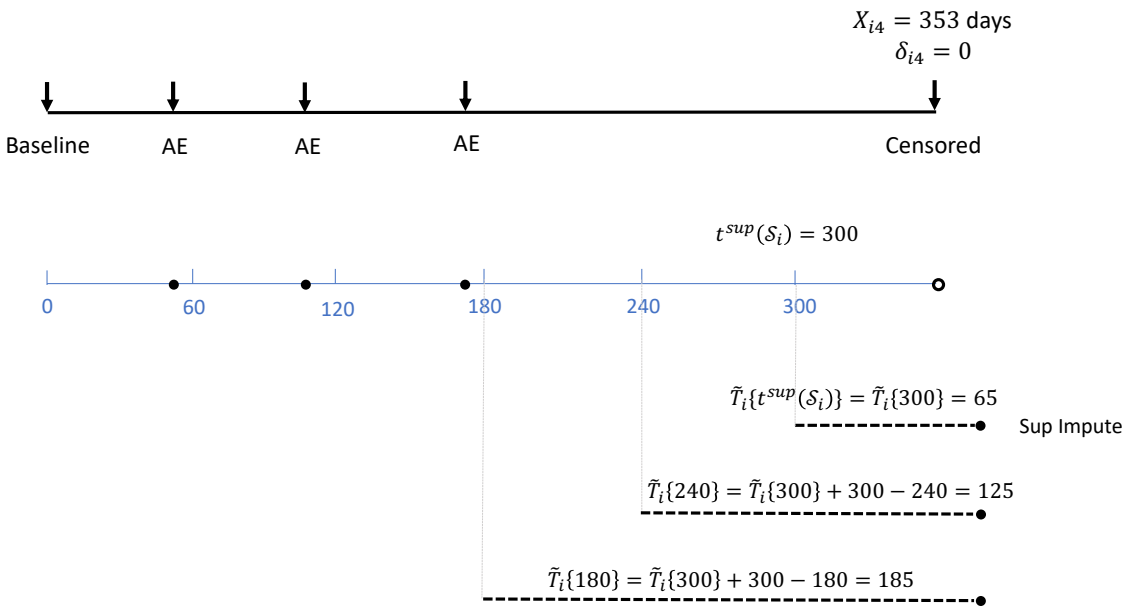


Figure D.1: Example Showing the Imputation of Event Times

D.3 Supplementary Table Corresponding to Figure 5.5

	PO				MI			
	$e^{\hat{\beta}}$	95% CI		P	$e^{\hat{\beta}}$	95% CI		P
Overall:	1.143	1.052	1.242	0.002	1.142	1.054	1.238	0.001
Sex:								
Male	1.088	0.983	1.204	0.104	1.087	0.984	1.201	0.100
Female	1.234	1.074	1.417	0.003	1.230	1.076	1.406	0.002
Smoking Status:								
Former	1.176	1.069	1.294	0.001	1.175	1.071	1.289	0.001
Current	1.032	0.875	1.218	0.707	1.040	0.884	1.223	0.639
Age:								
$\leq$ 65 years	1.067	0.949	1.198	0.278	1.069	0.955	1.198	0.246
$>$ 65 years	1.233	1.096	1.387	<0.001	1.226	1.093	1.374	<0.001
FEV1:								
$\leq$ 50 % predicted	1.095	0.994	1.208	0.067	1.094	0.995	1.202	0.063
$>$ 50 % predicted	1.291	1.104	1.510	0.001	1.289	1.107	1.502	0.001

Table D.1: Subset Analyses Comparing Azithromycin versus Placebo Using Proposed PO and MI Approaches with a GEE Model Fit of Equation (5.3)
(CI: Confidence Interval; PO: Pseudo-observation; MI: Multiple Imputation.)

BIBLIOGRAPHY

BIBLIOGRAPHY

Aalen, O. O., and E. Husebye (1991), Statistical analysis of repeated events forming renewal processes, *Statistics in medicine*, *10*(8), 1227–1240.

Albert, R. K., et al. (2011), Azithromycin for prevention of exacerbations of copd, *New England Journal of Medicine*, *365*(8), 689–698.

Andersen, P. K., and R. D. Gill (1982), Cox's regression model for counting processes: a large sample study, *Ann Stat*.

Andersen, P. K., and J. P. Klein (2007), Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies, *Scandinavian Journal of Statistics*, *34*(1), 3–16.

Andersen, P. K., M. G. Hansen, and J. P. Klein (2004), Regression analysis of restricted mean survival time based on pseudo-observations, *Lifetime data analysis*, *10*(4), 335–350.

Andrei, A.-C., and S. Murray (2007), Regression models for the mean of the quality-of-life-adjusted restricted survival time using pseudo-observations, *Biometrics*, *63*(2), 398–404.

Ashley, S. L., M. Xia, S. Murray, D. N. O'Dwyer, and E. Grant (2016), Six-somamer index relating to immune, protease and angiogenic functions predicts progression in ipf, *PloS One*.

Breslow, N. (1970), A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship, *Biometrika*, *57*(3), 579–594.

Calkins, K. L., C. E. Canan, R. D. Moore, C. R. Lesko, and B. Lau (2018), An application of restricted mean survival time in a competing risks setting: comparing time to art initiation by injection drug use, *BMC medical research methodology*, *18*(1), 27.

Chen, P.-Y., and A. A. Tsiatis (2001), Causal inference on the difference of the restricted mean lifetime between two groups, *Biometrics*, *57*(4), 1030–1038.

Cook, R. J., and J. Lawless (2007), *The statistical analysis of recurrent events*, Springer Science & Business Media.

Cook, R. J., and J. F. Lawless (1996), Interim monitoring of longitudinal comparative studies with recurrent event responses, *Biometrics*, *52*, 1311–1323.

Cook, R. J., and J. F. Lawless (1997), Marginal analysis of recurrent events and a terminating event, *Statistics in medicine*, *16*(8), 911–924.

Cook, R. J., G. Y. Yi, and K. A. Lee (2010), Sequential testing with recurrent events over multiple treatment periods, *Stat Biosci*, *2*, 137–153.

DeKosky, S. T., M. D. Ikonomovic, and S. Gandy (2010), Traumatic brain injuryfootball, warfare, and long-term effects, *New England Journal of Medicine*, *363*(14), 1293–1296.

Demets, D. L., and K. G. Lan (1994), Interim analysis: the alpha spending function approach, *Statistics in medicine*, *13*(13-14), 1341–1352.

DeMets, D. L., and J. H. Ware (1982), Asymmetric group sequential boundaries for monitoring clinical trials, *Biometrika*, *69*(3), 661–663.

Faucett, C. L., N. Schenker, and J. M. Taylor (2002), Survival analysis using auxiliary variables via multiple imputation, with application to aids clinical trial data, *Biometrics*, *58*(1), 37–47.

Fischl, M., L. Parker, C. Petinelli, and et al. (1990), A randomized controlled trial of a reduced daily dose of zidovudine in patients with the aquired immunodeficiency syndrome, *The New England Journal of Medicine*, *323*, 1009–1014.

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

Frome, E. L., M. H. Kutner, and J. J. Beauchamp (1973), Regression analysis of poisson-distributed data, *Journal of the American Statistical Association*, *68*(344), 935–940.

Gehan, E. A. (1965), A generalized wilcoxon test for comparing arbitrarily singly-censored samples, *Biometrika,*, *52*, 203–223.

Ghosh, D., and D. Lin (2000), Nonparametric analysis of recurrent events and death, *Biometrics*, *56*(2), 554–562.

Ghosh, D., and D. Y. Lin (2002), Marginal regression models for recurrent and terminal events, *Statistica Sinica*, pp. 663–688.

Graw, F., T. A. Gerds, and M. Schumacher (2009), On pseudo-values for regression analysis in competing risks models, *Lifetime Data Analysis*, *15*(2), 241–255.

Greene, W. H. (1994), Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

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

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

Hougaard, P. (1995), Frailty models for survival data, *Lifetime data analysis*, *1*(3), 255–273.

Hsu, C.-H., J. M. Taylor, S. Murray, and D. Commenges (2006), Survival analysis using auxiliary variables via non-parametric multiple imputation, *Statistics in Medicine*, *25*(20), 3503–3517.

Irwin, J. (1949), The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice, *Epidemiology & Infection*, *47*(2), 188–189.

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

Jiang, W. (1999), Group sequential procedures for repeated events data with frailty, *J Biopharm Stat*, *9*, 379–399.

Karrison, T. (1987), Restricted mean life with adjustment for covariates, *Journal of the American Statistical Association*, *82*(400), 1169–1176.

Karrison, T. G. (1997), Use of irwin's restricted mean as an index for comparing survival in different treatment groups?interpretation and power considerations, *Controlled clinical trials*, *18*(2), 151–167.

Kim, D. H., H. Uno, and L.-J. Wei (2017), Restricted mean survival time as a measure to interpret clinical trial results, *JAMA cardiology*, *2*(11), 1179–1180.

Klein, J. P., and P. K. Andersen (2005), Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, *Biometrics*, *61*(1), 223–229.

Lambert, D. (1992), Zero-inflated poisson regression, with an application to defects in manufacturing, *Technometrics*, *34*(1), 1–14.

Lan, K., and D. L. DeMets (1983), Discrete sequential boundaries for clinical trials, *Biometrika*, *70*(3), 659–663.

Lan, K. G., and D. L. DeMets (1989), Group sequential procedures: calendar versus information time, *Statistics in Medicine*, *8*(10), 1191–1198.

Lawless, J. F. (1987), Negative binomial and mixed poisson regression, *Canadian Journal of Statistics*, *15*(3), 209–225.

Lawless, J. F., and C. Nadeau (1995), Some simple robust methods for the analysis of recurrent events, *Technometrics*, *37*(2), 158–168.

Li, D. X. (1999a), On default correlation: A copula function approach.

Li, Z. (1999b), A group sequential test for survival trials: an alternative to rank-based procedures, *Biometrics*, *55*(1), 277–283.

Lin, D., W. Sun, and Z. Ying (1999), Nonparametric estimation of the gap time distribution for serial events with censored data, *Biometrika*, *86*(1), 59–70.

Lin, D. Y., L. J. Wei, I. Yang, and Z. Ying (2000), Semiparametric regression for the mean and rate functions of recurrent events, *Journal of Royal Statistical Society: Series B*, *62*(4), 711–730.

Little, R. J., and D. B. Rubin (1986), *Statistical analysis with missing data*, John Wiley & Sons, Inc., New York, NY, USA.

Liu, L. X., S. Murray, and A. Tsodikov (2011), Multiple imputation based on restricted mean model for censored data, *Statistics in medicine*, *30*(12), 1339–1350.

Logan, B. R., and S. Mo (2015), Group sequential tests for long-term survival comparisons, *Lifetime data analysis*, *21*(2), 218–240.

Mantel, N. (1963), Chi-square tests with one degree of freedom; extensions of the mantel–haenszel procedure, *Journal of the American Statistical Association*, *58*, 690–700.

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

Mazroui, Y., S. Mathoulin-Pélissier, G. MacGrogan, V. Brouste, and V. Rondeau (2013), Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data, *Biometrical Journal*, *55*(6), 866–884.

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

Nicolaie, M., J. van Houwelingen, T. de Witte, and H. Putter (2013), Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks, *Biometrics*, *69*(4), 1043–1052.

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

Ozga, A.-K., M. Kieser, and G. Rauch (2018), A systematic comparison of recurrent event models for application to composite endpoints, *BMC medical research methodology*, *18*(1), 2.

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

Pepe, M. S., and J. Cai (1993), Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates, *Journal of the American statistical Association*, *88*(423), 811–820.

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

Peto, R., and J. Peto (1972), Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society, Series A*, *135*, 185–207.

Pocock, S. (1977), Group sequential methods in the design and analysis of clinical trials, *Biometrika*, *64*(2), 191–199.

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

Prentice, R. L., B. J. Williams, and A. V. Peterson (1981), On the regression analysis of multivariate failure time data, *Biometrika*, *68*(2), 373–379.

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

Quenouille, M. H. (1949), Approximate tests of correlation in time-series 3, in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 45, pp. 483–484, Cambridge University Press.

Quenouille, M. H. (1956), Notes on bias in estimation, *Biometrika*, *43*(3/4), 353–360.

Rogers, J. K., A. Yaroshinsky, S. J. Pocock, D. Stokar, and J. Pogoda (2016), Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model, *Statistics in medicine*, *35*(13), 2195–2205.

Rondeau, V., S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran (2007), Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events, *Biostatistics*, *8*(4), 708–721.

Royston, P., and M. K. Parmar (2011), The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt, *Statistics in medicine*, *30*(19), 2409–2421.

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

Schaubel, D. E., and G. Wei (2011), Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring, *Biometrics*, *67*(1), 29–38.

Schaubel, D. E., and M. Zhang (2010), Estimating treatment effects on the marginal recurrent event mean in the presence of a terminating event, *Lifetime data analysis*, *16*(4), 451–477.

Schwartz, G. G., et al. (2018), Alirocumab and cardiovascular outcomes after acute coronary syndrome, *New England Journal of Medicine*, *0*(0), null, doi: 10.1056/NEJMoa1801174, pMID: 30403574.

Taylor, J. M., S. Murray, and C.-H. Hsu (2002), Survival estimation and testing via multiple imputation, *Statistics & probability letters*, *58*(3), 221–232.

Tayob, N., and S. Murray (2014), Nonparametric tests of treatment effect based on combined endpoints for mortality and recurrent events, *Biostatistics*, *16*(1), 73–83.

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

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

Tian, L., L. Zhao, and L. Wei (2013), Predicting the restricted mean event time with the subject's baseline covariates in survival analysis, *Biostatistics*, *15*(2), 222–233.

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, *Journal of the American Statistical Association*, *77*, 855–861.

Tukey, J. (1958), Bias and confidence in not quite large samples, *Ann. Math. Statist.*, *29*, 614.

Uno, H., et al. (2014), Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis, *Journal of clinical Oncology*, *32*(22), 2380.

Wang, X., and D. E. Schaubel (2018), Modeling restricted mean survival time under general censoring mechanisms, *Lifetime data analysis*, *24*(1), 176–199.

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

Wei, L.-J., D. Y. Lin, and L. Weissfeld (1989), Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American statistical association*, *84*(408), 1065–1073.

Wilcox, M. H., et al. (2017), Bezlotoxumab for prevention of recurrent clostridium difficile infection, *New England Journal of Medicine*, *376*(4), 305–317, doi: 10.1056/NEJMoa1602615, pMID: 28121498.

Xia, M., and S. Murray (2018), Commentary on Tayob and Murray (2014) with a useful update pertaining to study design, *Biostatistics*.

Xiang, F., and S. Murray (2012), Restricted mean models for transplant benefit and urgency, *Statistics in medicine*, *31*(6), 561–576.

Xiang, F., S. Murray, and X. Liu (2014), Analysis of transplant urgency and benefit via multiple imputation, *Statistics in medicine*, *33*(26), 4655–4670.

Zhang, M., and D. E. Schaubel (2011), Estimating differences in restricted mean lifetime using observational data subject to dependent censoring, *Biometrics*, *67*(3), 740–749.

Zhang, M., and D. E. Schaubel (2012), Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies, *Biometrics*, *68*(4), 999–1009.

Zhao, H., and A. A. Tsiatis (1997), A consistent estimator for the distribution of quality adjusted survival time, *Biometrika*, *84*(2), 339–348.

Zhao, L., B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. Wei (2016), On the restricted mean survival time curve in survival analysis, *Biometrics*, *72*(1), 215–221.

Zucker, D. M. (1998), Restricted mean life with covariates: modification and extension of a useful survival analysis method, *Journal of the American Statistical Association*, *93*(442), 702–709.