Adaptive designs in the time to event setting: The potential for benefit and risk

William Jen Hoe Koh

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Scott Emerson, Chair

Susanne May

Marco Carone

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Adaptive designs in the time to event setting:
The potential for benefit and risk

William Jen Hoe Koh

Chair of the Supervisory Committee:
Professor Scott Emerson
Department of Biostatistics

Group sequential designs (GSDs) have been the standard sequential approach to maintain scientific, ethical, and efficiency goals in any confirmatory Phase III studies. Over the past two decades, adaptive extensions to group sequential designs have been proposed to allow more flexible modification of aspects of the trial. However, such use of unblinded estimates computed from accruing clinical trial data has been viewed extremely cautiously by regulatory agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). In their Guidance to Industry on adaptive designs, the FDA has distinguished adaptive designs that are "less well understood" from the "well understood" GSDs. There is thus much interest in characterizing potential benefits (e.g., efficiency, flexibility), as well as the potential for harm (e.g., inflation of statistical error rates, introduction of operational bias), when using adaptive designs. Randomized clinical trials (RCTs) involving censored time to event data are of particular interest, as the aspects of adaptive designs that are "less well-understood" in that setting may have as much to do with how well we understand standard censored data models as with the general properties of unblinded adaptation.

A major focus of any sequential procedure is the appropriate control of statistical operating characteristics, including the type 1 error. A common requirement of all commonly

used sequential methods is the proper characterization of the growth of statistical information about the parameter of greatest interest. When sequentially analyzing time to event data, the censoring distribution can have great impact. It can affect both the choice of the distributional parameter used to summarize treatment effect (e.g., 5 year survival, median survival, hazard ratios) and the growth of statistical information over time. Further, efficiency of inference might need to consider not only the number of subjects accrued to the study, but also factors related to time as measured by both the typical time of patients on study ("study time"), as well as the calendar time needed from the start of accrual until the final analysis is performed. In this research, we investigate how these issues of information growth and time may impact (a) scientific interpretation, and (b) statistical credibility (control of type 1 error and study precision).

In the first part of the research, we focus on the proportional hazards setting wherein issues of (1) calendar time and (2) information growth are separable. We first investigate the efficiency of the adaptive weighting scheme as a consequence of changing the timing of the adaptation in prevention trials with potentially low background rates (either as a consequence of overestimating the event rate and/or high treatment efficacy). Noting that GSDs are better able to avoid any operational bias that might be introduced by the more flexible forms of adaptive designs, we compare our ability to preserve study precision solely through the use of blinded adaptations within prespecified GSDs versus the use of an unblinded adaptation which might better distinguish between low event rates versus extreme treatment effects. We next investigate how poor understanding of information growth (in the weighted logrank statistics) can impact the ability of adaptive procedures to preserve the overall Type 1 error. We examined scenarios whereby simply changing the censoring distribution can directly impact the ability of adaptive procedures to preserve the overall Type 1 error. We provide some recommendations from our findings.

In the second part of our research we consider settings in which we cannot presume a

parametric or strongly semiparametric probability model, for instance, when crossing survival curves are plausible. Under the weak null/non proportional hazards setting, calendar time and information growth are no longer separable. We investigate the degree to which the use of "less well-understood" statistics in presence of time varying treatment effect and censoring as induced either sequentially or accrual affects the degree to which we can control the probability of rejecting the null hypothesis when we may be concerned with the weak null.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

RCT: Randomized Controlled Trial.

GSD: Group Sequential Design

FSD: Fixed sample design

EMA: European Medicines Agency

FDA: Food and Drug Administration.

DMC: Data Monitoring Committee or also known as Data Safety and Monitoring Board

PREP: Pre-Exposure Prophylaxis

HPTN052: HIV Prevention Network 052 trial

NLST: National Lung Screening Trial

I-SPY-2: Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging And moLecular Analysis 2

BATTLE: Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination

ASN: Average Sample Size

ODAC: Oncologic Drugs Advisory Committee

OBF: O'Brien Fleming.

MLE: Maximum Likelihood Estimate

CLT: Central Limit Theorem

CHW: Cui, Wang, and Hung 1999 approach to critical value readjustment in presence of sample size changes

PH: Proportional hazards

TTE: Time to event

FTC: Emtricitabine

TRUVADA: tenofovir/FTC

ART: anti-retroviral therapy

UMP: Uniformly most powerful

LR: Logrank statistic

NA: Nelson-Aalen statistic

RMS: Restricted mean statistic

WKM: Weighted Kaplan Meier

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Professor Scott Emerson for his guidance, patience, support in enabling me to complete this dissertation. I am especially grateful for the amount of time he spent ensuring that my research continued in the right direction. I enjoyed his most honest, critical, and yet thoughtful statistical advice that makes this research challenging and yet enlightening. I am especially honored to have Scott as an advisor and friend during my PhD study.

I would also like to thank the members of my dissertation committee, Susanne May, Marco Carone, and Barbara McKnight for their valuable time and dedication to Biostatistics. I am especially grateful to the Department of Biostatistics for funding me throughout the PhD program. In particular, I am grateful to have the opportunity to work with Mary Emond, Peter Rabinovitch on the Cancer and Aging Project which gave me ample opportunity to work with cutting edge new research technology. I am also grateful to have the opportunity to work with Alice Arnold, Traci Bartz, and Chris Delaney at the Collaborative Health Study Coordinating Center who gave me immensely useful applied advice.

I am also grateful to make new friends inside this program: Amanda, Arthur, Bob, Caitlin, David B, David P, Leigh, Lisa, Phillip Keung, Jing, Tracey, Navneet, and Elisa. I am thankful for having the opportunity to share lots of discussions, both statistical and non-statistical with all of you. To Caitlin, David P, Leigh, Navneet, Elisa, and Tracey: I truly treasure the time spent that was either necessary or unnecessary, either indoors or outdoors, either studying or eating, either working or not working, or simply and purely for the sake of hanging out. I especially enjoy the company with all of you whether in the sun, rain, freezing cold, or mountains. To Don, Gale, Gitana, and Jerry: Your presence, laughter,

help, and friendship throughout my stay in Seattle has very much keep me sane.

To my dearest Mum and Dad, thank you for raising me up, providing me unwavering encouragement and support as I embark on this program. The countless midnight phone calls back home while both of you patiently wait for me to complete this program. I am particularly thankful to have my brother, Shukai, whom has provided me lots of administrative and emotional support, and supporting my parents and now his wife, Huiyun, and filling in the role of the breadwinner to support my Mum and Dad for the past few years while I complete this PhD study. Last but not least, I am especially grateful to my beloved friends in California, Larry and Loy, for providing me lots emotional support, guidance on various aspects of life, and making this PhD journey less arduous and lonely.

Chapter 1

Introduction

1.1 Goal of Clinical Trials

Before a new treatment/intervention can be approved by the regulatory bodies for public use, it undergoes a series of rigorous clinical trials to investigate its safety, efficacy, and effectiveness. Each of the phases of the drug discovery process may have different goals. Phase 1 studies involve administration of the treatment in screened human volunteers to investigate the appropriate dosing of the treatment, as well as to identify major safety concerns or side effects of the treatment that might preclude further study. Phase 2 clinical trials are conducted in a larger population of screened human volunteers to investigate preliminary safety and efficacy, and to refine dosing strategies of the treatment. Phase 3 clinical trials are typically confirmatory in nature, using a much larger sample of human volunteers. The goal of Phase 3 studies is to determine whether the treatment is safe, efficacious, and effective in the ultimate intended population.

The focus of this dissertation is largely concerned with Phase III confirmatory randomized clinical trials (RCT). Such trials may be directed toward superiority, (bio)equivalence, or non-inferiority of an experimental treatment relative to the current standard of care. The primary objective is to provide scientifically interpretable results with statistically reliable inference. It is vital that any confirmatory Phase III clinical trial be conducted in an ethical, efficient, and rigorous way to minimize any form of operational bias. Well-executed confirmatory trials can provide definitive proof to support the hypothesized claims and enable regulatory bodies (e.g., European Medicines Agency [EMA], US Food and Drug Administration [FDA])

to appropriately label new treatments, thereby facilitating clinicians' use of these treatments in an evidence-based setting.

In recent years, much focus has been on the way in which scientifically rigorous and statistically credible Phase 3 clinical trials can be designed to afford the greatest flexibility during the conduct of the RCT. Such flexible designs are targeted to satisfy a variety of optimality criteria, including the frequentist Type 1 and Type 2 statistical errors and statistical efficiency. In this introductory chapter, we present an overview of the topics to be covered as we explore the special issues presented by the use of adaptive clinical trials in the setting of censored time to event data.

1.2 Conventional Clinical Trial Designs

Historically, Phase III clinical trials were conducted using fixed sample designs (FSDs). In a FSD, we randomize and treat a predefined number of participants, and only analyze the data when all outcomes have been observed. We introduce the general principles behind FSDs in section 2.1 along with notation that will serve as the basis for later sections. While FSDs are still the most common and convenient RCT designs, they have both ethical and efficiency limitations. When a treatment provides definitive proof of superiority over placebo, we may want to adopt the treatment quickly so as to expedite delivery to patients and improve public health. Alternatively, if the treatment indicates evidence of harm, we want to stop randomizing subjects to the treatment as soon as possible in order to protect the interests of the patients on the study. Such might also allow them to participate in other trials with potentially effective treatments. In situations when the treatment is neither markedly effective nor harmful when compared to control, it is also advantageous to stop the trial early and reallocate resources to other trials. Thus, it is desirable to periodically monitor the accruing data during the conduct of a clinical trial to safeguard the interests of the human volunteers as well as improve the efficiency of the design. Such a process constitutes a sequential design.

Group sequential designs (GSDs), currently the gold standard for such monitoring, over-

come the shortcomings of FSDs by allowing for periodic looks at the data during the course of the trial after groups of subjects/events are accumulated. There is a rich statistical literature describing methods for group sequential sampling [Whitehead, 1997, Jennison and Turnbull, 1999]. Key to these methods is the quantification of the "information growth", which is inversely related to the variability of the test statistics at interim analyses compared to the variability of the test statistic at the planned final analysis [DeMets and Lan, 1995, Jennison and Turnbull, 1997, Scharfstein et al., 1997, Burington and Emerson, 2003]. A "stopping boundary" is first chosen to determine thresholds for making clinical decisions, where the threshold at any given analysis is based on the statistical information available at that analysis. A broad spectrum of "boundary shape function" have been characterized, each of which would control the type 1 and 2 statistical errors while addressing particular needs for caution at the earliest interim analyses or the need for efficiency with respect to average stopping time. Because such stopping rules have the potential to introduce bias into the classical statistical analysis methods, a variety of statistical procedures were also developed to enable frequentist inference to compensate for the bias in presence of early stopping [Tsiatis et al., 1984, Whitehead, 1986, Chang, 1989, Emerson and Fleming, 1990]. In section 2.2 and 2.3, we describe ways in which the structure of a group sequential stopping rule is carefully constructed to protect against multiple comparison issues. These methods provide a basis for understanding adaptive RCT designs.

Key to the use of GSD is the principled implementation of the stopping rule to avoid the bias of investigators influencing the schedule of analyses. Group sequential monitoring is overseen by an independent Data Monitoring Committee (DMC or Data Safety and Monitoring Board) whose main role is to protect the integrity of such unblinded use of interim results as well as the well being of participants [Ellenberg et al., 2003]. Guidance documents for DMC functioning are well established by both regulatory and academia to protect against operational bias during the conduct of a clinical trial [Whitehead, 1997, Jennison and Turnbull, 1999, Ellenberg et al., 2003, Food et al., 2006]. However, the design assumptions used in the selection of monitoring rules at the planning phase may often prove to be incorrect

during the conduct of the study. Thus, there is a desire to adapt the RCT design to reflect updated information about the exact schedule of analyses [DeMets and Lan, 1995, Burington and Emerson, 2003] or the variance of individual observations [Gould and Shih, 1998, Gould, 1992]. Proschan et al. investigated the extent to which data driven changes in the schedule of analyses in a GSD might affect the type 1 error and cautioned against indiscriminate alteration of the sampling plan. Section 2.4 touches on various aspects of blinded adjustments to the analysis schedule or sample size re-estimation that can be implemented in the context of GSD without adversely affecting statistical error rates.

However, there has been a desire to incorporate more extensive modification(s) to RCT designs during the conduct of a study in a flexible manner. In these generalized sequential sampling plans, conventional statistical adjustments may no longer be sufficient to protect against inflation of overall Type 1 error. This led to the explosion of literature in adaptive clinical trial design.

1.3 Adaptive RCT

There has always been considerable interest in the possibility of incorporating adaptive features into the design of the clinical trial with the goal of modifying either scientific or statistical aspects of the trial. Some of these motivations arise from the desire to use unblinded trial data when deciding to (1) re-power studies for unanticipated differences in treatment effect, (2) modify randomization ratios, (3) drop treatment arms, (4) modify the definition of the primary outcome based on interim data, and/or (5) modify eligibility criteria. Historically, such modifications have been performed sequentially across the different phases of clinical trials in a lengthy drug discovery process. More recent innovative strategies have considered a seamless Phase II-III design with the aim of reducing this calendar time (often referred to as "white space") between the conclusion of a Phase II study and the start of a Phase III study. Similarly, there has also been strong interest in making such modifications within a single phase of investigation.

One major focus of the early adaptive literature is the use of interim outcome data to

extend a trial beyond some pre-specified maximum stopping time. A common approach was to design some GSD, and then use the unblinded trial results at the penultimate interim analysis to choose the final, maximal sample size. In section 2.5, we illustrate how such an adaptive design differs from the classical GSD and describe how standard GSD software can allow selection of critical values when the adaptive rule is fully pre-specified. In that sense, the relevance of section 2.3 is highlighted, because these adaptive designs can be thought of as varying the timing of analyses and the conservatism of the boundary shape functions of a GSD.

It should be noted however, that a major interest of the adaptive RCT literature is to allow modifications to a RCT design that have not been fully planned in advance. Because it was well-known that unplanned sample size modifications can lead to an inflation of the overall Type 1 error [Proschan et al., 1992, Proschan and Hunsberger, 1995], the earliest statistical literature was primarily concerned with protecting the overall Type 1 error [Bauer and Köhne, 1994, Proschan and Hunsberger, 1995, Fisher, 1998, Cui et al., 1999, Denne, 2001, Chen et al., 2004, Müller and Schäfer, 2004, Gao et al., 2008, Mehta and Pocock, 2011]. In section 2.6, we provide a summary of these methods, which can be implemented so long as the information growth can be quantified at each stage of the design. Later authors noted that the same adjustments would further allow modifications to randomization ratios or subgroup enrichment.

Emerson [2006], Fleming [2006], and Emerson and Fleming [2010] pointed out numerous scientific, statistical and operational issues with unblinded adaptations. However, three issues that arise with the use of these flexible procedures are of particular interest. First, these procedures choose rules based on weighted statistics which violate the sufficiency principle, thus potentially leading to an unnecessary loss of efficiency [Mehta and Tsiatis, 2001, Jennison and Turnbull, 2003]. Second, there has not been a clear definition on what is considered a good or bad adaptive rule. Third, adaptive features made on the basis of unblinded data can not only inflate the overall Type 1 error through sequential testing, but can further introduce bias from sponsors/researchers who have potential conflict of interest. Such possibilities

might raise concerns about the validity of the trial results.

In FDA's "Guidance on Adaptive Design for Industry", FDA contrasts the "well-understood" designs of FSDs and GSDs with the "less well-understood" adaptive designs. We discuss some of the difficulties that must be addressed when using adaptive RCT and review some added challenges that arise in the setting of time to event analyses (section 2.7). When addressing these "less well understood" issues in the time to event analyses, it became important to better quantify additional properties of "well understood" GSD to enable us to characterize the specific aspects of adaptive designs that might be contributing to loss of efficiency. These new investigations are presented in Chapter 3.

1.4 Adaptive RCT in the Time To Event Setting

This dissertation is motivated by some unresolved statistical issues in the time to event setting with the use of adaptive designs. Our goal is to enhance the understanding of the benefits/limitations on the use of adaptive design in the broader context of time to event analysis. It is the thesis of this dissertation that much of what is considered "less well-understood" relates closely to the interplay between adaptive designs and the longitudinal nature of most RCT. That is, "less well-understood" analysis methods have much to do with "less well-understood" adaptive designs. We regard this to be especially true in the analysis of censored time to event data, and provide relevant properties of such analyses in Chapter 4.

1.5 Unmet Needs in the Time To Event Setting

In typical time to event RCTs, patients are accrued over some period of time. After being randomized to the treatment/placebo, they are then followed until some event of interest. During the RCT, interim analyses may be conducted prior to or after completion of accrual, giving rise to many issues related to the censoring distribution. Furthermore, in a sequential study, the censoring distribution typically varies substantially across interim analyses (sec-

tion 4.1). The time frame of observation that is of greatest interest should drive the choice of summary measure used to compare treatment groups (section 4.2), though the range of such parameters that can be estimated is similarly affected by the censoring distribution. This then often dictates whether a strong or weak null hypothesis is of greatest interest (section 4.4). When the contrast across treatment groups might be subject to time varying effects, for example, non proportional hazards (PH) when using Cox regression, the censoring distribution encountered in sequential analysis with incomplete accrual complicates inferential methods (section 4.3). We review how these issues affect the most common statistical analysis model (proportional hazards model and unweighted logrank statistic) used for such data in section 4.5, as well as the less common analysis methods (such as weighted logrank test, Nelson-Aalen, or Weighted Kaplan-Meier statistics) in section 4.6.

The original research presented in Chapter 3 and 5 - 7 in this dissertation is directed toward better understanding of the use of adaptive RCT with time to event data. In particular, we consider three important settings where current knowledge and methods are lacking:

1. A comparison of GSDs vs adaptive designs when event rates are low, but strong treatment effects are plausible,

2. The use of weighted log rank tests, as an example of a non-PH analysis, with adaptive RCTs, and

3. The sequential analysis of time to event data when weak null hypotheses are of greatest interest (e.g., when crossing survival curves are plausible, or in non inferiority trials).

Our organization in this dissertation is directed towards addressing the important settings as laid out first under the PH setting, which includes the strong null hypothesis of exact equality of entire survival distributions. We investigate

1. The (a) impact of early adaptations based on weighted statistics and (b) the impact of changing the analysis schedule as a function of the degree of early conservatism (Chapter 3).

2. The flexibility of (prespecified) adaptive procedures in the time to event setting in presence of extreme treatment effect (Chapter 5).

3. The impact of adaptive procedures on the overall Type 1 error and power of the (weighted) log-rank statistic when changing the censoring distribution (Chapter 6).

Under the setting of non PH/weak null hypothesis in Chapter 7, we investigate

1. how the censoring distribution can affect (a) the information growth across interim analyses, (b) the control of the probability of rejecting the null hypothesis, and (c) the test statistic/estimate as well as the interpretation in the presence of a time varying treatment effect, including the possibility of crossing survival curves,

2. the use of alternative summary measures to identify the better treatment under fixed sample and group sequential setting when considering a weak null hypothesis, and to illustrate the dilemmas faced by DMC in sequential adaptive monitoring.

Finally, in Chapter 8, we summarize our findings and describe the operational bias that arises as a consequence of adaptations, and we comment on the open statistical questions in adaptive designs in the time to event setting.

The scope of this dissertation did not investigate the potential availability of secondary statistical information during interim analysis that might be used to make adaptations. If adaptations were made on the basis of this secondary data, it is possible to inflate the overall Type 1 error even after adjusting for having looked at the primary outcome [Bauer and Posch, 2004]. Since accumulating data used for adaptations may include statistical information not captured by the sufficient statistics, generalized procedures that do not account for the correlation between the secondary data and primary endpoint would not adequately control the overall Type 1 error. Several authors have since explored some of the statistical issues [Jenkins et al., 2011, Irle and Schäfer, 2012, Magirr et al., 2014, Mehta et al., 2014, Magirr et al., 2016]. This issue is of special concern in the time to event setting in designing seamless Phase 2/3 trials or the use of biomarkers in enrichment designs. However, this is beyond the scope of this dissertation, though we do include comments in Chapter 8 on how our findings might magnify these issues.

Chapter 2

Background: Fixed Sample, Group Sequential and Adaptive Designs

We introduce the notation for the fixed sample, group sequential and adaptive design in the immediate outcomes settings. Because FSD can be considered a subset of GSD, and GSD can be considered a subset of adaptive design, in this dissertation we find it useful to distinguish adaptive designs that cannot be considered a GSD. Thus, we adopt the modified definition of an adaptive design from FDA as one whereby aspects of the study design may be modified on the basis of unblinded interim data. This definition precludes GSD but includes both pre-specified adaptive design and the more flexible fully adaptive design. We provide a simple example so to highlight differences between the 3 classes of designs. The notation developed in the context of immediately observed outcomes (or nearly immediate relative to accrual period) serves to transition to the time to event or longitudinal setting. We also describe various statistical aspects in the time to event setting that are typically not an issue in immediate settings. It is these special characteristics of time to event data analyses that present particular issues in both group sequential and adaptive designs.

2.1 Fixed Sample Design

In a classical fixed sample clinical trial, the data that are gathered on all participants randomized to the treatment and placebo arms are only analyzed when the study has concluded. During the planning of the RCT, the (bio)statistician collaborating on the clinical trial is asked to estimate the sample size required to address the scientific question of interest. The

answer to this question is governed by many factors such as the scientific question of interests, the primary endpoint(s), logistical and financial constraints, as well as calendar time. The statistician, thus becomes the middleman in balancing all of the constraints that are implicitly imposed by parties such as clinical researchers, regulatory bodies, ethicists, etc.

To determine the appropriate sample size, the clinical team must define not only the clinical variable that will be measured as the RCT's primary outcome, but also how the distribution of that outcome will be summarized (e.g., mean, median, hazard) within each treatment arm and contrasted (e.g., difference, ratio) across treatment arms. We use θ to represent the target parameter of interest that reflects the potential benefit/harm that a new treatment has over some current standard of care. Without loss of generality, we can assume that larger values of θ are indicative of superiority of the new treatment. At the time of designing the trial, we also indicate some null value θ_0 that potentially represents no advantage of the new treatment, and some design alternative θ_{Alt} that represents a clinically important improvement.

Unless otherwise stated, we will be primarily interested in the 1-sided hypothesis testing where large positive values of θ reflect the superiority of the new experimental treatment over the current standard. Less often, investigators may be interested in demonstrating non-inferiority or equivalence and involve testing a different hypothesis. We next describe notation in the setting of a primary response variable that is a continuous and immediately available outcome.

2.1.1 Notation

Consider the balanced two-sample design in Jennison and Turnbull [1999] where potential observations X_{Ai} randomized to treatment A and potential observations X_{Bi} randomized to treatment B are immediately observed in the clinical setting. X_{Ai} and X_{Bi} are independent and distributed with mean ω_A and ω_B respectively with known variance σ^2 for $i = 1, \cdots, N$. Our target parameter of interest, $\theta = \omega_A - \omega_B$, is the difference in the mean of the responses comparing subjects randomized to group A (experimental group) relative to subjects ran-

domized to group B (placebo group). We are concerned with testing the null hypothesis of $\mathbb{H}_0 : \theta \leq \theta_0$ against the 1-sided alternative of $\mathbb{H}_A : \theta \geq \theta_{\text{Alt}} > \theta_0$ at some one-sided level α.

When all the data have been gathered, we can construct the following test statistics with its respective asymptotically derived approximate distribution on the right:

1. Partial sum statistics: $S = \sum_{i=1}^{N}(X_{Ai} - X_{Bi})$ $\qquad S \stackrel{\cdot}{\sim} \mathcal{N}(N\theta, 2N\sigma^2)$
2. MLE estimate: $\widehat{\theta} = S/N$ $\qquad \widehat{\theta} \stackrel{\cdot}{\sim} \mathcal{N}(\theta, 2\sigma^2/N)$
3. Normalized Z statistic/Wald: $Z = (\widehat{\theta} - \theta_0)/\widehat{\text{se}}(\widehat{\theta} - \theta_0)$ $\qquad Z \stackrel{\cdot}{\sim} \mathcal{N}(\delta, 1)$
4. Fixed sample P-value statistic: $P = 1 - \Phi(Z) = 1 - \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp^{-u^2/2} du$

where $\delta = (\theta - \theta_0)\sqrt{N/2\sigma^2}$. Under $\mathbb{H}_A$, $\delta_A = (\theta_{Alt} - \theta_0)\sqrt{N/2\sigma^2}$ is commonly referred as our standardized design alternative. This standardized notation later comes into use in Chapter 7. The asymptotic statistical information based on 2N subjects is $\mathcal{I} = N/(2\sigma^2)$. $S/(2\sigma^2) \stackrel{\cdot}{\sim} \mathcal{N}(\theta\mathcal{I}, \mathcal{I})$ is the score statistic. The notation in this section can be generalized to other settings [Whitehead, 1997].

Our sample size for each group can be determined based on some given level α, and statistical power β, to discriminate between $\mathbb{H}_0 : \theta \leq \theta_0$ vs the alternative of interest $\mathbb{H}_A : \theta \geq \theta_A$ using the general formula

$$N = \frac{(z_{1-\alpha} + z_\beta)^2 V}{(\theta_{\text{Alt}} - \theta_0)^2}$$

where z_p denotes p^{th} quantile of a standard normal distribution for $p \in (0, 1)$, and V is the variance contributed by a single sampling unit. Based on the above setup and assuming a 1:1 randomization, $V = 2\sigma^2$. V will be used to describe the procedure when blinded revision of sample size is applied.

2.2 Group Sequential Designs

It is often desirable to make interim looks during the course of the trial to balance scientific, ethical and efficiency concerns. Armitage et al. [1969] quantified the inflation in type 1 error

that results if the accruing data were analyzed at multiple interim looks and assessed naïvely using the "standard" statistical critical values appropriate for FSD. Valid statistical inference instead needs to use critical values that account for the multiplicity of analyses, using the correlation structure induced by the repeated significance testing. Many methods have been proposed to derive boundaries for sequential clinical trials that protect against inflation of Type 1 error. In particular, two of the earliest described approaches for stopping boundaries [Pocock, 1977, O'Brien and Fleming, 1979] are often used as examples to illustrate the relative trade-offs between efficiency defined by average sample size(ASN) and the degree of early conservatism that would allow more precise assessment of safety and secondary endpoints. Some hybrids of the above are also used.

Group sequential designs allowing for interim monitoring of data for efficacy and futility balance the scientific, ethical, and efficiency goals in the design and conduct of confirmatory RCTs. At each interim analysis, a summary statistic or estimand of interest is computed. The test statistic or summary statistic is then compared to pre-defined critical values at the interim analysis. If the test statistic/summary measure is large and in the right direction suggestive of efficacy, the trial is stopped. Otherwise, if the test statistic is in the opposite direction and much smaller than the predetermined boundaries, the trial is stopped for futility.

In the remainder of this section, we introduce the notation for group sequential design, the stopping sets, the sampling density, and the general class of designs as described by the unified family [Kittelson and Emerson, 1999]. We then describe some of the asymptotic distribution of the test statistics and the frequentist inference following a sequential procedure. Later, in section 2.3, we elaborate on the choice of the sequential stopping rule based on either the scientific criterion or optimality criterion from a statistical standpoint.

2.2.1 Notation and Stopping Sets

It is often the case that clinical trialists plan a fixed sample design and then expand the design to incorporate sequential monitoring [Emerson et al., 2007]. Recall the hypotheses of

interest in the FSD: $\mathbb{H}_0 : \theta \leq \theta_0$ vs the alternative hypothesis, $\mathbb{H}_A : \theta \geq \theta_{\text{Alt}} > \theta_0$.

Suppose a total of J analyses are potentially conducted at sample sizes $N_1, \cdots, N_J$ accrued on each arm in the balanced setting for $j = 1, \cdots, J$ analyses. The capitalized notation for N and J are used to indicate that we allow the number and timing of analyses in a GSD to be random, though as discussed later, it is important that the schedule of analyses be independent of the unblinded estimates of treatment effect. We shall denote small letter, n_j, for realizations of random variables N_j, and for convenience in later formulas we denote the size of the groups accrued between analyses using an asterisk: $N_j^* = N_j - N_{j-1}$, with n_j^* denoting a particular realization of N_J^*.

At the j^{th} interim analysis, we use the accrued observations to compute the test statistic, $T_j = H(\boldsymbol{X}_{Aj}, \boldsymbol{X}_{Bj})$, where $\boldsymbol{X}_{Aj} = \{X_{A1}, \cdots X_{An_j}\}$ and $\boldsymbol{X}_{Bj} = \{X_{B1}, \cdots X_{Bn_j}\}$ and H to be some function of the summarized measure comparing group A and B. We can then partition the sample space T_j into stopping set $\mathcal{S}_j$ and continuation set $\mathcal{C}_j$ where $\mathcal{S}_j \cap \mathcal{C}_j = \varnothing$ and $\mathcal{S}_j \cup \mathcal{C}_j = \Re^1$. Beginning at the first interim analysis with $j = 1$, we obtain a test statistic T_j. Subsequently, if $T_j \in \mathcal{S}_j$, we stop the study and analyze the data. If $T_j \in \mathcal{C}_j$, we proceed to the $(j+1)^{\text{th}}$ interim analysis after accruing n_{j+1}^* subjects on each arm. By choosing $\mathcal{C}_J = \varnothing$, we guarantee that the trial will definitely stop at or before the J^{th} analysis. It can be shown that a minimal sufficient statistic for θ in this setting can be described on the partial sum statistic scale as (N_M, S_M), where stopping time $M = \min\{1 \leq j \leq J : S_j \notin \mathcal{C}_j\}$.

There are many ways of specifying the stopping sets and continuation sets. In particular, Kittelson and Emerson [1999] noted that continuation sets can be generalized to be of the form $\mathcal{C}_j \equiv \{(a_j, b_j] \cup [c_j, d_j)\}$ such that $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$. The boundary values a_j, b_j, c_j, and d_j typically represent the critical values that are used to make some decision rule on efficacy/equivalence/non efficacy/harm of the treatment relative to the placebo.

In a one-sided GSD investigating the efficacy/effectiveness of the treatment over the current standard, it is convenient to set $b_j = c_j$ in most situations. We shall be interested in comparing T_j with the boundary values of a_j or d_j to assess futility/non efficacy or efficacy respectively at the interim analysis. We will later describe how the stopping and continuation

sets can be further parameterized to obtain some of the common monitoring rules such as O'Brien Fleming (OBF) monitoring boundary.

On the basis of the accrued data at the j^{th} interim analysis, the test statistic T_j might correspond to any one of the following statistics: the partial sum statistic S_j, MLE $\widehat{\theta}_j$, standardized Z statistic Z_j, or fixed sample P value statistic P_j, each of which are defined in a manner analogous to that used in FSD. In addition, clinical trialists might consider the error spending scale [Lan and DeMets, 1983], a conditional probability scale, or a Bayesian predictive probability scale. As discussed in Emerson et al. [2007], when the variance σ^2 is known, these scales are 1 to 1 monotonic (usually nonlinear) transformations of each other. The interested reader is referred to that tutorial in Emerson et al. [2007] for more details about the exact relationships among these scales.

It is, however, of particular interest to comment further on the conditional power scale, because many authors of manuscripts on adaptive designs have proposed using adaptive rules in which a new adaptive sample size is a function of the conditional power (see section 2.6). The conditional power is most often computed at the j^{th} interim analysis as the estimated probability of achieving "statistically significant" results at the final J^{th} analysis, conditional on the observed trial results (at some j^{th} interim analysis where $S_j = s_j$) and some assumption θ^* about the true treatment effect θ. This is defined as follows:

$$C_j(d_J, \theta^*) = \Pr(S_J > d_J | S_j = s_j; \theta = \theta^*) = \Phi\left(\frac{d_J - s_j - \theta^* + [N_J - N_j]}{\sqrt{V[N_J - N_j]}}\right)$$

Common choices for θ^* might be the design alternative θ_{Alt} or the MLE at the j^{th} analysis $\hat{\theta}_j = \frac{s_j}{N_j}$. It should again be noted that when σ^2 is known, stopping boundaries chosen for one boundary scale can be easily transformed to stopping boundaries for another scale. However, the appropriateness of naively derived thresholds on the various scales is often not well understood by clinical trialists [Emerson et al., 2005, 2011b, Levin, 2013].

2.2.2 Independent Increment Structure

When defining some forms of the test statistics in the FSD setting, we appealed to approximate distributions based on asymptotic large sample results. In GSDs, these approximate distributions do not hold in general for the cumulative statistics due to the multiplicity of repeated significance testing and the sequential stopping rule. Rather, by appealing to the independence of the increments of observations accrued between interim analyses, the method of Armitage et al. [1969] can be used to compute the exact sampling distribution of the sequential test statistics.

To facilitate notation used later for the adaptive setting, we again note the use of the superscript * to represent the incremental data between interim analyses. We denote N_j^* to be the incremental sample size accrued between the $j-1^{\text{th}}$ and j^{th} analyses, i.e., $N_j^* = N_j - N_{j-1}$. The incremental statistics S_j^*, Z_j^*, θ_j^* used to denote data that are accrued during the j^{th} interim analysis can be expressed as follows:

$$
\begin{aligned}
S_j^* &= \sum_{i=N_{j-1}+1}^{N_j} (X_{Ai} - X_{Bi}) = S_j - S_{j-1} \\
\widehat{\theta}_j^* &= S_j^*/N_j^* \\
Z_j^* &= \sqrt{N_j^*}(\widehat{\theta}^* - \theta_0)/\sqrt{2\sigma^2} \\
P_j^* &= 1 - \Phi(Z_j^*)
\end{aligned}
$$

Conditional on the schedule of analyses, $S_j^*|(N_1, \ldots, N_j) \dot{\sim} \mathcal{N}(N_j^*\theta, 2\sigma^2 N_j^*)$. Under the incremental $\mathbb{H}_0^j$, the incremental statistics $Z_j^* \sim \mathcal{N}(0,1)$, $P_j^* \sim U(0,1)$ are conditionally independent of each other. However, the marginal and joint distribution of these statistics may not be well-characterized under alternatives. Our cumulative test statistics at interim analyses can be expressed in term of incremental statistics from each stage.

$$S_j = \sum_{i=1}^{N_j}(X_{Ai} - X_{Bi}) = \sum_{l=1}^{j} S_l^*$$

$$\widehat{\theta}_j = \frac{\sum_{l=1}^{j} N_l^* \widehat{\theta}_l^{\,*}}{\sum_{l=1}^{j} N_l^*}, \quad \text{since } \sum_{l=1}^{j} N_l^* = N_j$$

$$Z_j = \sqrt{\frac{N_j}{2\sigma^2}}\left(\frac{\sum_{l=1}^{j} N_l^* \widehat{\theta}_l^{\,*}}{N_j} - \theta_0\right) = \frac{\sum_{l=1}^{j} N_l^* (\widehat{\theta}_l^{\,*} - \theta_0)}{\sqrt{2\sigma^2 N_j}} = \sum_{l=1}^{j} \sqrt{\frac{N_l^*}{N_j}} Z_l^* = \sum_{l=1}^{j} w_l^* Z_l^*$$

where $w_l^* = \sqrt{N_l^*/N_j}$ for $l = 1, \cdots, j$.

A key assumption of a GSD is that the schedule and timing of interim analyses is independent of the estimates of treatment effect. In that setting, the above probability model is said to have "independent increments" in the sequential literature. Hence, the distribution of the minimal sufficient group sequential test statistic can be given in the form used by Armitage et al. [1969] as recursive convolutions of normal densities with a truncated density from the previous stage. (Later, in section 2.5, when we consider the adaptive setting, our incremental statistics may no longer be independent under the alternatives since the future interim analysis is dependent on the estimated treatment effect at previous analyses.)

For convenience, we describe the sampling distribution of our test statistics on the partial sum statistic scale. When independent increments hold, we can write the sampling distribution of the partial sum statistic based on the stopping sets for the observation $(M = m, S = s)$ that is defined recursively via Armitage et al. [1969] as

$$p_{M,S,\theta} = \begin{cases} f(m, s, \theta) & s \notin \mathcal{C}_m \\ 0 & \text{else} \end{cases}$$

where the (sub)density function $f(j, s; \theta)$ is further defined as:

$$f(1, s; \theta) = \frac{1}{\sqrt{n_1^* V}} \phi\left(\frac{s - n_1^* \theta}{\sqrt{n_1^* V}}\right)$$
$$f(j, s; \theta) = \int_{\mathcal{C}_{j-1}} \frac{1}{\sqrt{n_j^* V}} \phi\left(\frac{s - u - n_j^* \theta}{\sqrt{n_j^* V}}\right) f(j-1, u, \theta) du, \qquad j = 2, \cdots, J$$

with $\phi(x) = \exp^{-x^2/2} / \sqrt{2\pi}$ denoting the density of a standard normal distribution where $n_j^* = n_j - n_{j-1}$ and $n_0 \equiv 0$.

When a group sequential design terminates, statistical inference is desired to provide some quantification of the treatment effect, i.e, a point estimate, $100(1-p)\%$ confidence interval, and some p-value. As a consequence of the sequential sampling, the sampling density above is no longer an approximate normal distribution. Special inference procedures are required to account for this possibility of early stopping. This adjusted inference makes use of the above sampling distribution of the minimal sufficient statistic to derive bias-adjusted or median unbiased estimates, exact confidence intervals, and exact P values. Various authors have explored the relative behavior of alternative strategies in the GSD setting [Tsiatis et al., 1984, Whitehead, 1986, Chang, 1989, Emerson and Fleming, 1990].

2.2.3 Families of Designs

When later comparing the operating characteristics of particular adaptive designs, we will want to appeal to the current knowledge about GSD. In section 2.5, we see that commonly implemented adaptive designs can be viewed as stochastically switching between alternative GSDs. In these settings, the behavior of the adaptive design can sometimes be anticipated based on the properties of GSDs having different maximal sample sizes, different schedules of analyses, and different boundary shape functions. It is thus useful to consider the general behavior of GSD within a broad family of GSDs.

Families of GSDs can be defined on a variety of scales as noted above. Perhaps most commonly used are scales based on the MLE, Z statistic, or fixed sample P values, error spending

scales, or scales based on conditional or Bayesian predictive power. For our purposes, we are primarily interested in how "early conservatism" and schedule of analyses might affect the operating characteristics of GSDs, and thus it is immaterial which family of GSDs we might use for our investigations. For convenience we use a family defined on the MLE scale.

The unified family [Kittelson and Emerson, 1999] includes many of the most commonly used sequential sampling schemes (O'Brien Fleming, Pocock, Wang and Tsiatis [Wang and Tsiatis, 1987]). In the unified family, the "boundary shape function" linking the critical values across analyses is of the form $f(\Pi; A, P, R)G = (A + \Pi^{-P}(1-\Pi)^{-R})G$, where $\Pi_j = N_j/N_J$ is the proportion of maximal sample size observed at the j^{th} analysis, and parameters A, P, and R are typically viewed as controlling the "early conservatism" of the GSD. The values of G are typically obtained by a numerical search to ensure that the boundaries formed would provide the right level of α to test the null hypothesis, and to distinguish the alternative with some appropriate power. For certain choices of A, P, and R, we can choose the threshold at the earliest analyses to be so extreme that the GSD will only allow early stopping when there is compelling evidence of benefit or harm.

The boundary shape characterizes the degree of conservatism at early analyses, which in turn affects the efficiency of the design. The O'Brien Fleming (OBF) monitoring rule can be specified by setting $f(\Pi, 0, 1, 0) = \Pi^{-1}$ while the Pocock design is specified using $f(\Pi, 0, 0.5, 0) = \Pi^{-0.5}$. The OBF designs are the most common designs used in practice due to its early conservatism. The Pocock designs are also frequently used to explore GSD methodology, as they tend to be approximately efficient in terms of average sample size. While it is immaterial how the stopping rule is defined, it is of concern how the stopping rule can affect the operating characteristics of the study which in turn affects the scientific, ethical, and efficiency when competing designs are of interest.

2.3 Choice of Stopping Rules

In a fixed sample level α design, the conventional boundary on the Z scale is determined such that when the alternative is true, the power is roughly β. This enables one to determine

the sample size required to enable discrimination of the hypothesis in order to identify the treatment benefit at the time when all data are accrued and analyzed. In selecting an appropriate design, we often consider an optimality criterion that maintains the competing goals of science, ethics, and efficiency. In experimental designs, the best designs are selected under which "bias and variance are minimized", and "cost is minimal" [Sanchez, 2014].

With a sampling/monitoring rule, there is no single unique way to pick a design in this over-parametrized space. When considering alternative stopping rules at the design stage, there may be many criteria that we may choose to constrain or optimize. It is conceivable that one chooses to fix the overall Type 1 error, power for a specific alternative of interest, and the number/timing of interim analysis. Among designs with say 97.5% power at the hypothesized alternative, for a one sided alternative under a symmetric design, the unified family still includes a large class of stopping boundaries with varying degrees of early or late conservatism. Differences among such alternative designs with respect to other operating characteristics may constitute criteria that we want to further constrain or select to optimize.

One of the most common optimality criterion to minimize is the sample size. Since the sample size is a random variable, it is often of interest at the design stage to consider the average sample size in the study, because the number of subjects is directly related to the cost of the trial. However, other optimality criterion may be considered at design stage that more appropriately suit other aspects. This can include power under the hypothesized treatment effects (which includes the null hypothesis for type I error and alternative for the power), other aspects of the sample size distribution (75th percentile or maximum), or probability of early termination at each interim analysis. Just as with the statistical power afforded by the GSD, the sample size distribution and stopping probabilities vary as a function of the true treatment effect.

This last aspect relates to the probability of having an adequate sample size to be able to assess safety endpoints and other important secondary endpoints. Hence, we are sometimes interested in the degree of "early conservatism" that would ensure that the information on secondary endpoints is not compromised unless evidence about strong effects on the primary

endpoint is overwhelming. Because there is an infinite number of designs within the unified family, it is typically the role of the statistician to help distinguish the good and bad operating characteristics in order to balance the logistical, financial, ethical, scientific, and regulatory goals.

2.3.1 Degree of Early Conservatism: Holding Power Fixed

The degree of early conservatism is an important design property to consider. We generally want to be conservative in choosing our efficacy boundaries at early analyses to protect against random high bias when there is less information to judge whether the treatment is beneficial. When the treatment may exhibit some lack of benefit or indicate some harm, we may want to be anti-conservative in choosing our futility boundaries at early analyses in order to protect patients on the trial.

The simplest way to comparing designs is to hold power fixed for the alternative of interest and forcing an interim analysis to be made at 50% of the maximum sample size. The different monitoring boundaries in Figure 2.1 presents differential behavior of early conservatism of alternative GSD when compared to a FSD having sample size 1.0. The OBF boundary is most conservative relative to all other displayed boundaries as illustrated by the relatively extreme critical values for the estimated treatment effect that would be required in order to stop the trial at the first interim analysis. A key property of this one-sided OBF rule is that halfway through the study, i.e., at 50% of the maximum sample size, the trial would have excluded any value of $\theta \leq 0$. The Pocock boundary is less conservative than the OBF boundary, as seen by the less extreme critical values relative to that of the OBF, thus making it more easy to stop at the first analysis. An asymmetric boundary can however be chosen to take advantage of the early conservatism of OBF ($P = 1$) and the less conservative Pocock ($P = 0.5$) as the efficacy and futility boundary respectively (represented notationally by $P = c(0.5, 1)$ in Figure 2.2). Such a choice significantly reduces the maximal sample size compared to using a Pocock boundary but still ensure the high degree of early conservatism of an OBF monitoring boundary. We can flexibly vary the degree of early conservatism

for the futility boundary by choosing $P = 0.8$ which has a less aggressive futility boundary relative to a Pocock futility boundary but less conservative relative to an OBF futility. This is illustrated by $P = c(0.8, 1)$.

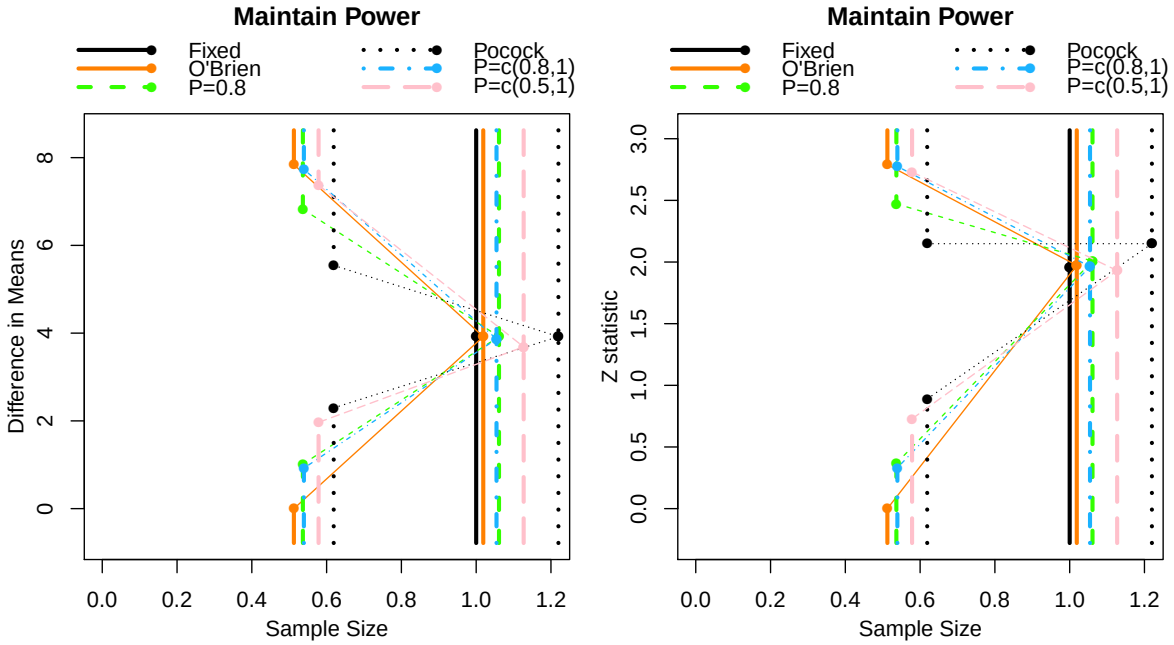

Figure 2.1: One-sided symmetric sequential boundaries [Emerson and Fleming, 1989] presented on the sample mean scale and Z scale. All boundaries illustrated have power of 97.5% at alternative $\theta = 7.84$. $P = c(0.5, 1)$ corresponds to the hybrid design with a Pocock futility and OBF efficacy boundary. $P = c(0.8, 1)$ corresponds to the hybrid design using a $P = 0.8$ as the futility boundary that is intermediate between $P \in (0.5, 1)$, and OBF efficacy boundary.

Holding the overall power constant for a hypothesized alternative no longer gives us the same maximum sample size. As such, even though the interim analyses are conducted at 50% of the way relative to this maximum sample size, the sample size at which the first interim analysis is conducted is no longer held constant for the different monitoring boundaries. Relative to a fixed sample design where a unit sample size is required with 97.5% power to discriminate the alternative of $\theta_{\text{Alt}} = 7.84$, the OBF rule results in a mild inflation of this

maximal sample size while the Pocock requires $> 20\%$ inflation. Generally speaking, for a symmetric design, as the level of early conservatism decreases, i.e., P decreases from ∞ (i.e., FSD) towards 0, the required maximum sample size to maintain power also increases.

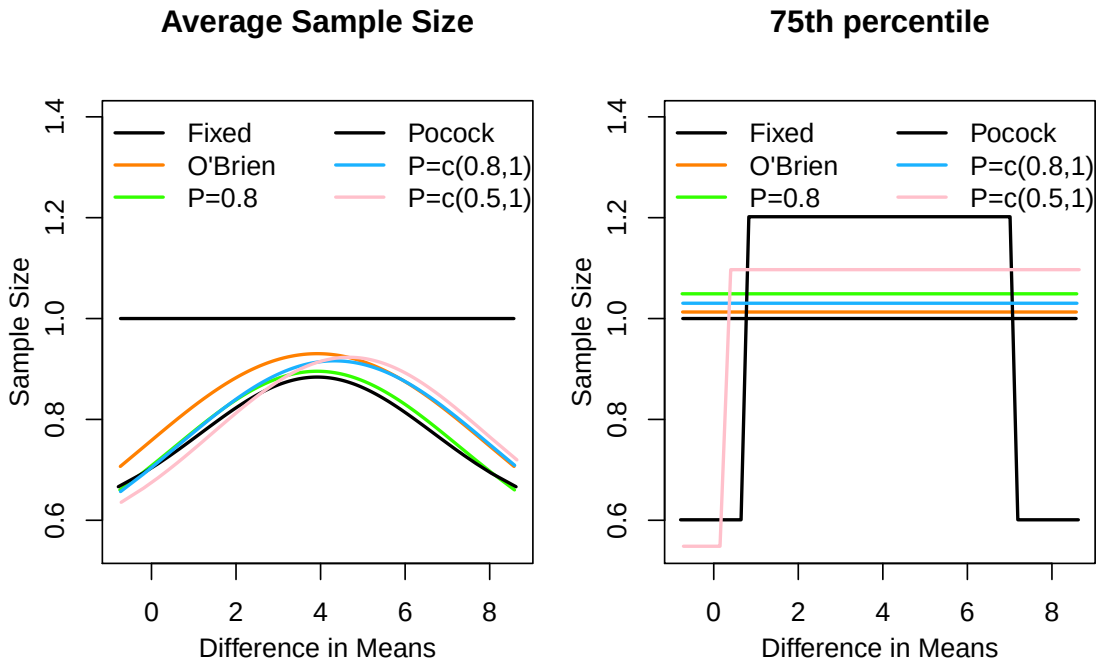

Figure 2.2: Average and 75$^{\text{th}}$ percentile of the sample size distribution for the various designs all having power of 97.5% at the design alternative $\theta = 7.84$ across different θ's between the null and alternative.

The different levels of early conservatism from these monitoring boundaries produce a spectrum of average ASN curves that are maximized at different values of potential alternatives. Among the symmetric boundaries, the less conservative Pocock boundary produced minimum ASN among values ranging between the null and alternative, while the OBF boundary averaged a slightly higher ASN relative to Pocock but a smaller ASN relative to the fixed sample design. The asymmetric rule with $P = c(0.5, 1.0)$ for futility and efficacy might better provide efficiency in discarding ineffective treatments, while having desirable

early conservatism that allows more detailed data on safety and secondary endpoints for effective therapies. The curves for the 75^{th} percentile of the sample size distribution also exhibit different behavior for the various monitoring boundaries across intermediate values of θ_0 and θ_{Alt}.

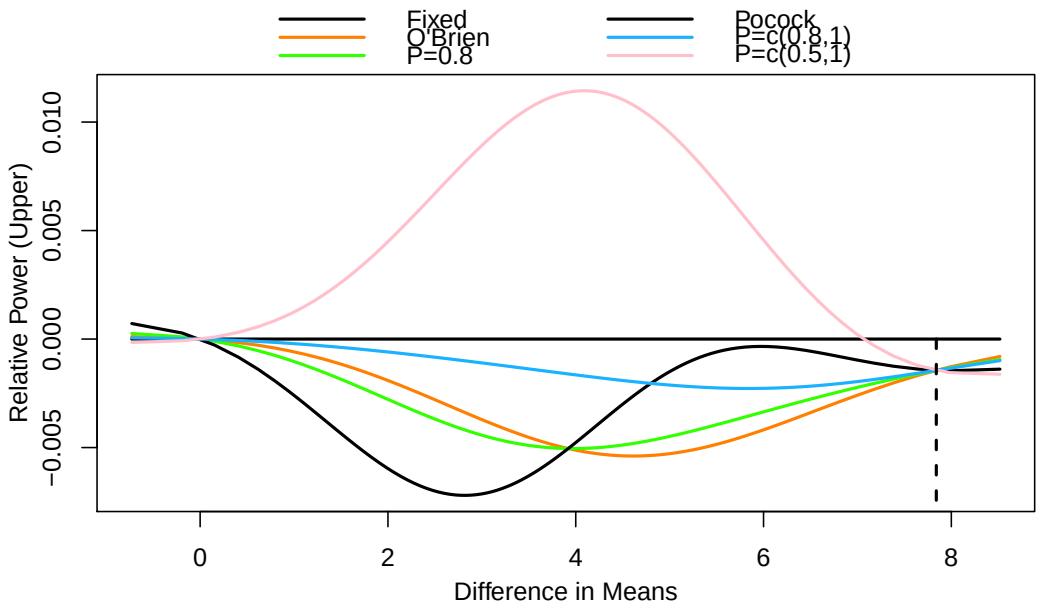

Figure 2.3: Difference in power relative to fixed sample design having the same maximal sample size as the OBF for the various designs used in Figure 2.1 with power of 97.5% at alternative $\theta = 7.84$ across different θ's between the null and alternative.

We can also compare the various GSDs with respect to their power curves. Figure 2.3 shows the difference in power curve relative to the FSD with the same maximum sample size as the OBF. Such a FSD is included in this comparison to enable us to assess the "cost" of including an interim analysis using an OBF design without increasing the maximal sample size. Because the OBF design is the most conservative GSD considered in Figure 2.3, the FSD provides greater power than all the other designs in this graph. It should be noted that all the GSDs were chosen to have an overall Type 1 error of 0.025 and power 0.975 at

the same design alternative. Hence, the only differences among the various GSDs are how the power curves differ at other alternatives. At other alternatives between the null and design alternative, the difference in power is negligible. The choice of the symmetric Pocock boundary ensures higher power at alternatives between half the hypothesized alternative to the alternative relative to all other monitoring boundaries presented. The OBF, on the other hand, beats the other monitoring boundaries between the null and half of the hypothesized alternative.

In summary, when holding power fixed, the unified family which consists of a spectrum of monitoring boundaries can have different operating characteristics. For "conservative early" designs like the OBF, the differences in the power curves will not differ substantially from a FSD with the same maximal sample size. However, for less conservative (but in the case of the Pocock more efficient on average) designs, there are more variations between the power curves relative to the OBF or FSD design.

2.3.2 Effect of Adding Interim Analyses: Holding Power Fixed

Often, the number of interim analyses affects the logistical and financial cost of the trial. We describe the impact of additional analyses with the use of the OBF monitoring rule while holding power fixed under the same alternative. With additional analyses, the maximum sample size is increased. The monitoring boundaries are interpolated more finely across interim analysis but remain approximately similar (Figure 2.4).

Additional interim analyses also reduce the ASN curves uniformly such that designs with fewer analyses averaged higher ASN curves across $\theta \in (\theta_0, \theta_{\text{Alt}})$ (Figure 2.5). The curves for the 75^{th} quantile of the sample size distribution are dominated differently depending on how close the true θ is to the alternative and the null.

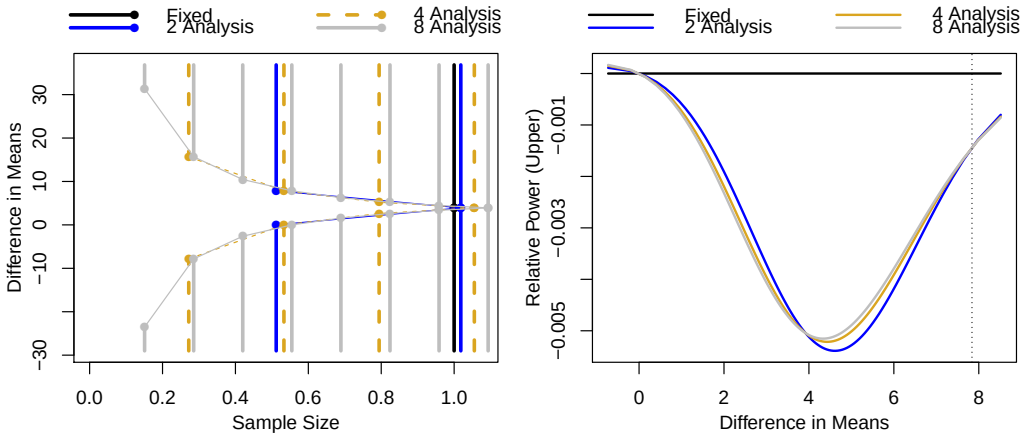

Figure 2.4: Sequential boundaries on the sample mean scale for the one-sided symmetric OBF design with different number of interim analyses and power fixed at 97.5% at alternative $\theta = 7.84$. (Right) Difference in power, relative to fixed sample design with the same maximal sample size for the two-stage OBF design, are shown for designs with different number of interim analyses while keeping power at 97.5% under design alternative $\theta = 7.84$.

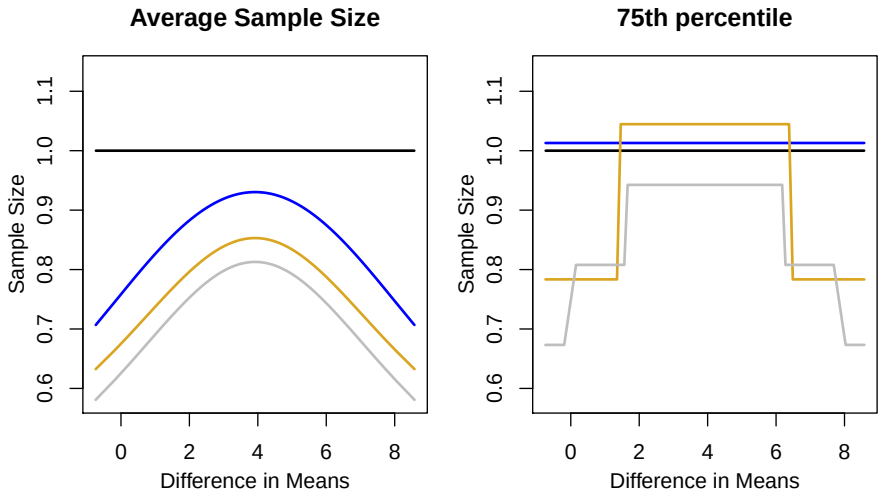

Figure 2.5: Average and 75$^{\text{th}}$ percentile of the sample size distribution for the various one-sided symmetric OBF designs with different number of interim analyses while keeping power at 97.5% under design alternative $\theta = 7.84$. The ASN curves in blue, gold, and gray correspond to 2, 4, and 8 analyses respectively.

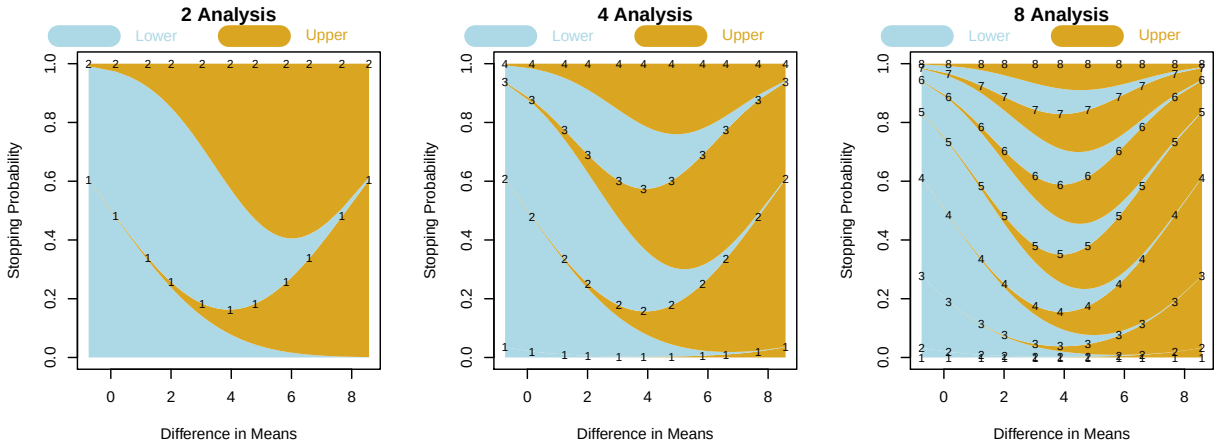

Figure 2.6: Probability of stopping for efficacy or futility for the various OBF designs with different number of interim analyses while keeping power at 97.5% under alternative $\theta = 7.84$. In gold, the region describes the probability of rejecting $\mathbb{H}_0$ for efficacy, while in light blue, the region describes the probability of rejecting $\mathbb{H}_0$ for futility/non-efficacy.

Figure 2.6 describes the degree of early conservatism as shown by the monitoring boundaries on the probability scale often interpreted as the probability of early stopping which ties back indirectly to the computation of the ASN curves. This figure can be interpreted as follows: Consider the leftmost sub-figure, the light blue colored region corresponds to the probability of rejecting $\mathbb{H}_0$ for futility/non-efficacy while the golden region corresponds to the probability of rejecting $\mathbb{H}_0$ for efficacy. The line indicated by the number k represents the cumulative stopping probability (sum of the light blue and gold region) at the k^{th} analysis across the different design alternatives ranging from $\theta = 0$ to $\theta_{\text{Alt}} = 7.84$.

Under the design alternative $\theta_{\text{Alt}} = 7.84$, when there are only 2 interim analyses, this cumulative probability of stopping by the first interim analysis is 50%. As we increase the number of interim analyses, this cumulative probability of early stopping at θ_{Alt} is subdivided and discretized by the interim analyses to allow more opportunities to stop even earlier when sufficient and statistically credible evidence has been established regarding the futility/non-efficacy or efficacy of the treatment/prevention strategy. This property of the

GSD is key when planning a prevention trial with the possibility of incorrect assumption of the background rates potentially coupled with extreme treatment efficacy in Chapter 5.

2.3.3 Effect of Adding Interim Analyses: Holding Maximum Statistical Information Fixed

In general, when holding the maximum statistical information fixed, the consequence of additional interim analyses at design stage will result in some loss of overall power. Depending on the choice of the monitoring procedure, this loss of power may be minimal. We consider the OBF and the Pocock boundary as illustrations (Figure 2.7 and 2.8). For this illustration, we hold fixed the maximum sample size for the same hypothesized θ_{Alt} while increasing the number of equally spaced interim analyses.

There are minor differences in the OBF monitoring boundaries, with slight loss of statistical power under the design alternative, as well as across the various θ's between the null and alternative. The critical value at the final analysis for the Pocock monitoring rule is more extreme as we increase the number of interim analyses. However, the consequence of holding the maximum sample size fixed and increasing the number of interim analyses for the Pocock boundary also changes the relative behavior of the power curves vastly for intermediate values of θ between the null and alternative.

The ASN curves generally decrease when interim analyses are added as seen in Figure 2.8. The curves for the 75% quantile of the sample size distribution are not necessary dominated at all $\theta \in (\theta_0, \theta_{\text{Alt}})$ with the addition of interim analyses.

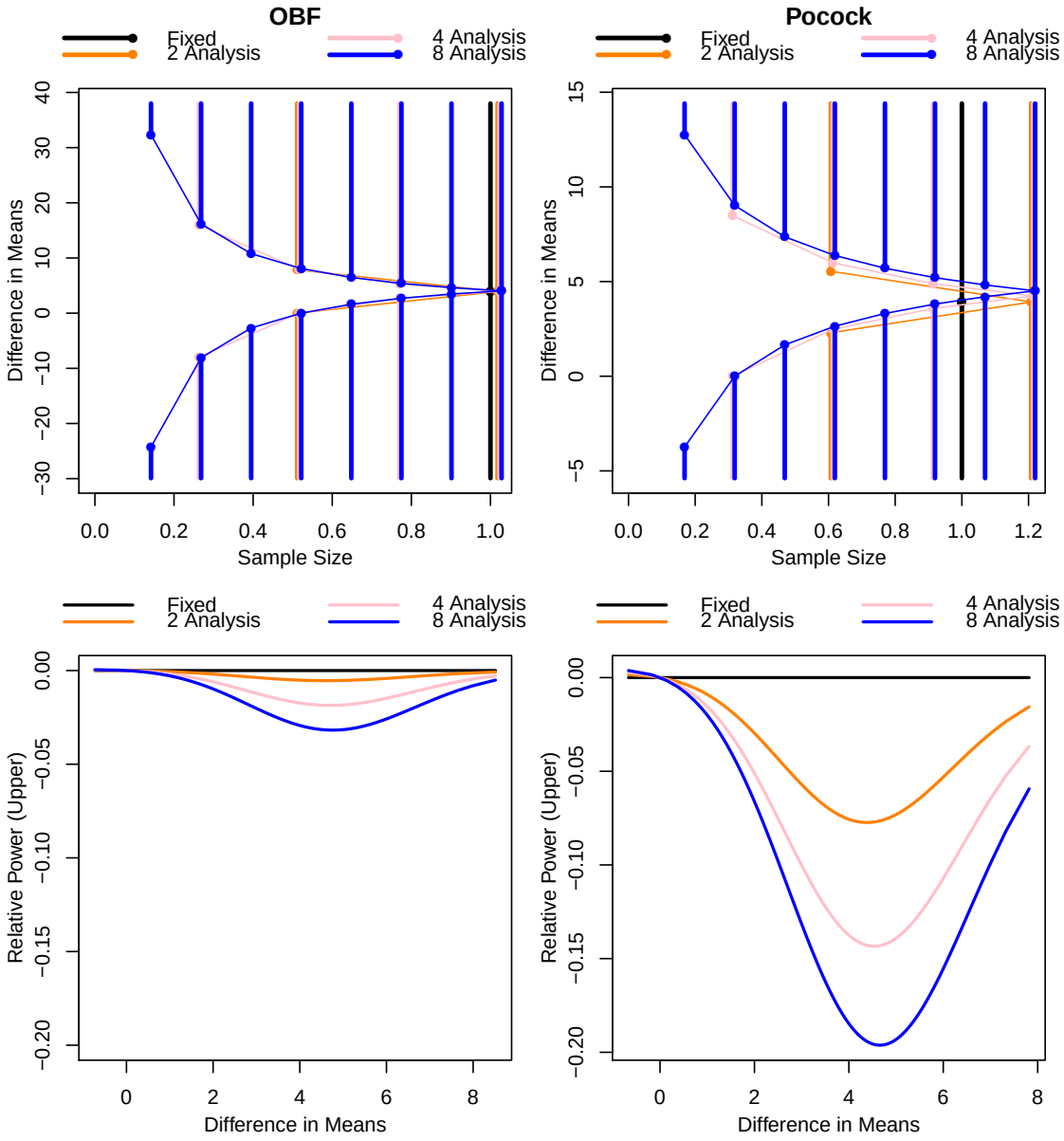

Figure 2.7: Monitoring boundaries and differences in power relative to the fixed sample design for O'Brien Fleming and Pocock designs when increasing the number of interim analyses while maintaining the statistical information under the same alternative $\theta = 7.84$. The monitoring boundaries are slightly different for Pocock and generally show bigger loss of power when holding maximum statistical information fixed but increasing the number of the interim analysis relative to the OBF class of designs.

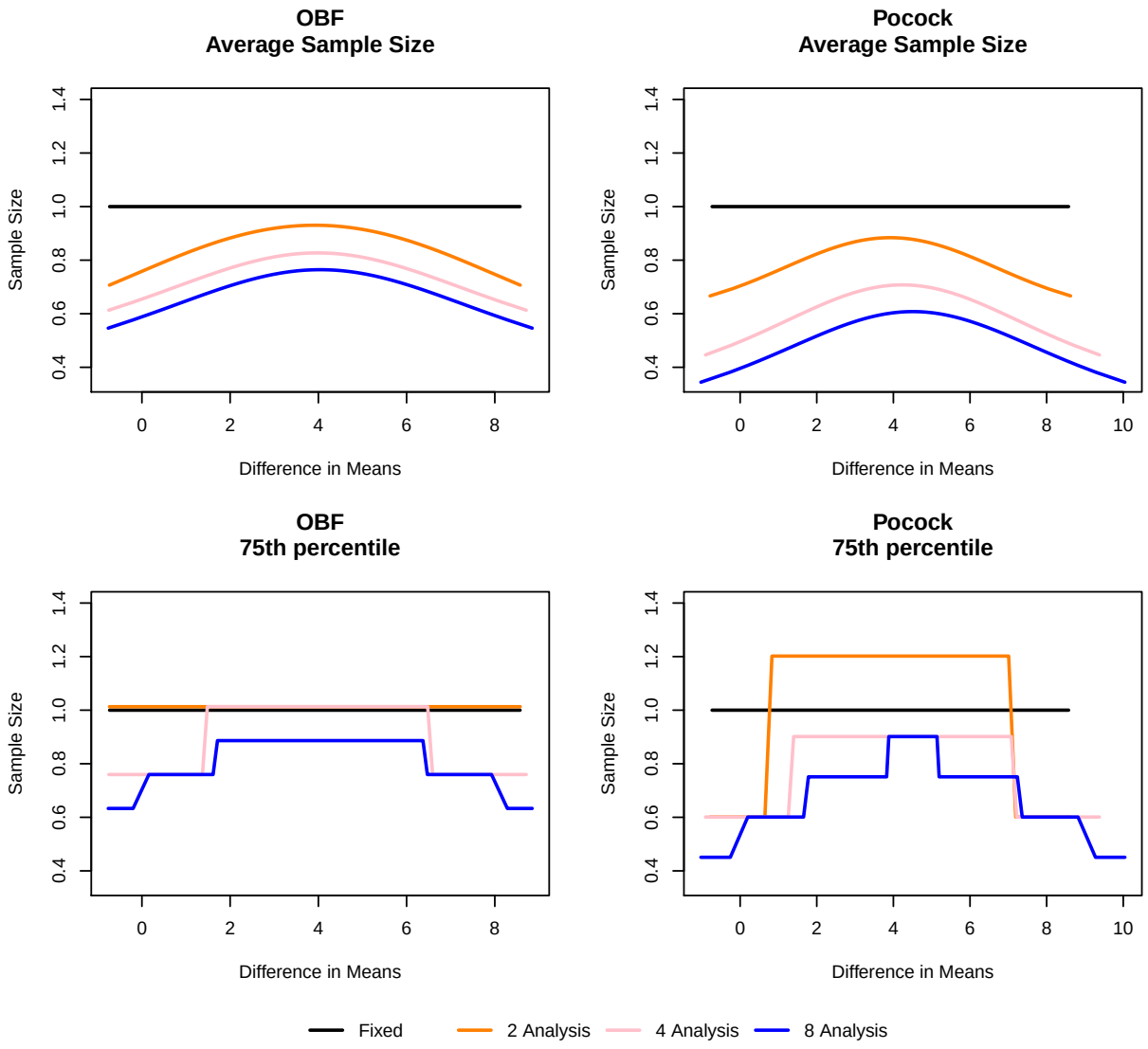

Figure 2.8: Average and 75$^{\text{th}}$ percentile of the sample size distribution for the O'Brien Fleming and Pocock design are shown as we increase the number of interim analyses while maintaining statistical information for the same alternative $\theta = 7.84$. The fixed sample design has both the maximum and average sample size at 1.

2.3.4 Summary

In summary, when holding power fixed, common operating characteristics of the group sequential designs such as maximum statistical information, ASN, 75^{th} percentile of the sample size, and the probability of early stopping vary depending on the degree of early conservatism of the monitoring rules. When adding interim analysis (holding power fixed)

- ASN decreases
- Maximum statistical information increases
- Probability of stopping at earlier interim analysis increases

We rarely choose the optimal design based purely from a statistical standpoint. In practice, the choice of the monitoring procedure is based on science, logistics, or even ethics which may be less tangible to parametrize. However, the large class of designs that satisfies at least general concerns such as the overall Type 1 error, power for a specific alternative, or the degree of early conservatism is sufficient to facilitate picking operating characteristics that balance science, ethics, logistics, etc. In this over parameterized space, we see the difficulty of choosing a single best design since slight modifications may no longer allow us to fairly compare another design of choice.

2.4 Blinded Sample Size Revision: Information Based Approach

There are many methods that have been proposed to allow for sample size revisions to maintain statistical power [Wittes and Brittain, 1990, Gould and Shih, 1992, 1998]. However, it is vital that when sample size revision is performed during interim analysis, the integrity of the trial not be compromised as a consequence of unwittingly unblinding the treatment groups. Thus, blinded procedures that do not disclose information about the treatment groups assignments are typically preferred and regarded as "well-understood" procedures by FDA. For these reasons, it is generally preferred that blinded revisions not be performed by a DMC that has seen unblinded interim results. We describe interpretations of the sample

size formula that give rise to some of these procedures. Later, we describe procedures that choose to unblind the study to repower the trial to a different alternative (section 2.6).

2.4.1 Notation

In the fixed sample setting, we are interested in discriminating the null hypothesis of $\mathbb{H}_0 : \theta \leq \theta_0$ vs the alternative $\mathbb{H}_A : \theta \geq \theta_{\text{Alt}}$ in a hypothesis test with an overall 1-sided Type 1 error of α and power β. The general sample size formula may be written as $N = \frac{\delta^2_{\alpha\beta} V}{\Delta^2}$ where $\delta_{\alpha\beta} = (z_{1-\alpha} + z_\beta)$ and $\Delta = \theta_A - \theta_0$ [Emerson, 2000]. We refer to $\delta_{\alpha\beta}$ as the standardized alternative under the fixed sample design having maximal statistical information of 1.0. This can be extended to group sequential design by making some multiplicative adjustment such that $\delta_{\alpha\beta} R$ is the maximum sample size for the group sequential design where R is this multiplicative adjustment described in Jennison and Turnbull [1999].

In the general setting, V may be a function of the summary measure, θ, as well as other parameters that may be independent of the summary measure θ. Many times, V may not be precisely known at the design stage and is instead chosen based on a reasonable guess from prior studies. Because an incorrect specification of V can often lead to underpowered or overpowered studies, it is sometimes written into the protocol that the sample size will be revised when a more precise estimate of V is obtained.

Under the immediate outcome setting used as our example, V is equivalent to $2\sigma^2$ based on a 1:1 randomization. Equations 2.1 & 2.2 present alternative formulations of the usual sample size formula.

$$\text{Maintain sample size:} \quad N = \frac{(z_{1-\alpha} + z_\beta)^2}{\frac{(\theta_A - \theta_0)^2}{V}} \tag{2.1}$$

$$\text{Maintain statistical information:} \quad \frac{N}{V} = \frac{(z_{1-\alpha} + z_\beta)^2}{(\theta_A - \theta_0)^2} \tag{2.2}$$

Equation 2.1 can provide an interpretation of our RCT design where we choose to maintain the sample size despite observing variability $\widehat{V}$ different than that used when planning

the RCT. When the observed $\widehat{V}$ during the trial differs from our guess V, our test statistic would not attain the same power β to detect the designed difference Δ. Instead, when evaluating power under various alternatives, we are comparing power curves based on alternatives that are measured in terms of "standard deviation", i.e., $\Delta/\sqrt{V}$. At the planning phase, sensitivity analyses may be desired under different assumed variability V [Emerson, 2000]. For example, when our sample size is 1 unit, V is assumed to be 1 at design stage with level $\alpha = 0.025$, power $\beta = 0.975$, and $\Delta/\sqrt{V} = 3.9199$. By holding our sample size constant, i.e., at 1 unit, when V is truly 1.25, holding α and β fixed, we can only discriminate our hypothesis when $\Delta/\sqrt{V} = 3.919928/\sqrt{1.25} = 3.50609$. Alternatively, one may interpret having a lower power ($\beta^* = 93.89\%$) if one chooses to further hold Δ constant.

Equation 2.2 illustrates an interpretation for the blinded revision of sample size strategy where one chooses to maintain a possibly prespecified maximal statistical information N/V. By keeping our statistical information fixed, the right hand side can be interpreted as choosing to maintain our level α and power β to discriminate our original hypothesis Δ. Hence, when V is incorrect, we can modify our sample size N to approximately hold the right hand side constant, while maintaining the same power to discriminate Δ. Because modification of the sample size is conditional upon this observed variability $\widehat{V}$ during the course of the trial, as long as the treatment effect is blinded, we consider $\widehat{V}$ to be ancillary and thus asymptotically independent of the estimated treatment effect.

Whitehead et al. [2001] investigated the use of this strategy to maintain statistical information for GSD without unblinding. Other authors such as Mehta and Tsiatis [2001] described the general approach of using information based monitoring but did not fully specify how to keep the design blinded at interim analysis to revise estimates of the nuisance parameters in their examples. Other monitoring approaches to revise boundaries based on accrued statistical information by Lan and DeMets [1983] or Burington and Emerson [2003] are similarly generally considered "well-understood" within the context of GSD. In Appendix B, we present such an example of blinded adaptation and its use during the course of monitoring. This general approach will later be relevant to the results presented in Chapter 5.

In the setting when our responses are continuous, the use of the overall variance, σ^2, is typically preferred to maintain blinding. [Whitehead, 1997] advised against the use of the pooled variance which can result in breaking the blind of the treatment effect. In particular, when our primary outcome is the difference in probability, we can use $\widehat{p} = \sum_{i=1}^{N} \sum_{w=0,1} X_{iw}/(2N) = (Rp_1 + p_0)/(R+1)$ to solve for the event rate under the null where R is the randomization ratio, and $w = 0, 1$ to denote treatment group 0 and 1 respectively. By plugging in $p_0 = p_1 + \theta$, we can obtain $\widehat{p} = p_0 - R\theta/(R+1)$ to compute the variance. Table 2.1 displays a summary of estimators that are used to perform blinded adaptations in more general settings.

Table 2.1: Test statistics provided when sample size review is performed in the general setting.

Setting	Blinded	Assumption
Difference in proportions	$\overline{p}$	$p_0 = p_1 + \theta$
Difference in means	$\sigma^2 = \sum_{i,w}(X_{iw} - \overline{X})^2$	$\theta_0 = \theta_1 + \theta$
Log-rank analysis	Number of events	$\log(\theta_{\text{Hazard Ratio}})$

The use of blinded procedures can adequately address incorrect assumptions at design stage that are applied during the sample size calculation. As per FDA guidance documents [FDA, 2010, 2015], blinded procedures that do not unblind the study generally do not introduce bias. It is also recommended that these adjustments be made as late as possible when a more precise estimate of the baseline rate is obtained. While it is not recommended that these adjustments be used to decrease sample size, blinded procedures may be considered in conjunction with group sequential methods to possibly revise the sample size for early stopping. There are however times whereby trials are not adequately designed with the appropriate statistical power, or lack a thorough evaluation of all assumptions during planning stage that lead to more attractive options to "repower" aspects of the trial design based on unblinded interim analyses. These form the basic arguments for most of the statistical development in designing more flexible, adaptive trials.

2.5 Prespecified Adaptive Designs

Group sequential monitoring was first used to overcome many of the ethical and efficiency concerns present in a fixed sample design. Bauer and Köhne [1994] introduced the idea of added flexibility for modifications to mid-trial designs in confirmatory setting. The concept of "self-designing" trial envisioned by Fisher [1998] considered adaptations to possibly (1) drop treatment arms, (2) re-power studies for unanticipated differences in treatment effect, (3) modify the definition of primary outcome based on interim data, (4) modify eligibility criteria, and/or (5) modify randomization ratios.

Group sequential monitoring can adequately address most statistical concerns in (1) and (2). Other adaptations in 3-5 are more difficult to interpret and less acceptable as they involve complex modification of the scientific hypotheses of interest [Fleming, 2006, Emerson, 2006, Emerson and Fleming, 2010]. Nonetheless, there are differences between the newer adaptive designs and the classical GSD that might provide some additional advantage in efficiency of clinical testing of new treatments.

It is possible to pre-specify an adaptive sampling scheme in a manner to allow inference based on minimum sufficient statistics [Levin, 2013]. With a pre-specified adaptive sampling scheme, a known sampling distribution can be used to perform frequentist inference, thus enabling evaluation of operating characteristics. This notion is similar to the class of "sequentially planned decision procedures" [Schmitz, 1993]. Since GSD is a special class of pre-specified adaptive design, choosing adaptive rules based on minimum sufficient statistics among this larger class of designs should theoretically allow us to be more efficient so long we pre-define the opportunities to make changes to the sampling plan that increase the sample size [Jennison and Turnbull, 2006a]. This can more appropriately allow us to investigate the degree of efficiency gain with judicious choice of sampling plan based on specific optimality criteria that balance scientific and ethical constraints.

We introduce notation for the prespecified adaptive design, which can be defined by a sequence of group sequential designs with different maximum statistical information. Be-

cause these adaptive rules are prespecified, the full sampling distribution can be written down, enabling frequentist inference to be computed in a manner analogous to that used for frequentist inference in GSD. For purposes of this dissertation, we shall be concerned with statistical adaptations in the time to event setting. However, such statistical adaptations become complex when they inherently affect the scientific interpretation of the results.

2.5.1 Notation

Using the notation in section 2.2.1, we consider a single adaptive interim analysis is made based on the GSD. We summarize the necessary notation for this setting based on Levin et al. [2014]. We apply the same set up and will generally be concerned with the following test statistics, namely: the partial sum statistic, $S_j = \sum_{i=1}^{N_j}(X_{Ai} - X_{Bi})$; the MLE estimate, $\widehat{\theta} = S_j/N_j$; the normalized Z statistic/Wald, $Z_j = \sqrt{N_j}(\widehat{\theta} - \theta_0)/\sqrt{2\sigma^2}$; the fixed sample P_j-value statistic, $1 - \Phi(Z_j)$. In addition, our incremental statistics are $S_j^* = \sum_{i=N_{j-1}^*+1}^{N_j^*}(X_{Ai} - X_{Bi})$, $\widehat{\theta}_j^* = S_j^*/N_j^*$, $Z_j^* = \sqrt{N_j^*}(\widehat{\theta}^* - \theta_0)/\sqrt{2\sigma^2}$, and $P_j^* = 1 - \Phi(Z_j^*)$.

Following Levin et al. [2013], at the adaptive interim analysis h such that $\{h : 1 \leq h < J\}$, future incremental sample sizes are modified as follows: We partition the continuation region $\mathcal{C}_h$ at the adaptive interim analysis h into r mutually exclusive continuation sets, denoted by $\mathcal{C}_h^k$ for $k = 1, \cdots, r$ where $\mathcal{C}_h^{k'} \cap \mathcal{C}_h^k = \varnothing$ if $k = k'$, and $\cup_{k=1}^r \mathcal{C}_h^k = \mathcal{C}_h$. Let each continuation set, $\mathcal{C}_h^k$, which is made at the adaptive analysis, to correspond to a future group sequential path k with a maximum of J^k interim analyses. Then, each continuation region $\mathcal{C}_{h+1}^k, \cdots, \mathcal{C}_{J^k}^k$ will correspond respectively to some potential future sample size $n_{h+1}^k, \cdots, n_{J^k}^k$. Also, let the continuation set $\mathcal{C}_h^k$ at the adaptive analysis to correspond to a symmetric continuation region. We can define the random sample path variable K for values $0, 1, \cdots, r$ made after the adaptive interim analysis. We now have a three dimensional statistic (J, S, K) where J is the stage, S is the partial sum statistic calculated at the stopping, and K to be the sequential path that led to the stopping. This statistic has been shown to be minimal sufficient by Levin [2013]. For notational convenience, we shall suppress the superscript for J.

Consider Figure 2.9 as an illustration. In this design, we have a two stage ($J = 2$) pre-

specified adaptive design where an adaptive interim analysis is conducted at the first interim analysis ($h = 1$), with $K = 3$ different maximum statistical information n_2^1, n_2^2, and n_2^3. This two stage design has $r = 3$ mutually exclusive continuation regions $\mathcal{C}_1 = \cup_{k=1}^{3} \mathcal{C}_1^k$ such that each continuation region can be described as a group sequential design. More explicitly, one such two-stage group sequential design has an interim analysis conducted at n_1 subjects with continuation region $\mathcal{C}_1 = (a_1, d_1)$ leading to a maximum statistical information n_2^1 as described by the blue lines. Similarly, another two-stage group sequential design has an interim analysis conducted at n_1 but a continuation region $\mathcal{C}_3 = (a_3, d_3)$ that leads to a much bigger maximum statistical information n_2^3 relative to n_2^1 or n_2^2 as described by the yellow lines.

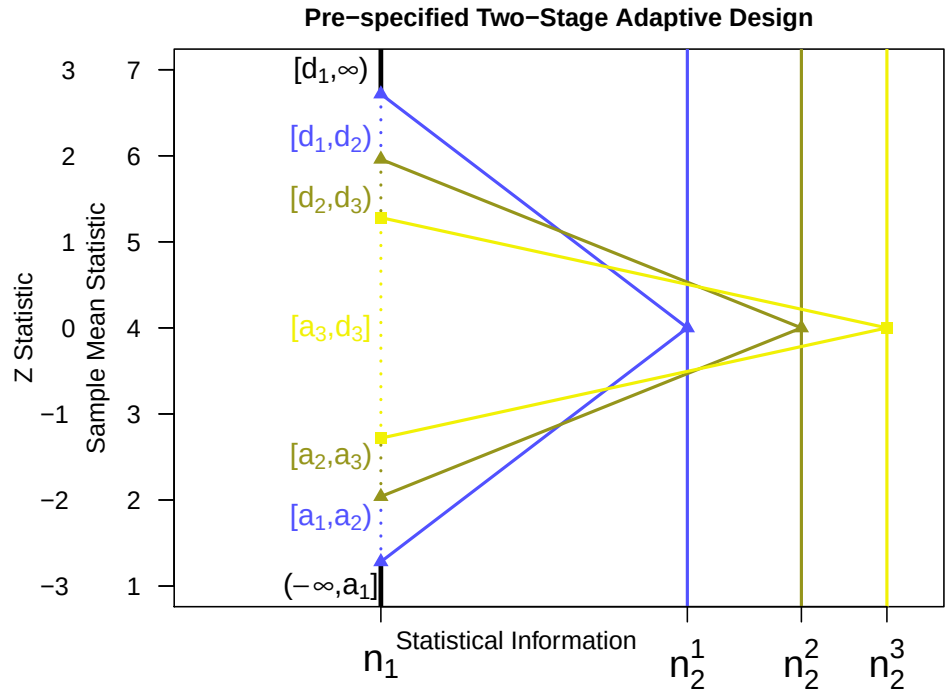

Figure 2.9: A pre-specified adaptive design with a single adaptive interim analysis where the different continuation regions correspond to different final statistical information. The above can be described as a series of group sequential design with different continuation regions or more formally as a stochastic hybrid of group sequential designs.

Such a prespecified adaptive design as shown in Figure 2.9 can otherwise be described as a *stochastic hybrid of group sequential designs* that *randomly* switches between the three different maximum statistical information $n_2^1, n_2^2,$ and n_2^3, comprising of the respective continuation sets $\mathcal{C}_1^k = (a_k, a_{k+1}) \cup (d_{k+1}, d_k)$ for $k = 1, 2$, and $\mathcal{C}_1^3 = (a_3, d_3)$. It is this random switching of group sequential designs with different maximum statistical information that allows the sampling density to be characterized.

The sampling density for the class of prespecified adaptive designs can be considered as an extension of the notation described for the sampling density of the GSD in section 2.2.1. Frequentist operating characteristics can be obtained with the extension of the GSD sampling density. We refer the reader to Levin [2013] for more details on the sampling density, characterization of minimum sufficiency, as well as the evaluation of inferential procedures.

2.6 (Fully) Adaptive Designs

Adaptive features made on the basis of unblinded data can not only inflate the overall Type 1 error through sequential testing but can further introduce bias from sponsors/researchers who have potential conflict of interest. This possibility may raise concerns about the validity of the trial results. Since there is a regulatory need for "adequate and well-controlled" investigations (Kefauver-Harris, 1962) in confirmatory studies to enable appropriate labeling of any new treatment, adaptive designs that were not fully specified ("fully adaptive") received less regulatory support from both FDA and EMA. Recent guidance documents from EMA [2007] and FDA [2010, 2015] provided substantial discussions on what aspects of adaptation are deemed acceptable, less acceptable, or outside the scope of discussion in the context of major confirmatory Phase III settings. Despite minor differences between each agencies' definition of adaptive designs, both agencies recognize the need for adequate and well-controlled confirmatory trials, and that appropriately planned and well-executed adaptive designs in earlier phase studies could lead to improvements in the drug discovery process.

Regulatory agencies are concerned with unplanned adaptations for reasons besides scientific interpretability of results. One of the biggest issue with such procedures was first

examined by Proschan et al. [1992] and elaborated in more detail in Proschan and Hunsberger [1995]. In their two stage setting, a sample size modification is made based on the interim estimate of the treatment effect using the function $n_2^* = h(z_1, n_1)$ where z_1, n_1, n_2^* are the interim stage one z statistic, first stage sample size, and incremental second stage sample size respectively. As might be expected, the authors demonstrated that failure to adjust for the multiplicity of analyses results in an inflation of the overall Type 1 error. However, the surprising result was that the overall Type 1 error can more than double and the usually conservative Bonferroni correction assuming two analyses fails to protect us against such inflation. At first glance, the results appear paradoxical. However, knowing the unblinded interim results, one can effectively avoid actually performing analyses unlikely to achieve statistical significance, and try to hone in on sample sizes that maximize the chance of achieving statistical significance. In effect, then, a user has considered many more analyses than the two that were actually performed, and a Bonferroni correction based on only two analyses would be incorrect.

In the following sections, we summarize the proposed approaches for controlling the overall Type 1 error when modifying the sample size using unblinded interim results in the adaptive setting. These approaches can be classified into (1) combination function, (2) conditional error, (3) re-weighting of test statistic approaches.

2.6.1 Combination Approaches

Bauer and Köhne [1994] proposed the combination approach via Fisher's method of meta-analysis. Incremental p values obtained independently from different stages of the discovery pipeline are combined via a pre-specified function $h(p_1^*, p_2^*)$. Under $\mathbb{H}_0 : \theta = 0$, $p \sim U(0, 1)$. By distribution theory, we know that $-2 \log p \sim \chi_2^2$. This thus defines a rejection region R such that $\alpha = \Pr[h(p_1^*, p_2^*) \in R | \mathbb{H}_0] = \int \int_R h(p_1^*, p_2^*)\, dP_1^* dP_2^*$.

In a two stage setting, Fisher's criterion allows one to combine incremental p_1^*, p_2^* values obtained from results based on two independent stages, with the rejection region defined as $p_1^* p_2^* \leq c_\alpha = \exp[-\frac{1}{2}\chi_4^2(1-\alpha)]$. At stage one, reject $\mathbb{H}_0$ for efficacy if $p_1^* \leq \alpha_1$, or reject $\mathbb{H}_0$

for futility if $p_1^* \geq \alpha_0$. If not, continue to stage two. At stage two, reject $\mathbb{H}_0$ for efficacy if $p_1^* p_2^* \leq c_\alpha$ where α_0, α_1 is solved by

$$\Pr[p_1^* \leq \alpha_1] + \Pr[(p_1^* \in (\alpha_1, \alpha_0)) \cap (p_1^* p_2^* \leq c_\alpha)] = \alpha.$$

They evaluated the loss of power by comparing their approach to the UMP test of the design based on the total sample size obtained separately from the two stages. Additionally, their numerical evaluation considered the degree of this loss of power based on different fractions of this total sample size. Their evaluation has several limitations. The original sampling scheme for the two stage design is unknown and based upon the sample size of $n_1 + n_2^*$ that is regarded to be the true design. Naïvely, this second stage is always conducted regardless of the results obtained from the first stage. Thus, one may consider this as expanding some FSD with an unknown maximum sample size based on an administrative look.

A similar approach to combining incremental p^* values was also investigated by Lehmacher and Wassmer [1999] where they proposed using pre-specified weights to combining these p values obtained from independent stages. Under the null hypothesis, these p^* values can be combined via $Z = \omega_1 \Phi^{-1}(1 - p_1^*) + \omega_2 \Phi^{-1}(1 - p_2^*)$ with $\omega_1^2 + \omega_2^2 = 1$. Such approach has an intuitive and linear mapping of the weight function on the Z scale and differs from the combination approach of Bauer and Köhne [1994] that has a non-linear mapping of the weights. We note that when sample sizes for the second stage are dependent on the first stage results, these p^* are no longer independent under the alternatives.

2.6.2 Conditional Error Approaches

The conditional error approach was first suggested by Proschan and Hunsberger [1995] where they examined the worst case setting of using unblinded treatment results to make sample size adaptations in a two stage setting. They proposed the use of pre-specifying a conditional error function $A(z_1^*)$ to preserve the overall type I error where $A(z_1^*) = \Pr_{\mathbf{H}_0}[Z_2 > k | Z_1^* =$

$z_1^*, n_1)] \in [0, 1]$. $A(z_1^*)$ is defined as the probability of incorrectly rejecting the null hypothesis conditional on some observed interim statistic $Z_1 = z_1^*$. Proschan and Hunsberger proposed several forms of conditional error functions. The linear $A(z_1^*) = \Phi(az_1^* + b)$ and the circular function $A(z_1^*) = 1 - \Phi(\sqrt{h^2 - z_1^{*2}})$, where constants h, or a and b are chosen to maintain the α at some desired level, are more popular in the adaptive literature.

Müller and Schäfer [2001] and Müller and Schäfer [2004] described the more general procedure of using GSDs to make sample size adaptations by preserving the conditional error. This generalized procedure allow the use of unplanned modifications in the context of GSDs with multiple interim analysis by preserving the conditional Type 1 error at the adaptive interim analysis. Their approach first conditions out the interim estimated treatment effect and re-evaluates the future modified boundaries based on the remaining conditional error. Even though a prespecified conditional weighting is established, one does not know the future course of the stopping boundary until the interim analysis. Because the sampling rule may not be prespecified in advance, there is difficulty in providing frequentist estimation after the trial stops. Despite that, Brannath et al. [2009], Gao et al. [2013] and Levin et al. [2014] have proposed methods on evaluating the adjusted point estimates and confidence intervals accounting for unplanned adaptations.

2.6.3 Re-weighting of Test Statistic

Methods proposed by Fisher [1998], Shen and Fisher [1999], Cui et al. [1999], Schäfer and Müller [2001], and Shen and Cai [2003] approached the problem by re-weighting the test statistic to control for the overall Type 1 error after making an unplanned unblinded adaptation. Consider the design with the original sample size $n = n_1^* + n_2^*$. At the first interim analysis, when n_1 subjects are accumulated, based on accumulated data summarized by Z_1, the trial is modified to increase the stage two sample size from n_2^* to $\widetilde{n}_2^*$, with $\widetilde{n}_2^* = \gamma(n_1^*)$. Although $Z_1^* \sim \mathcal{N}(0, 1)$ and $\widetilde{Z}_2^* \sim \mathcal{N}(0, 1)$ under the (incremental) null hypothesis, where $\widetilde{Z}_2^*$ is the incremental statistics based on the new n_2^* subjects, the naïve sampling distribution of $Z_2 = \frac{n_1^*}{n_1^* + \widetilde{n}_2^*} Z_1^* + \frac{\widetilde{n}_2^*}{n_1^* + \widetilde{n}_2^*} \widetilde{Z}_2^*$ based on weighting all the data equally no longer have the usual

standard normal distribution since $\widetilde{n}_2^*$ depends on Z_1^*. Thus, the naïve use of the test statistic based on the standard normal distribution will result in an inflation of overall Type 1 error.

To correct for this data adaptive look, Cui et al. [1999] (CHW) considered re-weighting the later part of the data when the sample size is increased at the adaptive interim analysis. Later, Chen et al. [2004] defined the use of "promising regions" such that when the estimated treatment effect is within some range of conditional power, it is considered promising to extend the trial by more than doubling the sample size to target a different effect size than planned. Gao et al. [2008] and Mehta and Pocock [2011] extended such notion to characterize larger regions via conditional power where the naïve Type 1 error procedure may be used even when the sample size is more than doubled. Gao et al. [2008] claimed such an approach does not "downweight data accumulated in different periods".

It is easy to see that Gao et al. [2008]'s claim is incorrect. (1) When we choose to use the naïve test statistic, we apply the re-weighting scheme by CHW; Or (2), when we re-weight our test statistic, we use the naïve critical value. Consider the J-look GSD whereby an adaptation is made at the penultimate interim analysis to increase the original statistical information from $\mathcal{I}_J$ to $\mathcal{I}_{\text{New}}$. Let the final adapted Z statistic with the naïve weighting be $Z_{\text{Naïve}} = \sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_{\text{New}}}} Z_{J-1} + \sqrt{\frac{\mathcal{I}_{\text{New}} - \mathcal{I}_{J-1}}{\mathcal{I}_{\text{New}}}} Z_J^{\text{New}^*}$ where $Z_J^{\text{New}^*}$ denotes the incremental Z statistic between the $J-1$ and J analysis.

The naïve critical value d_J based on the original GSD is no longer acceptable when an unblinded adaptation using the interim estimated treatment effect $\widehat{\theta}_{J-1}$ is used to change the design. We denote $d_J^\star$ to be the CHW critical value after such a design change is made to increase $\mathcal{I}_J$ to $\mathcal{I}_{\text{New}}$. The adjustment procedure using CHW is as below

$$Z_{\text{Naïve}} > d_J^* = \frac{1}{\sqrt{\mathcal{I}_{\text{New}}}} \left[\frac{\sqrt{\mathcal{I}_{\text{New}} - \mathcal{I}_{J-1}}}{\sqrt{\mathcal{I}_J - \mathcal{I}_{J-1}}} \left(d_J \sqrt{\mathcal{I}_J} - Z_{J-1} \sqrt{\mathcal{I}_{J-1}}) \right) + Z_{J-1} \sqrt{\mathcal{I}_{J-1}} \right] \qquad (2.3)$$

Substituting $Z_{\text{Naïve}}$ into 2.3, we get

$$\begin{aligned}
&\left(\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_{\text{New}}}}Z_{J-1}+\sqrt{\frac{\mathcal{I}_{\text{New}}-\mathcal{I}_{J-1}}{\mathcal{I}_{\text{New}}}}Z_J^{\text{New}^*}\right)\sqrt{\mathcal{I}_{\text{New}}}-Z_{J-1}\sqrt{\mathcal{I}_{J-1}}> \\
&\qquad\left[\frac{\sqrt{\mathcal{I}_{\text{New}}-\mathcal{I}_{J-1}}}{\sqrt{\mathcal{I}_J-\mathcal{I}_{J-1}}}\left(d_J\sqrt{\mathcal{I}_J}-Z_{J-1}\sqrt{\mathcal{I}_{J-1}})\right)\right] \\
\implies &\sqrt{\frac{\mathcal{I}_{\text{New}}-\mathcal{I}_{J-1}}{\mathcal{I}_{\text{New}}}}Z_J^{\text{New}^*}\sqrt{\mathcal{I}_{\text{New}}}\frac{\sqrt{\mathcal{I}_J-\mathcal{I}_{J-1}}}{\sqrt{\mathcal{I}_{\text{New}}-\mathcal{I}_{J-1}}}>d_J\sqrt{\mathcal{I}_J}-Z_{J-1}\sqrt{\mathcal{I}_{J-1}} \\
&\qquad\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J}}Z_{J-1}+\sqrt{\frac{\mathcal{I}_J-\mathcal{I}_{J-1}}{\mathcal{I}_J}}Z_J^{\text{New}^*}>d_J \\
&\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J}}Z_{J-1}+\sqrt{\frac{\mathcal{I}_{\text{New}}-\mathcal{I}_{J-1}}{\mathcal{I}_J}}\sqrt{\frac{\mathcal{I}_J-\mathcal{I}_{J-1}}{\mathcal{I}_{\text{New}}-\mathcal{I}_{J-1}}}Z_J^{\text{New}^*}>d_J \qquad (2.4)
\end{aligned}$$

Equation 2.4 contradicts Gao et al. [2013]'s argument that the procedure of Cui et al. [1999] does not downweight any aspects of the data based on the adapted design.

2.6.4 Variance Spending Approach

Fisher introduced the more general concept using the "variance-spending" method that allows one to modify various aspects (sample size, primary endpoint, eligibility criteria, etc) of the trial at any point during the trial as long as the variance spending function is not used up. In addition, the statistical properties of their approach are preserved as long as the weights are chosen prospectively for future observation based on the prior data but before the next observation enters.

2.6.5 Notation

We describe the notation for the fully adaptive design based on the notation used in GSD where, without loss of generality, we assume an unblinded interim analysis is made at the j^{th} interim analysis to modify statistical aspects of the design. There are two interpretations of such an adaptation. We can (1) consider the pre-specified adaptive rule in the previous section as being fully flexible; or (2) completely assume these adaptations to be unplanned

and thus require additional adjustments at the unblinded interim analysis.

In general, the J-look adaptive design has continuation sets, $\mathcal{C}_j = [a_j, b_j] \cup [c_j, d_j]$, and stopping sets, $\mathcal{S}_j = \mathcal{C}_j^c$, described previously has a specified unblinded interim analysis at the h^{th} interim analysis for $h : \{1 \leq h < J\}$. At the h^{th} interim analysis, given the observed $\widehat{S}_h$, the conditional error of rejecting $\mathcal{H}_0$ based on the observed data as far, α_{Cond}, can be expressed as

$$\alpha_{\text{Cond}} = \Pr(S_{h+1} \in \mathcal{S}_{h+1} | S_h = \widehat{S}_h; \theta = 0) + \sum_{j=h+2}^{J} \Pr((\cap_{l=h+1}^{j-1} S_l \in \mathcal{C}_l) \cap (S_j \in \mathcal{S}_j) | S_h = \widehat{S}_h; \theta = 0)$$

The general approach of preserving the conditional rejection probability, also known as the conditional rejection principle (CRP) [Müller and Schäfer, 2001], has a Markovian interpretation that if the future course of a trial is altered in such a way that the Type 1 error conditional on the data observed so far remains the same for the original and altered trials, then the unconditional Type 1 error of the original and altered trials is also preserved under the null hypothesis [Gao et al., 2013]. Thus, conditional on the data observed so far, we can define the altered course with continuation and stopping sets, $\mathcal{C}_h^*$ and $\mathcal{S}_h^*$ as in the setting of section 2.5 respectively. Unlike the prespecified setting where the adaptive rule is known in advance, here, because the design change is unplanned and based on the data observed so far, there is a need to further adjust using the CRP principle. We thus need to control the conditional Type 1 error α_{Cond} by evaluating

$$\Pr(S_{h+1} \in \mathcal{S}_{h+1} | S_h = \widehat{S}_h; \theta = 0) + \sum_{j=h+2}^{J} \Pr((\cap_{l=h+1}^{j-1} S_l \in \mathcal{C}_l) \cap (S_j \in \mathcal{S}_j) | S_h = \widehat{S}_h; \theta = 0) =$$

$$\alpha_{\text{Cond}} =$$

$$\Pr(S_{h+1} \in \mathcal{S}_{h+1}^* | S_h = \widehat{S}_h; \theta = 0) + \sum_{j=h+2}^{J} \Pr((\cap_{l=h+1}^{j-1} S_l \in \mathcal{C}_l^*) \cap (S_j \in \mathcal{S}_j^*) | S_h = \widehat{S}_h; \theta = 0)$$

This approach to monitoring the unplanned adaptations will be used to evaluate fully flexible

adaptive designs in Chapter 5.

2.6.6 Equivalence of Methods under the Two Stage Setting

Jennison and Turnbull [2003] provided a comprehensive review of the adaptive sample size procedures described earlier and demonstrated equivalence of these proposed methods under the two stage setting. In their review, they demonstrated that these flexible procedures do not attain the level of efficiency compared to a GSD. Similar results were obtained from Tsiatis and Mehta [2003] in a more restrictive setting. These early comparisons were often unclear in terms of specifying the optimality criterion. Jennison and Turnbull [2006a,b] re-evaluated the class of adaptive designs from a decision theoretic framework similar to Schmitz [1993] that shed light on the advantages and disadvantages of adaptive procedures.

2.7 "Well understood" vs "Less well-understood" Designs

We highlight some of these current issues with the use of adaptive designs that have further implications in the time to event setting [Gillen and Emerson, 2005, Emerson et al., 2011a, Levin et al., 2013, Levin, 2013, Shoben and Emerson, 2014, Garcia, 2015].

2.7.1 Adaptive Sample Size Re-estimation

Levin et al. [2013] evaluated the class of pre-specified adaptive designs relative to GSDs using the ASN as their optimality criterion. They found negligible improvement with the use of pre-specified adaptive design over the best GSD in terms of ASN when a total of two analyses are allowed. In addition, when the pre-specification was relaxed, group sequential design is almost fully efficient compared to the best analogous adaptive design with ad-hoc unplanned adaptations. Other authors [Tsiatis and Mehta, 2003, Jennison and Turnbull, 2003, 2006a] also found efficiency gains in the use of GSDs over adaptive designs using the maximum statistical information as their optimality criteria. However, in these comparisons, the operating characteristics between the designs were not similar in terms of either the

number or the timing of the interim analyses.

Currently, adaptive modification of statistical information in the immediate setting has not shown any marked advantage over current standard designs with respect to efficiency as defined using ASN. However, in the time to event setting, when the overall cost of the trial is related to the total number of subjects as well as the calendar time of obtaining all relevant events, there may be compelling reasons (such as medical ethics) to consider adaptive modification of patient accrual depending on trends in the treatment effect.

Emerson et al. [2011a] found scenarios among a limited spectrum of pre-specified adaptive designs where the best GSD averaged a higher sample size over pre-specified adaptive designs without compromising the trial duration in the censored setting under proportional hazards (PH). They found potential for benefit in the use of adaptive designs when $\theta < 0.8$ with slow accrual. In fact, adaptive designs that allow for the reduction of the accrual of subjects come with the cost of a slight extension of the study duration. When the accrual rate is fast, adaptive designs do not appear to gain any advantage as the matched GSD achieves shorter average duration over adaptive design with similar number of subjects. While it was uncertain that such benefit may persist with the use of weighted statistics that require further adjustment to control the overall Type 1 error, their research provided some glimmer of hope that adaptive strategies may play a bigger role in the improvement of efficiency over standard designs under time varying treatment effects.

2.7.2 Information Growth

In sequential analysis, monitoring boundaries are approximated by the information fraction. As such, the concept of maximum statistical information plays an important role in making distinctions between group sequential, blinded revision of sample size, and adaptive strategies. In a group sequential design or blinded revision of sample size, we maintain maximum statistical information at some pre-specified level. It is this property that distinguishes these "well-understood" strategies from adaptive strategies when interim analyses allow adaptation to a different maximum statistical information or sequential paths based on unblinded

interim results. In any well-understood design, at the design stage, we know the rule for defining the maximum statistical information. Adaptive procedures designed to test multiple alternatives typically have unknown maximum statistical information at design stage until the unblinded interim analysis when the adaptation is made.

Many such adaptive approaches that adjust this maximum statistical information require specifying the statistical information at the interim analysis as well as at the possibly revised final analysis [Bauer and Köhne, 1994, Proschan and Hunsberger, 1995, Cui et al., 1999, Posch and Bauer, 1999, Chen et al., 2004, Jennison and Turnbull, 2006a,b, Gao et al., 2008, Mehta and Pocock, 2011, Levin et al., 2013]. Characterization is easy in the immediate setting when the sample size is a surrogate measure of statistical information. In the time to event setting, this actual information growth may be unknown unless presuming the use of the logrank statistic under PH.

In longitudinal settings, information growth depends upon the statistical method of choice as well as the underlying assumptions dictating the hypothesis of interest [Gillen and Emerson, 2005, Shoben and Emerson, 2014]. Under the strong null, i.e., when we have exact equality of distributions at all moments, the censoring weights do not affect the statistical validity of the (unweighted) logrank test statistic. However, the use of weighted versions of the logrank statistics in the setting of strong null, or the use of logrank statistics (both weighted and unweighted) in the presence of time varying treatment effects is affected by this censoring distribution [Gillen and Emerson, 2005]. This consequence affects directly the scientific credibility of the trial results and most importantly gives rise later to issues in our ability to precisely quantify information growth as well as the maximum statistical information. Additionally, these statistical issues in the longitudinal setting can be considerably exaggerated with the use of inefficient estimators [Shoben et al., 2010].

Often, we may consider the use of weighted logrank statistics or other test statistics to gain efficiency when we desire to emphasize clinical importance of earlier/later survival time. In such situations, information growth under the strong null is no longer linear as the risk sets are dependent on the censoring and underlying survival curves [Gillen and Emerson,

2005, Shoben, 2010, Brummel and Gillen, 2014]. Even though such a choice of an efficient weighting scheme is made to gain efficiency under non proportional hazards alternatives, characterization of the control of the overall Type 1 error needs to be evaluated under the strong null hypothesis of proportional hazards. Furthermore, any common methods of analyzing time to event data may induce a time varying treatment effect across time or stage of the clinical trial unless using the Cox regression or the unweighted logrank statistics, and that the proportional hazards assumption is valid.

2.7.3 Inference after Adaptations

With a sampling rule, it is convenient to understand the operating characteristics to enable comparisons with GSDs. Standard frequentist procedures using the sampling density in section 2.5 can be applied. In these settings, where the expected length of confidence intervals for the parameter of interest is the optimality criterion, methods using the minimal sufficient statistics based on the LR ordering or stage-wise ordering can be compared to adaptive trial methods such as those proposed by Brannath et al. [2009]. Levin et al. [2014] provided a comprehensive evaluation of these approaches when prespecified adaptations are made. As might be anticipated based on statistical theory for FSD, in a broad spectrum of adaptive settings, methods based on the minimal sufficient statistic and using the likelihood ratio ordering were found to be more efficient than those adaptive methods that cannot be implemented based on the minimal sufficient statistic alone. However, it should be noted that these observations are empirical: *No general theory about uniformly most accurate CI is available in the sequential setting.*

In fully adaptive procedures, the lack of pre-specification generally means that the inference procedure in presence of early stopping is more limited to approaches by Brannath et al. [2009] or Gao et al. [2013]. The degree to which fair comparisons can be made are more difficult in this over parameterized space when issues of information growth and schedule of analyses further impact the monitoring rule.

2.7.4 Adaptive Enrichment

Several authors recognize that methods described in section 2.6 can be used to allow adaptive modifications to randomization ratios or subgroup enrichments. In enrichment/adaptive randomization ratio settings, when secular trends exist in accrual of subjects, Garcia [2015] cautioned that there is a need to adjust for the analysis when estimating the treatment effect as a consequence of confounding as introduced through the randomization ratio. More research is needed to better understand statistical issues when these adaptive procedures are already in place such as I-SPY-2 [Barker et al., 2009] or BATTLE [Kim et al., 2011]. Because of the close interplay between multiple comparison adjustments and finding the right subgroup, adapting randomization ratios in presence of secular trends can further bias the results away from the null, potentially affecting the predictive values of the treatment in future Phases.

2.7.5 Operational Bias

One key distinction between group sequential or blinded revision of sample size as opposed to adaptive strategies is the preservation of blinding of treatment groups to all groups other than the DMC and the statistical center performing the analysis for the DMC. In adaptive strategies, this blinding is broken and may be conducted by another statistical center independent from the main statistical center performing analysis for the DMC [Fleming, 2006]. There are additional concerns in ensuring confidentially in adaptive designs further discussed in FDA [2010]. Confusion on the roles of each party involved may compromise the understanding of the clinical trial results [Emerson, 2006, Fleming, 2006, Emerson and Fleming, 2010].

2.7.6 Patient-wise Separation

In the immediate outcomes setting, response measurements are made after having been on the treatment/placebo for a relatively short period of time (e.g., weeks) relative to the entire

time during which the RCT is conducted (e.g., years). In these settings, participants are accrued and randomized in groups, and all participants' responses are measured at clear defined time points before the next sequential group of patients are randomized.

In other clinical settings, patients who are randomized to treatment/prevention strategies to treat a particular disease condition may experience the effect of the intervention long after receiving the treatment. Many of these clinical settings are concerned with irreversible outcome(s) that take longer time than what can be quickly evaluated within a matter of weeks. Such outcomes can include mortality in cardiovascular trials or seroconversion of HIV status in HIV prevention settings. Thus, interim analyses may be conducted when the majority of the subjects do not have the outcome of interest but may be required to be monitored further past the interim analysis for the outcome to mature.

Bauer and Posch [2004] pointed out an issue with the use of unblinded interim analyses in the time to event setting particularly when surrogate or short term data collected on the patients is available. The use of unblinded secondary outcome data that is predictive of the patients' response to the treatment becomes more prominent in the settings with delayed ascertainment of outcomes. At interim analyses, some participants randomized to the treatment/placebo may not have the outcome of interest but do have data available on a surrogate outcome correlated with the primary outcome. Other participants may not yet been accrued into the study. This differential length of follow-up constitutes what is known as "patient-wise" separation at interim analyses. The use of unblinded procedures based on surrogate secondary outcomes for adaptations can potentially increase the risk of inflation of the overall Type 1 error [Bauer and Posch, 2004]. Naïve re-weighting of critical values based only on information from the primary outcomes would not necessarily control for the additional correlation of the surrogate endpoint with the primary test statistic. Subsequent literature by Jenkins et al. [2011], Irle and Schäfer [2012], Magirr et al. [2014] and Magirr et al. [2016] have proposed approaches to account for this correlation, though none of these literature have comprehensively evaluated the cost in efficiency relative to the use of GSDs.

2.8 Summary

In summary, we presented some of the distinctions between group sequential designs as opposed to these more recently described adaptive strategies. In particular, the immediate outcomes settings provide notational convenience for setting up the foundations of both group sequential design or pre-specified adaptive design. The notation in the immediate outcomes settings in this Chapter extends easily to the time to event setting. The issues highlighted in section 2.7.1 and 2.7.2 are of focus in this dissertation. While we will later address issues related to operational bias in section 2.7.5, an assumption made throughout this dissertation is that surrogate outcomes are not considered when making unblinded adaptations in section 2.7.6.

Chapter 3

Impact of Analysis Schedule on Operating Characteristics of Designs

GSDs often present a natural setting for us to gain intuition on aspects of the operating characteristics of adaptive designs. As described in Chapter 2, an adaptive design can be viewed as a stochastic hybrid of two or more group sequential designs. In section 2.3, we described some "well-known" properties of GSDs and how changing aspects of the design, such as the degree of early conservatism, or the schedule of interim analyses, can affect other operating characteristics such as average efficiency, maximum sample size, or power.

For example, with one-sided symmetric designs, when holding power fixed and allowing the maximum statistical information to vary, the known property of Pocock boundary shape functions being more efficient in terms of ASN over OBF is true when interim analyses are equally spaced and the true treatment effect is such that the probability of rejecting the null is between 0.001 and 0.999. Many of these "well-known" assumed properties of GSD, however, are often characterized based on equally spaced interim analyses. These characteristics may no longer hold true when the analysis schedule is no longer equally spaced, as would be typical under many proposed adaptive RCT.

Considerations to modify aspects of a RCT design are often planned at a specified "adaptive analysis". When adapting based on a FSD with no early stopping at the adaptive analysis, n is typically modified to $\widetilde{n}$ to increase statistical power β. When adapting based on a GSD, the popular approach to sample size re-estimation is often proposed at a planned penultimate interim analysis. This future sample size is modified from a previously specified

maximum statistical information N_J to $\widetilde{N}_J$, where $\widetilde{N}_J$ is some function of interim treatment estimate $\widehat{\theta}_{J-1}$. Changing this maximal sample size, however, also changes the boundary shape function. That is, with an adaptation from N_J to $\widetilde{N}_J$, we also change the schedule of analyses. The stopping boundaries implemented prior to the adaptive analysis must now be considered relative to the new maximal statistical information.

The degree of early conservatism for the choice of the revised monitoring boundary is dependent upon the choice of $\mathbf{\Pi} = (\Pi_1, \cdots, \Pi_J), A, P, R$, and N_J at design stage, where $\Pi_j = N_j/N_J$. When the maximum sample size is no longer maintained with the original monitoring boundary, then the desired degree of early conservatism may be appropriate based on another GSD with $\mathbf{\Pi'} = (\Pi'_1, \cdots, \Pi_J), A', P', R'$, and N'_J.

We motivate the above with an example based on a modified version of the schizophrenia trial described in Mehta and Pocock [2011]. We assume a two-look, one-sided symmetric GSD with an OBF boundary that has 79.7% power to detect a mean Negative Symptoms Assessment score difference of $\delta_1 = 2$ with a one-sided level 0.025 using known standard deviation of 7.5. In this GSD (*OBF442*), a maximal sample size of 442 was initially chosen with an interim analysis scheduled after 208 subjects' data were available (47.1% of the maximum statistical information). In *OBF442*, an observed interim treatment effect greater than or equal to 2.9868 would suggest early termination in favor of efficacy while an observed interim treatment effect less than or equal to -0.1757 would suggest early termination for "futility". Any observed estimates between -0.1757 and 2.9868 would suggest continuing to the maximal sample size of 442.

We then suppose that investigators want to consider adapting to a larger sample size when interim results are "promising", but not as large as initially anticipated. In their example, Mehta and Pocock considered an adaptive rule that would increase the conditional power of the study up to 80% within some "promising zone", but not exceed a maximal sample size that is doubled of what was originally planned (so up to 884). As illustrated by Levin et al. [2013], nearly the same efficiency can be achieved by considering a more discrete adaptive process, and we use such an approach in our illustration.

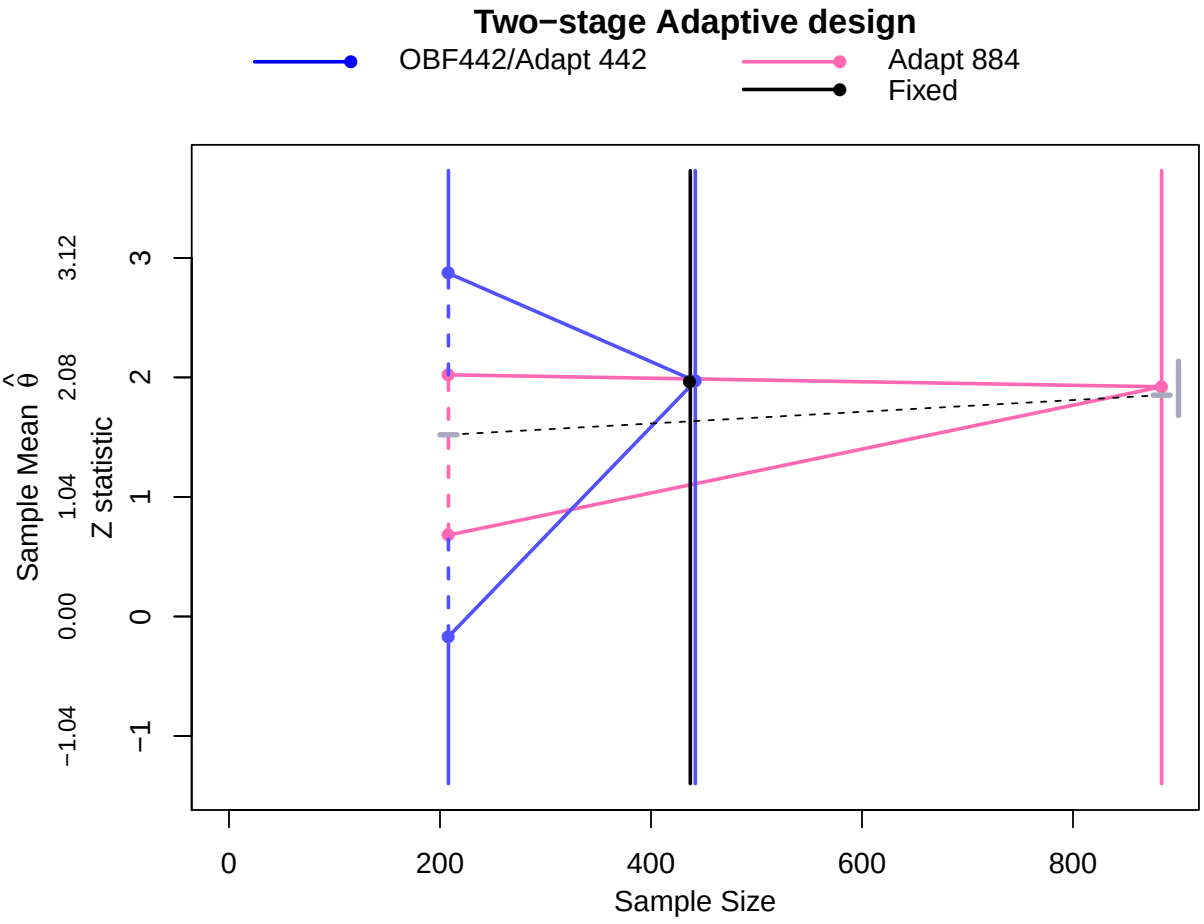

Figure 3.1: Adaptive design as a stochastic hybrid of two GSDs with the continuation regions (as represented by the dotted lines) in blue to 442 subjects, and in pink to a maximum sample size of 884. In gray are the plausible critical values based on CHW.

Hence, we arbitrarily presume that an interim estimated treatment effect between 0.7082 and 2.1029 will lead to an adaptive increase in the maximal sample size to 884 subjects (23.5% of the maximum statistical information based on 884). Other estimates in the continuation region (in Figure 3.1) will leave the maximal sample size unchanged at 442. This adaptive design (a stochastic hybrid of *OBF442* in blue, and *Adapt884* in pink as shown in Figure 3.1) has an experiment-wise Type 1 error rate of 0.025, and has 89.968% power to detect the design alternative of 2.0 with known standard deviation of 7.5. We note that with this arbitrarily

chosen adaptive rule, there is a 22.6% chance of increasing the sample size under the null hypothesis. As noted by Emerson et al. [2011b], there is rarely a reason to use an adaptive rule that would double the maximal sample size in this manner.

It is of interest to examine the resulting adaptive design relative to competing GSDs that are in some way comparable. We want to consider designs that have the first interim analysis conducted with 208 subjects, and have power of 89.968% to detect the alternative of 2.0. This then allows us to vary how we define the boundary shape function and the maximal sample size. Based on these constraints, we can construct several other GSDs for comparison, all of which are one-sided level 0.025, with 89.968% power to detect the design alternative of 2.0, and a total of two analyses with the first analysis conducted at 208 subjects. These alternative GSDs can include:

1. *OBF90*: Boundary shape function of a one-sided symmetric OBF yielding a maximal sample size of 592, with critical values (on the Z statistics scale) at the interim analysis of -0.9834 for futility and 3.3103 for efficacy. In this design, the interim analysis occurs at $\Pi_1 = 208/592 = 35.1\%$ of the total statistical information.

2. *MOD90*: Boundary shape function is chosen such that the critical values (on the Z statistics scale) at the interim analysis agree with the original OBF442, yielding maximal sample size of 608.32. The design with these constraints is no longer within the family of one-sided symmetric designs. The boundary shape function for the efficacy boundary is $P = 0.8532$, a value intermediate to the OBF ($P = 1$) and the Pocock ($P = 0.5$). In this design, the interim analysis occurs at $\Pi_1 = 208/608 = 34.2\%$ of the total statistical information.

3. *MOD884*: Boundary shape function with $P = 0.37105$ corresponds to a one-sided symmetric design with analyses conducted at 208 and 884. Such a boundary shape function is less "conservative early" than a Pocock design, which generally indicates a loss of average efficiency. In this design, the interim analysis occurs at $\Pi_1 = 208/884 =$

23.5% of the total statistical information.

Note that we can further consider a GSD with a maximal sample size of 884 subjects (*MODMatched884*, not shown here), and constrain this design to have the same boundaries as *OBF442* conducted at the interim analysis of 208 subjects. Such a design has an overall power of 96.3% to detect the design alternative of 2. The boundary shape function for these designs are shown in Figure 3.2 with the respective ASN curves and relative power as shown in Figure 3.3.

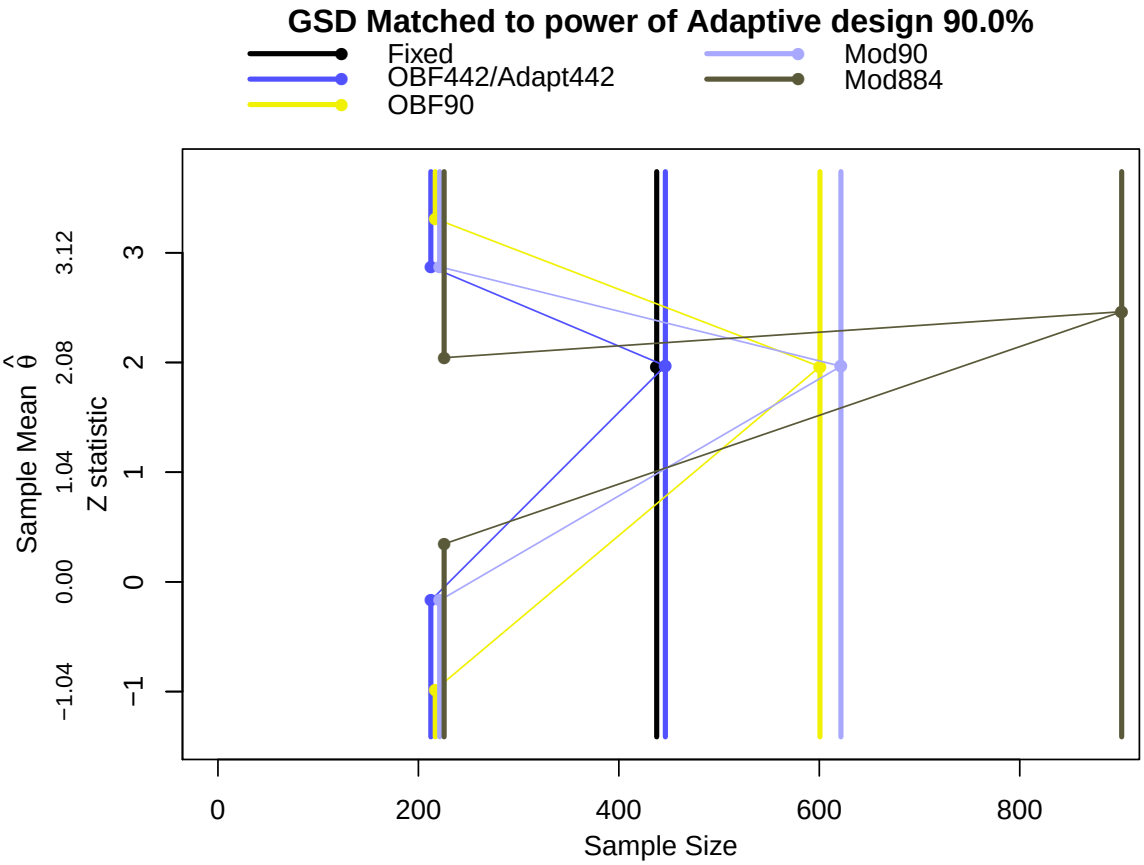

Figure 3.2: Sequential boundaries for OBF442/Adapt442 (79.7% power) and other GSDs as shown. The GSDs corresponding to *Mod884*, *GSD90*, and *Mod90* are matched at 89.97% power to the adaptive design that comprises of *OBF442* and *Adapt884* under the design alternative of 2, and known standard deviation of 7.5.

The above illustration highlights the difficulty in trying to compare GSDs with adaptive designs. As we change the maximum statistical information, we generally change our relative schedule of analyses. It thus becomes unclear whether our schedule of analyses should be described relative to the new maximum sample size or the original design as this reflects different degrees of early conservatism. In the above example, when we preserve the level of early conservatism based on *OBF442* at the first interim analysis, the design *Mod90* beats *OBF90* in terms of ASN that is more efficient for all alternatives between 0 and 2. We will appeal to some of these difficulties as we try to identify aspects of the potential adaptive rules that might explain differences in operating characteristics. That is, we will appeal to how the adaptive rule might be viewed as changing the maximal sample size, power function, relative timing of analyses, and the boundary shape function.

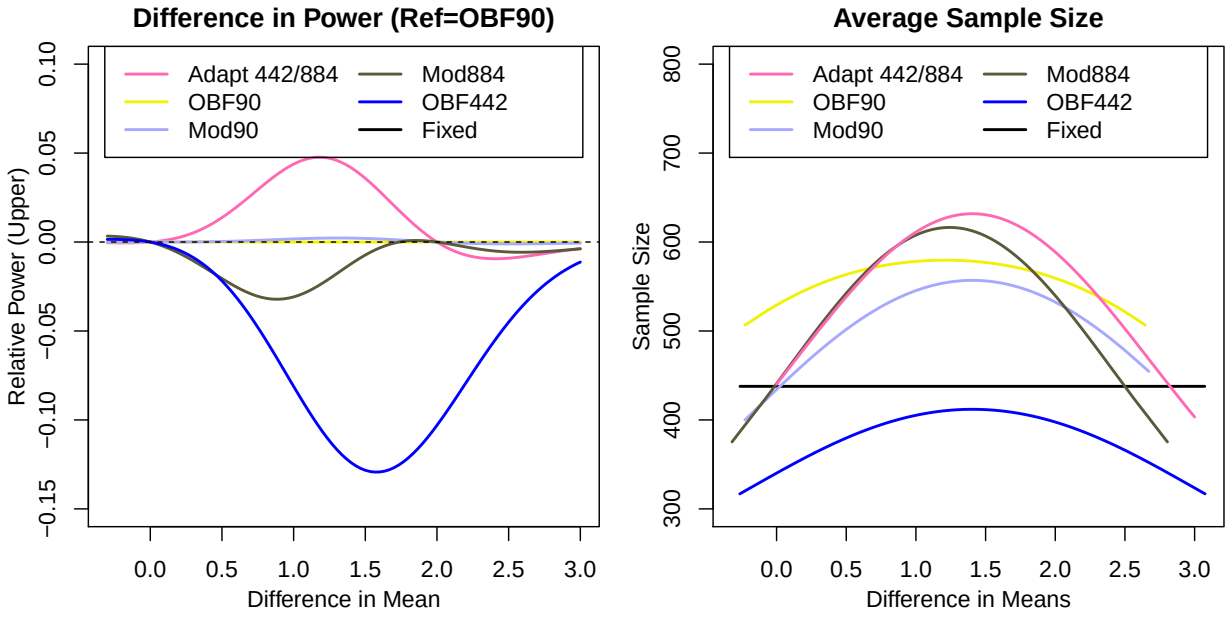

Figure 3.3: Relative power curves calibrated to *OBF90* on the left and ASN curves on the right for the various designs (*OBF90*, *Mod90*, and *Mod884*) are matched to the power of the pre-specified adaptive design that comprises of *Adapt442* and *Adapt884* with an overall power 89.97% under the design alternative of 2, and known standard deviation of 7.5.

It should be noted that the above results all presumed a pre-specified design and analyses based on the minimal sufficient statistic. If, however, the adaptive rule was not adequately pre-specified, the critical value chosen at the final analysis would have to be determined by, say, the CHW approach. In Figure 3.1, the grey line depicts the range of plausible critical values that might be used depending upon where the interim estimate fell in the continuation region leading to an adaptive increase in the sample size. When using such variable critical values, the power of the adaptive design decreases negligibly to 89.90% (the corresponding pre-specified rule gives a simulated power of 89.97%, the *OBF442* design gives 79.67% power based on 10,000,000 simulations).

This chapter is divided into two sections. In the FSD setting with no early stopping at the adaptive analysis, we investigate the impact on the overall power of the design when the timing of the adaptive analysis to increasing/decreasing the maximum statistical information is varied. More specifically, when there is either logistical difficulty in accrual, or event rates differ from trial assumptions, adaptations on the basis of unblinded interim analysis may have to be made based on results gathered from early interim analysis as opposed to late occurring analysis.

Various operating characteristics within the family of GSDs can often be improved upon by increasing the number of analyses (at the expense of increasing the maximum statistical information) or increasing the maximum statistical information while holding other aspects of the design fixed. When the total number of interim analyses has not been changed, the degree of early conservatism based on the original design is no longer preserved when modifications are made to the maximum sample size. This interplay of changing both the maximum statistical information and analysis schedule in a GSD contributes to the difficulty of understanding the advantages and disadvantages of adaptive designs. In order to better separate the efficiency issues associated with the use of a pre-specified adaptive sampling scheme vs having to perform adjusted analyses for fully adaptive designs, it is vital that we understand the "game theory" that the adaptive clinical trialists might be able to use. To do so, we rely on the properties of the GSDs to describing this bivariate relationship by holding

other operating characteristics constant.

In a fully flexible design, the adaptive trialists make adaptations based on some knowledge of the unblinded treatment effect that re-weights the incremental test statistics differently across stages. These strategies are often evaluated as if only one of the potential sample paths is chosen in a flexible design. Thus, in order to characterize this degree of uncertainty, we consider a prespecified version of this fully adaptive sampling scheme in order to understand the best operating characteristics one would have obtained.

3.1 Effect of Unequally Spaced Interim Analyses on ASN

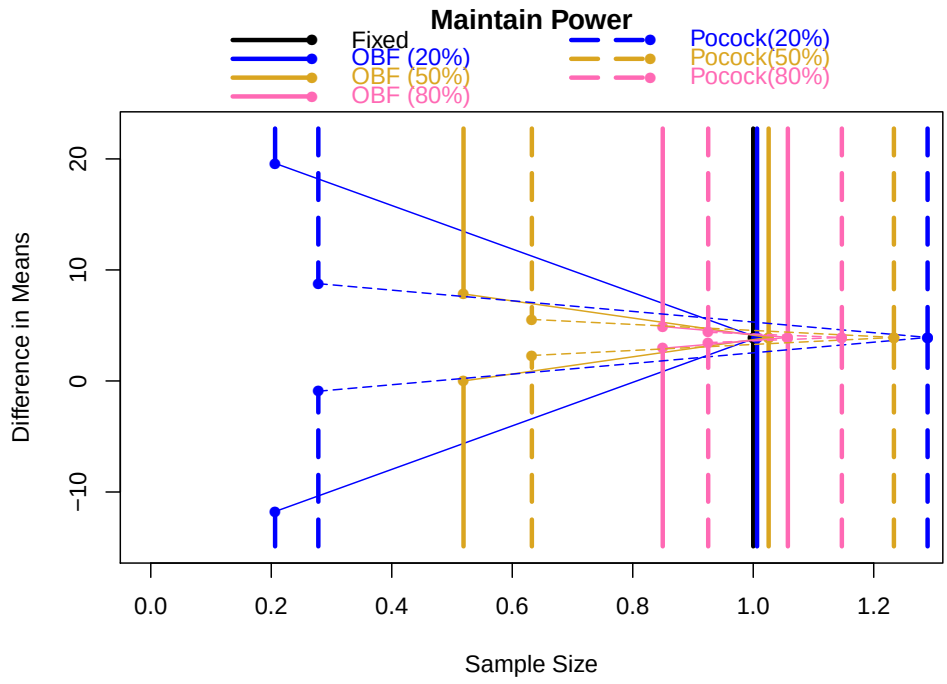

Figure 3.4: Sequential boundaries for the two-stage, one-sided symmetric, OBF and Pocock designs with interim analyses conducted at 20%, 50%, 80% of the maximum statistical information. The maximum statistical information for the Pocock boundaries are inflated least relative to the fixed sample design when interim analyses are conducted as late as possible. The maximum statistical information for OBF boundaries are inflated more when the interim analyses are conducted later.

Our knowledge in the GSD literature is such that the ASN based on the Pocock rule typically dominates the ASN curve for the OBF rule $\forall\ \theta \in (0, \theta_A)$ (Figure 3.5). Consider the OBF and Pocock monitoring boundaries where an interim analysis is conducted at either 20%, 50%, or 80% of the maximum statistical information. The equally spaced analyses are often favored and chosen at design stage for convenience. In this case, the OBF with an interim analysis conducted at 50% of the maximum statistical information would have stopped the trial for either futility or non-efficacy when the interim estimated treatment effect of less than 0 is obtained (Figure 3.4).

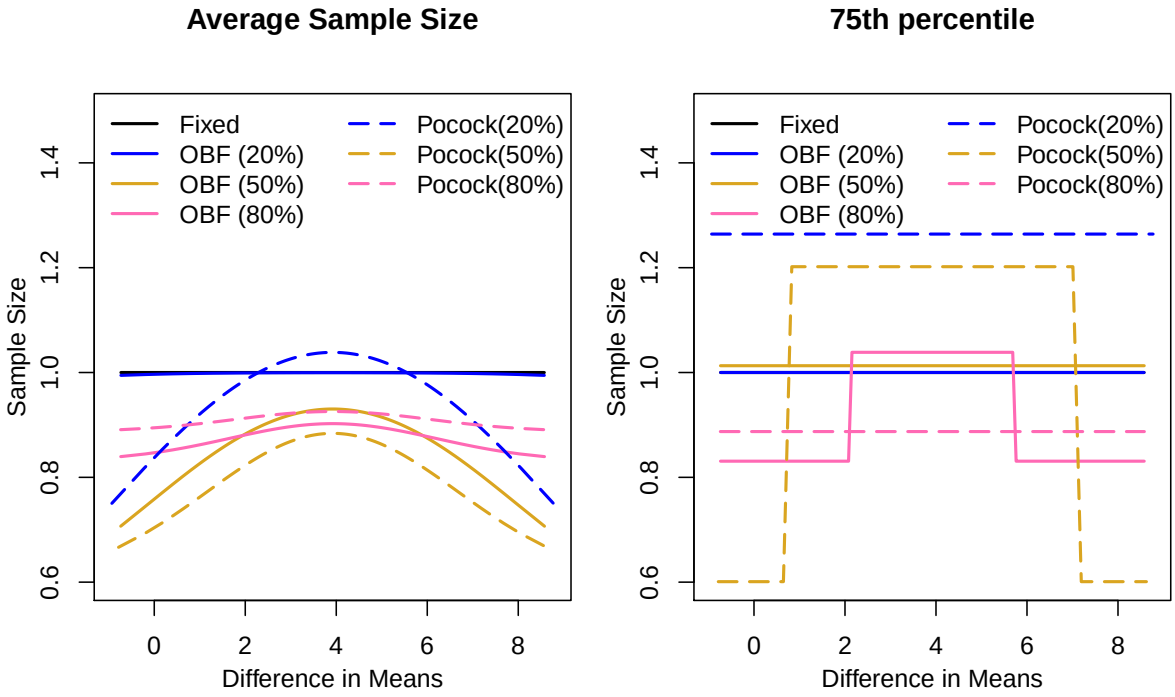

Figure 3.5: Average and 75$^{\text{th}}$ percentile of the sample size distribution for the one-sided, symmetric two-stage OBF and Pocock with interim analyses conducted at either 20%, 50%, or 80% of the maximum statistical information. The common assumption that the Pocock design has a lower ASN relative to the OBF design holds when an interim analysis is conducted at 50% of the maximum statistical information. When the interim analyses are no longer equally spaced, this known property may not be true for values between the null and alternative.

However, when the schedule of analyses is no longer equally spaced for the OBF and the Pocock rule, this known property of Pocock being more efficient in terms of ASN within the null and alternative values relative to the OBF boundary no longer holds true. When an interim analysis is conducted at 20% of the maximum sample size, the ASN curve for the Pocock boundary is no longer dominating the ASN curve for the OBF rule $\forall\ \theta \in (0, \theta_A)$ (Figure 3.5). In contrast, when the schedule of analyses is moved such that this interim analysis is conducted at 80% of the maximum statistical information, the OBF design is now more efficient than the Pocock boundary $\forall\ \theta \in (0, \theta_A)$.

In practice, when flexible modifications are made to the statistical information of the design at some interim analyses, the pre-defined level of early conservatism may no longer be held constant after an adaptation.

3.2 Timing of Interim Analyses

Previous research [Jennison and Turnbull, 2003, Tsiatis and Mehta, 2003, Jennison and Turnbull, 2006a, Levin et al., 2013] have not found adaptive designs to be markedly more efficient than GSD in the best case. This is in part due to the use of statistics that violate the minimal sufficiency principle, leading to inefficient weighting of the available data [Bauer and Köhne, 1994, Proschan and Hunsberger, 1995, Cui et al., 1999]. However, clinical trialists also choose inefficient adaptations in practice where they either more than double the sample size at the penultimate analysis or choose a poor sampling scheme (in terms of timing at which the adaptation is made) [Chen et al., 2004, Gao et al., 2008, Mehta and Pocock, 2011]. Limited research has attempted to address how each of the individual components when chosen optimally can affect the efficiency of the design. In situations when the sampling scheme is not pre-specified, the impact of the role of inefficient statistics is unclear.

In this section, we investigate the efficiency loss using overall power as our optimality criterion to better understand the impact of the timing of analysis schedule when the adaptive sampling scheme is not pre-specified. Weighted statistics are thus used in these situations to adjust for the flexible adaptations. In contrast, we compare these flexible designs to adap-

tive designs that use minimal sufficient statistics at the time of termination as in section 2.5. When the target of inference is the difference in incidence rates, or the sample size is dependent upon disease rates in the time to event setting, a lower than anticipated event rate may necessitate "flexible" adaptations based on unblinded interim results to facilitate design changes.

3.2.1 Notation

Without loss of generality, we consider the two-stage FSD where an interim analysis is made with no early stopping under the immediate setting when the target of inference is the difference in treatment means. Recall the notation setup in section 2.1, let potential pairs of observations for the prespecified first stage and second stage sample size be n_1^*, n_2^* respectively with the total sample size of the trial be $n = n_1^* + n_2^*$. Let $\theta = \omega_1 - \omega_2$ be the difference in treatment means with variance σ^2, i.e., each treatment group has variance $\sigma^2/2$. The respective distributions for the estimate of the difference in treatment means for each stage are $\overline{X}_{n_j^*}|n_j^* \sim (\theta_i, \frac{\sigma^2}{n_j^*})$ for $j = 1, 2$. The hypotheses of interest are $\mathbb{H}_0 : \theta \leq \theta_0$ vs $\mathbb{H}_A : \theta \geq \theta_A > \theta_0$. We shall suppress * for Z_1, n_1 for the rest of this section.

The best linear unbiased estimator $\overline{X} = w_1\overline{X}_{n_1} + w_2\overline{X}_{n_2^*}$ has the most efficient weighting under this sampling scheme with $w_1 = \frac{n_1}{n_1+n_2^*}$ and $w_2 = \frac{n_2^*}{n_1+n_2^*}$. In general, n_1, n_2^* are fixed under this optimal procedure since the sample size has been predetermined at the beginning of the trial. Thus, $Z_1|n_1$ is independent of $Z_2^*|n_2^*$ under $\mathbb{H}_0$ and $w_1Z_1 + w_2Z_2^* \overset{\mathbb{H}_0}{\sim} N(0, 1)$.

Consider the flexible two-stage design in Fisher [1998] where we prespecified recruiting n subjects. After recruiting n_1 subjects, our estimator at interim $\overline{X}_{n_1}$ based on the data collected so far has mean θ and variance $\frac{\sigma^2}{n_1}$. Suppose a design modification is made on the basis of the estimated treatment effect obtained thus far. Conditional on the estimated treatment effect $\overline{X}_{n_1}$, and the first stage sample size, n_1, we can determine the number of second stage subjects $\widetilde{n}_2^*$ need to be recruited to detect the same design alternative. We can preserve the overall Type 1 error by applying the remaining weights, $w_2 = n_2^*/n$, on each of the $\widetilde{n}_2^*$ subjects. Each subject is given weights $\frac{1}{n} \times \frac{n_2^*}{\widetilde{n}_2^*}$ rather than $\frac{1}{n}$.

Alternatively, we can optimize this adaptation at design stage by pre-specifying the same rule that the flexible design uses that led to $\widetilde{n}_2^*$ subjects. Then, this design would be optimal by theory of best linear unbiased estimator since we would efficiently weigh our estimator by the proportion of sample size from each stage, i.e., $w_1 = \frac{n_1}{n_1+\widetilde{n}_2^*}$ and $w_2 = \frac{\widetilde{n}_2^*}{n_1+\widetilde{n}_2^*}$ with the respective variance as illustrated in Table 3.1.

Table 3.1: Table of the weights for each estimator and the (conditional) variance as a result of adapting.

Designs	Sample size	Weights	Estimator	Variance
Original	$n = n_1 + n_2^*$	$w_1 = \frac{n_1}{n};\ w_2 = \frac{n_2^*}{n}$	$w_1\overline{X}_{n_1} + w_2\overline{X}_{n_2^*}$	$\frac{\sigma^2}{n}$
Fully Adaptive	$\widetilde{n} = n_1 + \widetilde{n}_2^*$	$w_1 = \frac{n_1}{n};\ w_2^* = \frac{n_2^*}{n}$	$w_1\overline{X}_{n_1} + w_2^*\overline{X}_{\widetilde{n}_2^*}$	$\frac{n_1 + \frac{(n_2^*)^2}{\widetilde{n}_2^*}}{n^2}\sigma^2$
Prespecified Adaptive	$\widetilde{n} = n_1 + \widetilde{n}_2^*$	$w_1^* = \frac{n_1}{\widetilde{n}};\ w_2^{**} = \frac{n_2^*}{\widetilde{n}}$	$w_1^*\overline{X}_{n_1} + w_2^{**}\overline{X}_{\widetilde{n}_2^*}$	$\frac{\sigma^2}{\widetilde{n}}$

In either the original or the prespecified design, the total sample size (n and $\widetilde{n}$ respectively) is fixed. Hence, the conditional variance of the estimator is also the unconditional variance of the estimator. The variance under the flexible adaptive design is conditional on the final sample size $\widetilde{n}$ since n_1 is fixed and $\widetilde{n}_2^*$ is conditionally random depending on the estimated treatment effect at stage 1. Because $\widetilde{n}_2^*$ is some function of $\overline{X}_{n_1}$ and n_1, so long as the total weights in the second stage is spent on the $\widetilde{n}_2^*$ subjects, the overall Type 1 error is preserved following theoretical arguments by Fisher [1998]. The conditional variance of the adaptive estimator is no longer equivalent to the unconditional variance of the estimator. The unconditional variance estimator needs to account for all other potential sample sizes the adaptive design would have chosen that did not occur.

3.2.2 Relative Efficiency

We can evaluate the relative efficiency of the conditional variance of the estimator obtained from the flexible adaptive design relative to the (conditional) variance of the estimator based on the prespecified adaptive design. The conditional relative efficiency can be expressed as

$$\begin{aligned}\text{Relative Efficiency} &= \frac{\text{Var}\left[\overline{X}_{\widetilde{n}}\middle| n_1, \widetilde{n}_2^*\left(\overline{X}_{n_1}\right)\right]}{\text{Var}\left[\overline{X}_{\widetilde{n}}^{\text{Opt}}\middle| n_1, \widetilde{n}_2^*\right]} = \frac{\widetilde{n}}{n^2}\left(n_1 + \widetilde{n}_2^*\frac{(n_2^*)^2}{(\widetilde{n}_2^*)^2}\right) \\ &= \frac{(\gamma+\theta)\left(\gamma+\frac{1}{\theta}\right)}{(\gamma+1)^2} \geq 1 \quad \text{unless } \widetilde{n}_2^* \equiv n_2^*\end{aligned}$$

The last line above follows by parameterizing $\theta = \widetilde{n}_2^*/n_2^*$ and $\gamma = n_1/n_2^*$. We can interpret θ as the fraction of the increase/decrease in the revised second stage data with respect to the second stage data based on the original design and γ being the ratio of sample sizes of the first stage to the second stage from the original design. The contour plot for the above relative efficiency is as shown in Figure 3.6.

For each fixed γ, when θ increases ($\theta > 1$), the relative efficiency increases. When θ (< 1) decreases to 0^+, we see that this relative efficiency also increases. This translates directly to a loss of efficiency in the variance estimator of the flexible design as opposed to the optimal design. Interestingly, at bigger values of γ, i.e., at late adaptations when we have a more reliable estimate of our treatment effect, the relative efficiency increases in a non linear fashion. This increase in relative efficiency also indicates a less efficient weighting with this flexible adaptation.

We can further re-parametrize the above by defining k to be the relative increase/decrease of the total sample size n of the original design. Let the total sample size of the flexible adaptive design be revised from n to $\widetilde{n}$ which is k times the original sample size n. Then $n_1 + \widetilde{n}_2^* = kn = k(n_1 + n_2^*)$. Via the substitution $\theta = (k-1)\gamma + k$, the relative efficiency as

a function of γ and k can be re-expressed as follows:

$$\text{Relative Efficiency} = k\frac{1+(k-1)\gamma}{k+(k-1)\gamma}$$

with the restriction that $\theta > \frac{1}{n_2^*} > 0$, i.e., the trial must sample at least 1 more participant.

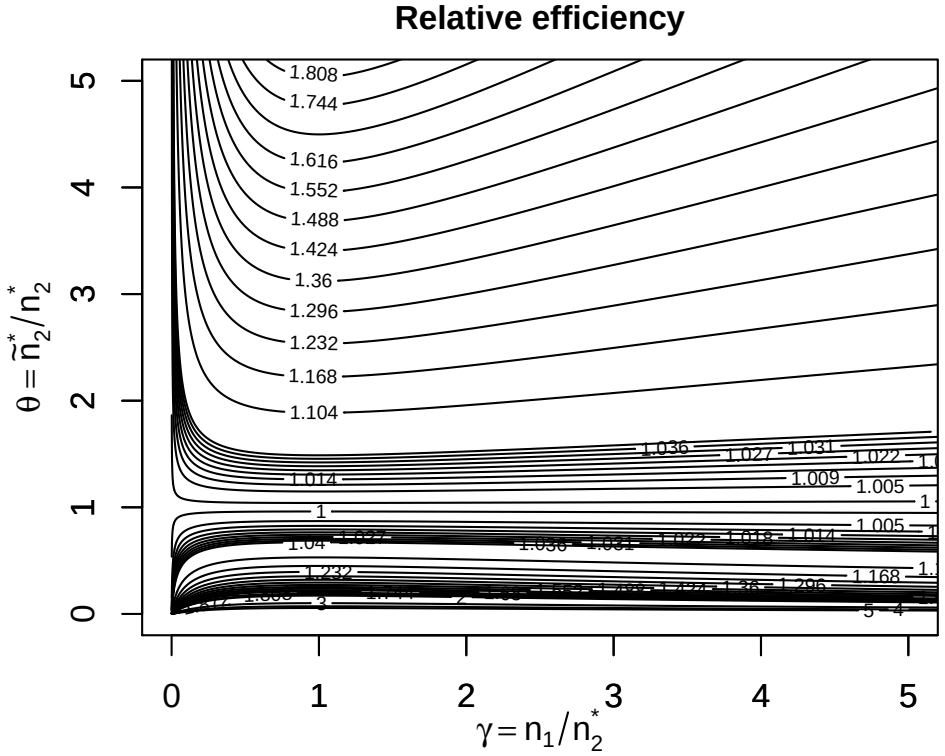

Figure 3.6: Contour plot for θ vs γ comparing the relative (conditional) efficiency of the variance estimators of the flexible design vs that of the optimal estimator for the design with $n_1 + \widetilde{n}_2^*$ subjects.

The contour plot of the relative efficiency (Figure 3.7) presented for k vs γ provides another interpretation of Figure 3.6. We now see that when $\gamma < 1/9$, the relative efficiency is close to 1, indicating that any form of adaptation leading to a larger sample size is essentially reweighting the second stage data to be similar to the equal weighting scheme from the

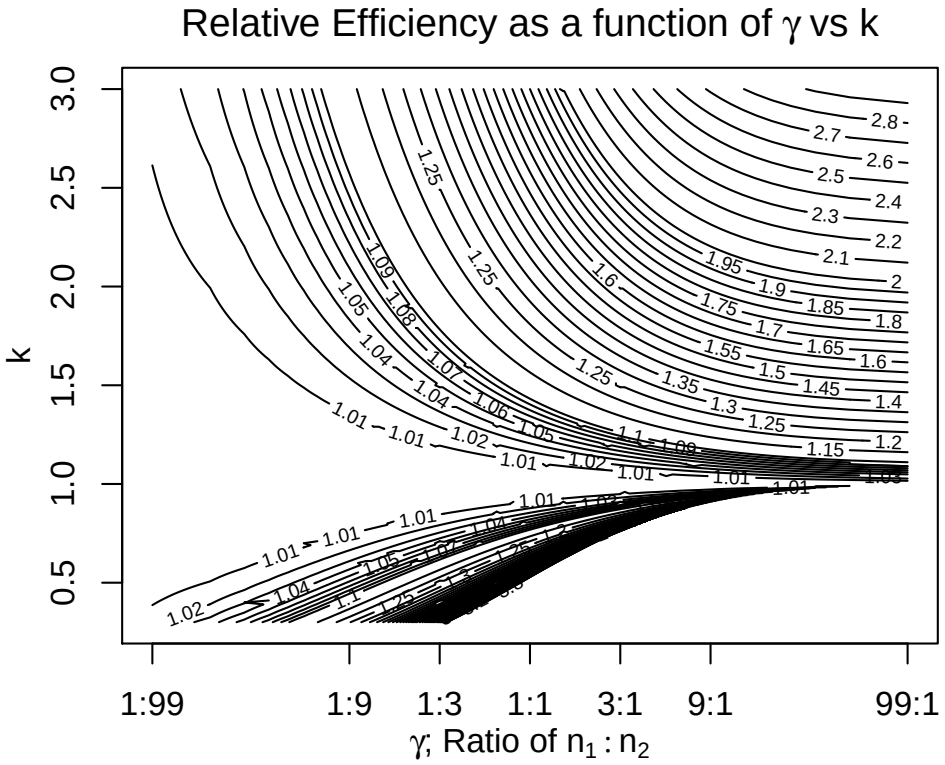

Figure 3.7: Plot of relative efficiency as a function of k vs γ when increasing the original sample size to kn. Note that certain contours may only exist if $\theta > \frac{1}{n_2^*} > 0$.

optimal design. However, when these adaptations are made earlier on to decrease the sample size, this may lead to a bigger loss of efficiency depending on k. If $k = 0.5$, and an adaptation is made at 1/10 of n, then the second stage data is weighted heavily using 9/10 of the remaining weights that is more than twice the weight of the optimal design. When $\gamma > 9$, adaptations are close to the maximal sample size, any form of adaptation to increase the second stage sample size by $k > 1$ can lead to a much bigger loss of efficiency since this remaining, revised larger sample size is allocated a smaller weight as compared to the first stage data.

It is of use to quantify how the relative efficiency based on unplanned adaptations can

translate directly to the overall power of the design when the probability of adaptation changes. The above characterization makes the assumption that the unblinded adaptation is made all the time without considering other potential adaptations. To appropriately compare between unplanned adaptations with FSDs or designs with known sampling scheme, we must thus hold fixed the probability of an unplanned adaptation to modify the final sample size from n to $\widetilde{n}$.

3.2.3 Design Settings

A prespecified sampling scheme has some known probability p of adapting to either a larger/smaller sample size. With blinded analyses, p is independent of the estimated interim treatment effect under majority of the settings. With unplanned flexible adaptations, this probability is unknown, and direct comparison with competing designs with known sampling schemes is difficult. This is because there is an infinite number of decision rules one can make to determine the second stage sample size, ranging from stopping after n_1 subjects are accumulated to recruiting a finitely large sample size for the second stage [Emerson, 2006].

Consider the following design scenarios in the superiority setting with the following maximum statistical information that is dependent on some sample size n to detect a particular alternative at some level α and power β with suitable assumptions of the variance estimates (population variance, event rates, etc).

1. $\mathcal{I}_n$.
2. $\mathcal{I}_{\widetilde{n}}$ ($\mathcal{I}_{\widetilde{n}}$ may or may not be known at planning)
3. $\mathcal{I}_n$ subjects with probability p of adapting to statistical information $\mathcal{I}_{\widetilde{n}}$ in a blinded analysis
4. Study design planned with statistical information $\mathcal{I}_n$, and an unplanned adaptation at some unblinded interim analysis leads one to revise the statistical information to $\mathcal{I}_{\widetilde{n}}$

where $\mathcal{I}_{\widetilde{n}}$ may be bigger or smaller than $\mathcal{I}_n$.

Study design 1 is typical of the setting where a sponsor may plan their design based on an assumed treatment effect of interest, population variance of the estimated treatment effect, or expected event rate of the placebo group in event driven studies to achieve some specific power.

When trial assumptions differ from practice, study design 1 may be underpowered and the planned statistical information $\mathcal{I}_n$ may not provide the desired power as anticipated. Therefore, in study design 2, this corresponds to a design that correctly specifies the variance of the treatment effect, or the true event rate of the placebo group. In time to event settings, or settings when we are interested in the difference in incidence rates, the total statistical information may be prespecified in terms of calendar time. Hence, study design 2 corresponds to the setting of having the optimal statistical information $\mathcal{I}_{\widetilde{n}}$ when we know all the correct parameters for the trial.

Recall previously in section 2.4, we described information based approaches that are used to revise design assumptions such as statistical variance or event rates based on blinded interim results. (Recall the example in Appendix B.) Study design 3 describes the setting whereby blinded sample size revisions are made based on either the pooled event rates or aggregate sample variance to maintain our statistical information by revising n to $\widetilde{n}$ to detect the same design alternative and maintaining statistical power. If one truly knows the variance of the treatment effect/event rates, the study design would be optimal in terms of statistical power to detect the alternative as well as maintain overall Type 1 error. As such, most trial designs may in fact be powered somewhere between design 1 and 2.

Study design 4 is related to the setting whereby an unplanned, adaptive element is incorporated into design 1 with the intention to possibly modify statistical information based on unblinded interim results. In this setting, an adjustment using Cui et al. [1999] has to be applied to control for any potential inflation of the overall Type 1 error as demonstrated by Proschan and Hunsberger [1995].

3.2.4 Statistical Criterion for Comparison across Designs

In experimental design, the optimal design is one that is superior to other existing study designs with respect to some statistical criterion. It is thus of interest for one to compare designs that are optimal when "bias and variance are minimized, efficiency is maximized, and cost is reduced" [Sanchez, 2014]. To achieve the multiple goals during clinical trial planning, the optimal design has to address the competing goals of science and ethics.

To do so, we tend to focus on comparing designs with similar ASNs since this is very often the limiting factor that affects patient accrual as well as the overall cost of the trial. After holding the average sample size fixed, we can isolate what constitutes a design that is optimal according to some criteria (in terms of operating characteristics, stopping rules) vs one that may be sub-optimal. The design with similar ASN and operating characteristics that maximizes the efficiency will be most important to clinical investigators.

In group sequential testing that allows early stopping, we can characterize the operating characteristics of a design/test with respect to the distribution of sample sizes at the time of study termination. Often this distribution is characterized by the expected number of subjects accrued prior to study termination, the average sample number (ASN). Other summary measures of the sample size distribution such as median, 75th percentile, or 90th percentile may be more useful in other situations. We may thus write the distribution of the sample size as a function of the stopping boundaries and the value of the true mean θ. This may be written as $F_N(n;\theta) = \sum_{j:N_j \leq n} \Pr(M = j|\theta)$ where M is the analysis time. A consequence of sampling variation (dropouts or recruitment) or conducting interim analyses on the calendar time is that we may observe $n_1^{'}, n_2^{'}$ rather than n_1, n_2 in practical settings.

3.2.5 Simulation Study

We consider the setting when there is no early stopping. To compare across designs, we hold fixed the timing of this interim analysis as well as the ASN. We consider the class of designs with similar ASNs and thus constrain all flexible adaptive designs to have some probability

p of modifying the final sample size. A numerical search can be performed to maximize the overall power of this class of flexible adaptive design. To examine specifically the loss/gain in power, numerical simulations were performed to compare between blinded vs unblinded designs.

We simulated 1,000,000 clinical trials, each with a sample size of $n = 100$ subjects equally randomized into two groups, with 90% power to detect the design alternative of $\theta = 0.4584195$, and known variance $\sigma^2 = 0.5$ for each group. Let p be the probability of adapting to a final sample size of $\widetilde{n}$. Thus, $1-p$ is the probability of staying the course with a final sample size of n. The ASN is $n(1-p) + p\widetilde{n}$. We let the adaptation be conducted at an interim analysis corresponding to kn for $k = \{0.1, 0.2, \cdots, 0.9\}$.

We optimize the overall power for this fully, flexible adaptive design using a grid search that has some probability p of adapting under the same design alternative. We then adjust the critical value based on Cui et al. [1999] to control the overall Type 1 error in this fully adaptive setting when using the unblinded treatment results to make an adaptation.

Simulation/Optimization algorithm

1. Simulate $S_1, S_2^*,$ and $\widetilde{S}_2^*$ from $N(\theta n_1, n_1), N(\theta n_2^*, n_2^*),$ and $N(\theta\widetilde{n}_2^*, \widetilde{n}_2^*)$ respectively.

2. Compute $Z_1, Z_2, \widetilde{Z}_2$ via $Z_1 = S_1/\sqrt{n_1}$, $Z_2 = (S_1 + S_2^*)/\sqrt{n_1 + n_2^*}$, and $\widetilde{Z}_2 = (S_1 + \widetilde{S}_2^*)/\sqrt{n_1 + \widetilde{n}_2^*}$ respectively.

3. Compute the CHW critical region at stage 2 regardless of whether an adaptation was made.

4. Search for the adaptive rule that maximizes the overall power subject to a fixed known probability p of adapting based on CHW.

5. Compute the respective (unadjusted) power based on this adaptive rule in 4 that has some known probability p of adapting.

6. To obtain the true (adjusted) power such that this optimal rule controls the overall Type 1 error, we recalibrate the usual critical value z_α to $z_\alpha^\ddagger$ under the null to fix α. We then recompute the power (adjusted) of this prespecified rule using the new critical value $z_\alpha^\ddagger$.

The optimal rule obtained from part (4) of the above algorithm is used to prespecify Study Design 3. This then allows us to obtain the corresponding power of an adaptive design based on minimal sufficient statistic when the study terminates.

Commonly proposed procedures in the adaptive literature often allow modifications to the sample size at slightly less than one-half of the sample size. We explore the setting where we adapt to a smaller sample size at some interim analysis by finding the optimal adaptive rule that maximizes the overall power. We then prespecify this adaptive rule so that we can use the minimal sufficient statistic at the end of the trial. Having held fixed the adaptive rule, we can then evaluate the potential loss in power when we decide on either an early or late adaptation relative to the original design. We describe results when adapting to a smaller sample size and refer the interested reader to Appendix C for additional results.

3.2.6 Simulation Results for Adapting to a Smaller Sample Size

Consider the setting when we adaptively decrease the total number of subjects in the study to 50 based on some known probability of $p = 0.5$ (blue lines of Figure 3.8). At early adaptations, we do not obtain much improvement in overall power when we prespecify the adaptive rule from this grid search. For example, at 10% of the sample size of the original design, when we adapt to a smaller sample size based on the best sampling scheme in a fully adaptive design, the overall power is 78.76% (Table 3.2, 3.3) which is negligibly lower ($\sim$0.7%) relative to an adaptive design using minimum sufficient statistics. This indicates that at early interim analyses, our estimates may be so unreliable that any form of adaptation is purely random. In such situations, the remaining weights that are not used is reallocated to a smaller number of subjects as compared to the original second stage sample size.

Consider the same scenario where instead we decide to make a late adaptation with the same known probability of 50%. Relative to an early adaptation based on a fully adaptive design, a late flexible adaptation has negligible gain in power. However, when we prespecify the best flexible adaptive rule and thus eliminate the need to use weighted statistics, we obtain a substantial gain in overall power (represented by the blue solid line) relative to

the power based on a late flexible adaptation. This is because, as we gather more reliable information about our treatment estimate to effectively make the right adaptation to stop the trial with a smaller number of subjects, we need to pay a much bigger penalty for having make this late adaptive look based on CHW, leading to a bigger loss of power (7.6%). As the probability of adaptation increases and the adaptation is made later in the study, the loss of power becomes pretty substantial when we compare between the prespecified adaptive design and the fully adaptive design.

Consequently, this means that when we obtain more reliable information about the estimate of the treatment effect, we should be making smarter adaptations if we prespecify the rule. However, by choosing to use weighted statistics, i.e., to make unplanned, data adaptive looks to modify our design, we lose more power since the inefficiency of the weighting scheme leads to a substantial penalty paid for the use of this late unplanned look. This in turn translates to a huge loss of efficiency as quantified by the loss of power relative to prespecified adaptations.

We can compute the assumed sample size for a fixed sample design based on the power obtained from the flexible strategy vs the prespecified strategy in column $CHW^{\ddagger}$ and $Adj^{\triangle}$ (Table 3.2 and 3.3) respectively. By doing so, the sample sizes computed in columns $SS^{\ddagger}$ and $SS^{\triangle}$ allow us to contrast the benefit of what would have been the required sample size had we preplanned the entire study envisioning this amount of power was required. At early interim analyses, the prespecified adaptive rule provides some gain in statistical information (up to 2.5% gain) over an unplanned fully adaptive design at halfway through the study. However, when this prespecified adaptation is made later, the efficient weighting scheme can provide up to 25% gain in overall power over the use of a fully adaptive rule.

In our grid search, we also considered other probabilities of adaptations (30%, 50%, 80%, 90%). We did not find a setting for which the unplanned adaptive rule provided additional benefit in terms of achieving higher statistical power relative to a prespecified rule when choosing to decrease the final sample size. Across the different probability of adaptations used, the relative loss of power is generally larger when late adaptations are made.

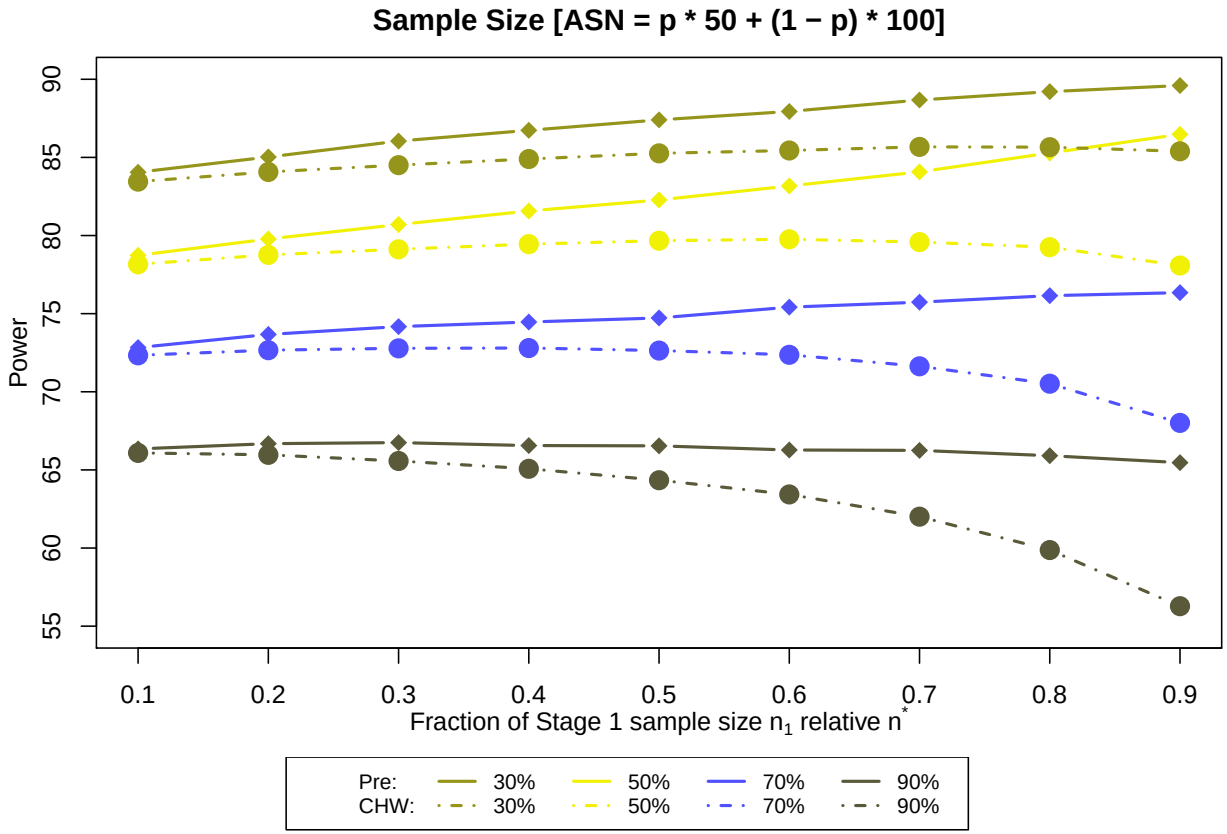

Figure 3.8: Plot of the overall power for adaptive design based on minimal sufficient statistics (Pre) vs the use of weighted statistics (CHW) when we consider various probabilities of decreasing the sample size from $n = 100$ to 50. The prespecified adaptive design has consistently higher power across various probability of decreasing the final sample size compared to the fully adaptive design requiring further adjustments using CHW.

Table 3.2: Simulation summary under the setting of decreasing the total sample size by half based on the probability of adaptation p where ASN $= 50p + 100(1-p)$. The original design has at least 90% power to detect the design alternative of $\theta \geq 0.4584195$.

						n=100	$\widetilde{n} = 50$				Overall Power				Sample size (FSD)**		
	n_1	n_2^*	$\widetilde{n}_2^*$	γ	RE	Orig	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	SS‡	SS$^{\triangle}$	SS$^{\triangle}$/SS‡
$p = 30\%$	5	95	45	0.05	1.028	89.99	62.53	62.94	62.89	1.006	83.45	84.74	84.05	1.007	81.82	83.19	1.017
	10	90	40	0.11	1.062	89.98	62.14	63.01	62.74	1.01	84.07	85.93	85.03	1.011	83.23	85.51	1.027
	15	85	35	0.18	1.107	89.99	61.53	63	63.04	1.024	84.51	86.91	86.04	1.018	84.26	88.08	1.045
	20	80	30	0.25	1.167	90.02	60.91	63.1	63.1	1.036	84.9	87.75	86.73	1.022	85.2	89.92	1.055
	25	75	25	0.33	1.25	90.03	60	62.96	62.9	1.048	85.26	88.49	87.4	1.025	86.09	91.79	1.066
	30	70	20	0.43	1.375	89.98	58.88	62.98	62.93	1.069	85.44	89.06	87.94	1.029	86.55	93.35	1.079
	35	65	15	0.54	1.583	90.04	57.37	62.96	62.93	1.097	85.67	89.65	88.67	1.035	87.13	95.6	1.097
	40	60	10	0.67	2	90	55.24	63.09	63.13	1.143	85.66	89.99	89.21	1.041	87.08	97.31	1.117
	45	55	5	0.82	3.25	90.03	51.48	63	62.94	1.223	85.39	90.15	89.6	1.049	86.42	98.62	1.141
$p = 50\%$	5	95	45	0.05	1.028	89.98	62.6	62.98	63.08	1.008	78.16	79.78	78.73	1.007	71.33	72.35	1.014
	10	90	40	0.11	1.062	90.08	62.07	62.97	63.03	1.015	78.76	81.31	79.78	1.013	72.4	74.27	1.026
	15	85	35	0.18	1.107	89.99	61.53	62.97	62.91	1.022	79.12	82.47	80.7	1.02	73.05	76.06	1.041
	20	80	30	0.25	1.167	90.02	60.85	62.97	63.07	1.036	79.45	83.65	81.57	1.027	73.66	77.8	1.056
	25	75	25	0.33	1.25	90.03	60.04	63.04	62.97	1.049	79.67	84.79	82.28	1.033	74.07	79.26	1.070
	30	70	20	0.43	1.375	90.02	58.92	62.99	63.06	1.07	79.76	85.91	83.17	1.043	74.25	81.2	1.094
	35	65	15	0.54	1.583	89.95	57.37	62.98	62.87	1.096	79.58	87.03	84.07	1.056	73.91	83.24	1.126
	40	60	10	0.67	2	90.01	55.2	63.04	63.19	1.145	79.25	88.33	85.28	1.076	73.3	86.14	1.175
	45	55	5	0.82	3.25	90.01	51.4	62.96	63.06	1.227	78.08	89.69	86.48	1.108	71.19	89.23	1.253

Sample size (FSD)** is the average sample size of a FSD for each of the respective approaches based on the power under column "Overall Power" attained.
RE: (Conditional) Relative efficiency based on the relative variances of the flexible adaptive design to the pre-specified adaptive design.
CHW‡: Power after adjusting for the unplanned adaptation.
Unadj: Power computed based on naïve overall Type 1 error of $\alpha = 0.025$.
Adj$^{\triangle}$: Adjusted power for fixed overall Type 1 error of $\alpha = 0.025$.
SS‡: Sample size of a FSD based on the power obtained using CHW.
SS$^{\triangle}$: Sample size of a FSD based on the power obtained based on the (Adj)usted test for fixed overall Type 1 error of $\alpha = 0.025$.

Table 3.3: Simulation summary under the setting of decreasing the total sample size by half based on the probability of adaptation p where ASN $= 50p + 100(1-p)$. The original design has at least 90% power to detect the design alternative of $\theta \geq 0.4584195$.

						n=100	$\widetilde{n} = 50$				Overall Power				Sample size (FSD)**		
	n_1	n_2^*	$\widetilde{n}_2^*$	γ	RE	Orig	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	SS‡	SS$^{\triangle}$	SS$^{\triangle}$/SS‡
$p = 80\%$	5	95	45	0.05	1.028	90	62.59	62.99	63.06	1.008	69.31	70.69	69.76	1.006	57.81	58.41	1.01
	10	90	40	0.11	1.062	89.97	62.19	63.07	63.2	1.016	69.44	71.7	70.25	1.012	57.98	59.09	1.019
	15	85	35	0.18	1.107	90	61.54	63.05	63.13	1.026	69.32	72.53	70.46	1.016	57.83	59.36	1.027
	20	80	30	0.25	1.167	90	60.89	63.03	62.89	1.033	69.09	73.22	70.58	1.022	57.52	59.54	1.035
	25	75	25	0.33	1.25	90.01	60	63.06	63.03	1.05	68.68	73.93	70.84	1.031	56.98	59.89	1.051
	30	70	20	0.43	1.375	89.99	58.97	63.06	63.08	1.07	68.03	74.52	70.84	1.041	56.12	59.9	1.067
	35	65	15	0.54	1.583	90	57.33	62.98	62.9	1.097	66.9	75.09	70.79	1.058	54.68	59.82	1.094
	40	60	10	0.67	2	90.01	55.17	63	62.93	1.141	65.29	75.69	70.67	1.082	52.69	59.66	1.132
	45	55	5	0.82	3.25	89.98	51.48	62.98	62.98	1.223	62.23	76.08	70.5	1.133	49.1	59.43	1.210
$p = 90\%$	5	95	45	0.05	1.028	90.01	62.63	63.04	62.99	1.006	66.09	67.12	66.34	1.004	53.67	53.99	1.006
	10	90	40	0.11	1.062	90.01	62.19	63.06	63.2	1.016	65.97	67.73	66.68	1.011	53.52	54.41	1.017
	15	85	35	0.18	1.107	90.04	61.59	63.02	63.18	1.026	65.57	68.05	66.75	1.018	53.04	54.49	1.027
	20	80	30	0.25	1.167	90.01	60.91	63.05	63	1.034	65.07	68.42	66.56	1.023	52.44	54.26	1.035
	25	75	25	0.33	1.25	89.96	60.01	63.01	63.15	1.052	64.34	68.6	66.54	1.034	51.56	54.23	1.052
	30	70	20	0.43	1.375	89.98	58.96	63.04	62.95	1.068	63.43	68.79	66.27	1.045	50.49	53.91	1.068
	35	65	15	0.54	1.583	89.96	57.4	63	63.13	1.1	62.01	68.85	66.25	1.068	48.85	53.87	1.103
	40	60	10	0.67	2	89.97	55.2	63.03	63.08	1.143	59.87	68.76	65.91	1.101	46.48	53.45	1.150
	45	55	5	0.82	3.25	89.96	51.48	63.04	62.88	1.222	56.28	68.58	65.47	1.163	42.7	52.91	1.239

Sample size (FSD)** is the average sample size of a FSD for each of the respective approaches based on the power under column "Overall Power" attained.
RE: (Conditional) Relative efficiency based on the relative variances of the flexible adaptive design to the pre-specified adaptive design.
CHW‡: Power after adjusting for the unplanned adaptation.
Unadj: Power computed based on naïve overall Type 1 error of $\alpha = 0.025$.
Adj$^{\triangle}$: Adjusted power for fixed overall Type 1 error of $\alpha = 0.025$.
SS‡: Sample size of a FSD based on the power obtained using CHW.
SS$^{\triangle}$: Sample size of a FSD based on the power obtained based on the (Adj)usted test for fixed overall Type 1 error of $\alpha = 0.025$.

3.2.7 Summary

Other scenarios such as increasing the sample size at different interim analysis in a two-stage setting were also explored. These results are in Appendix B. In these situations, where we choose to increase the sample size based on such flexible adaptation, there is slight but negligible gain in overall power relative to the prespecified adaptive design. This occurs at very early interim analyses. When early adaptations are made to decrease the sample size based on a fully flexible rule, there is generally negligible gain in overall power as compared to a prespecified adaptive design. However, when late adaptations are considered to either increase or decrease the sample size, there is generally substantial power loss when using a flexible adaptive rule as compared to having prespecify the same rule at design stage.

In summary, unplanned adaptive rules to decrease sample size can come at a substantial loss of power when these adaptations are made late during the study. When adaptations are made early, this loss of power does not appear as large relative to the power loss as a consequence of late adaptations. Even though our simulation study is focused on the fixed sample setting, the results in this section provide some intuition on the impact of unplanned adaptive rules on the overall power of the trial. Later, in the censored time to event setting, we are interested in adaptations made in the presence of monitoring rules that allow early stopping. In the time to event setting, certain situations may favor an adaptation that is conducted at early interim analyses.

3.3 Impact of Interim Analyses on Efficiency of Group Sequential Designs

Adaptive designs with the aim of increasing the maximum statistical information are often compared with respect to GSDs by adding more interim analyses so as to beat the operating characteristics of the adaptive design in an unfair manner. Currently, few authors have attempted to separate out what defines a good/bad sampling scheme in presence of such adaptations when the number of interim analyses is held fixed. Since GSD is a special case

of prespecified adaptive design, it is of use to consider the properties of GSD to understand how prespecified adaptive design may behave later when we choose to modify the maximum statistical information. Earlier on, we see that the analysis schedule for an adaptive design may change when an adaptation is made. This modification of the maximum statistical information can affect the degree of early conservatism when judged by the maximal sample size the revised monitoring boundary. To better understand this, we now describe the difficulty by characterizing how changing the schedule of analyses based on a GSD can affect other operating characteristics as defined using the ASN.

Several authors have investigated finding the best sequential design based on some optimality criterion. Jennison [1987] considered group sequential tests that minimize the sample size based on several competing null and alternatives of interest. Extensions to finding efficient sequential designs were later formulated using a Bayesian decision theoretic framework by placing point mass priors on alternatives of interest [Eales and Jennison, 1992, Barber and Jennison, 2002]. Eales and Jennison [1992] found negligible improvement in ASN when choosing between non optimal vs optimal schedule of analyses.

For scientific reasons, we often choose the OBF design since we are often reluctant to inflate our maximal sample size relative to a FSD, and we want to maximize safety information in the presence of more modest treatment benefits. However, in an adaptive design, as we have seen from the schizophrenia example, when we increase our maximum statistical information, we no longer maintain the same degree of early conservatism.

We highlight some of the potential issues by considering a two-stage GSD with interim stopping. As more interim analyses are added, the optimality of the monitoring schedule becomes less clear. Results for the three-stage GSDs were also investigated and we also refer the interested reader to Appendix B.

3.3.1 Characterizing the ASN of Two-stage Designs

We consider a 2-stage GSD under the unified family framework. At level $\alpha = 0.025$, we assume a FSD that has power $\beta = 1 - \alpha$ to detect the alternative of 0.1, and a common

known standard deviation of 1. Such a design requires a total sample size of 6,146 subjects. We enumerate all possible stopping rules within the unified family framework by allowing P to vary, while setting A and R to 0 to search for the critical value G. Using ASN as our optimality criterion, we characterize the ASN contours under the alternative among the class of one-sided symmetric, two-sided symmetric, and hybrid GSDs (using a fixed OBF efficacy boundary while varying the futility boundary). We note that there is a wider variety of designs that may be considered, and that the designs that are found to be "best" according to our constraints may no longer be "best" when we expand the search to a bigger class of designs, or consider other more complex optimality criterion. However, these 3 classes of GSDs are sufficient to illustrate our point.

We can find the "best" design within each class that has minimal ASN within this bivariate space characterized by the function of the schedule of analyses and the parameter space for $P > 0$. The contour plots of this bivariate space for each class of designs explored are shown in Figure 3.9. The blue points (in the top row) show the global minimum within this bivariate space. We describe the results for the one-sided symmetric designs (Table 3.4). Among all possible combinations of P and schedule of interim analyses, the best one-sided symmetric GSD occurs when $P \approx 0.54$ with a minimum ASN of 4213 conducted at roughly 42.7% of the maximum statistical information (Table 3.4). This value of P is close to a Pocock design.

In this space, we can also characterize the ASN among all equally spaced interim analyses for each $P > 0$. Alternatively, we can fix the spacing of the schedule of interim analyses to characterize the ASNs across various P parameters. Likewise, such characterization can be done by holding P fixed. Of note, $P < 0.5$ tends to be less efficient with respect to a Pocock rule ($P = 0.5$) and thus we choose to focus on describing the discretized space for $P \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$.

The Pocock design with an equally spaced analysis schedule has an ASN of 4,325, and a maximum sample size 20.2% higher than that of a FSD. Had we chosen the optimal schedule of interim analyses that is conducted at 42.9% of the total statistical information using

the Pocock boundary, then we obtain an even lower ASN of 4218, at the cost of inflating our maximum statistical information to 22.2% relative to the FSD. An OBF design, on the other hand, with an equally spaced schedule of analyses has an ASN of 4,662 (higher than the optimal Pocock design), and a maximum sample size of only 1.3% higher than the sample size of a FSD. This maximum statistical information is substantially lower than the Pocock design with either the equally spaced analyses, or optimized spacing of analyses. The optimal schedule of analyses for the OBF design takes place at 58.3% of the total statistical information, which attains a minimum ASN of 4,509 at the cost of a slight inflation of the maximum sample size relative to the FSD.

In practice, however, we often choose P to reflect the degree of early conservatism in order to balance scientific, ethical and efficiency concerns. Thus, one may alternatively want an optimal schedule of analyses for some fixed P that minimizes the ASN. However, given this best schedule of interim analyses, this bivariate space is poorly behaved in the sense that one can find some other parameter P within the same class of design to obtain an even lower ASN based on this same schedule of interim analysis. Each of these optimization results are considered local minimums (relative to the global minimum) since we only hold fixed either the schedule of analyses or the P parameter. This aspect is illustrated in the bottom row of Figure 3.9 where each profile curve for the ASN corresponds to a particular choice of early conservatism for $P \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$.

For example, we can consider the symmetric one-sided OBF design. The optimal schedule for the two-stage OBF design has an interim analysis conducted at 58.3% of the maximum sample size (ASN=4520). This increases the maximum sample size by 3% relative to a FSD. However, for this same schedule of interim analysis, the design with $P = 0.76$ beats the ASN for this two-stage OBF design, with a slightly higher maximum sample size (6.9%) relative to the FSD.

In general, relative to the FSD, the ASN decreases marginally as P decreases over the investigated range of P. Similar trends are observed for the two-sided symmetric designs. We see that for the one/two-sided symmetric monitoring rules, the contour plots are pretty

similar. The ASN tends to increase as the timing of the interim analysis approaches 1.

Table 3.4: Optimal spacing of analysis schedule for designs with a total of two analyses. A FSD having 97.5% power requires $N = 6146$ to detect a design alternative of $\theta = 0.1$.

		Equally Spaced			Holding P fixed				Holding $\mathcal{I}_1^P$ fixed			
	P	ASN	Max	$\frac{\text{Max}}{N}$	ASN^P	Max^P	$\frac{\text{Max}^P}{N}$	$\mathcal{I}_1^P$	P_{Opt}	ASN_{Opt}	Max_{Opt}	$\frac{\text{Max}_{\text{Opt}}}{N}$
One-Sided	0.5	4325	7387	1.202	4218	7510	1.222	41.4	0.53	4214	7328	1.192
	0.6	4266	6966	1.133	4222	6973	1.134	44.7	0.56	4217	7149	1.163
	0.7	4277	6660	1.084	4272	6650	1.082	48.1	0.61	4241	6933	1.128
	0.8	4352	6447	1.049	4346	6459	1.051	51.6	0.66	4284	6773	1.102
	0.9	4483	6309	1.026	4432	6348	1.033	55.3	0.71	4348	6657	1.083
	1	4662	6226	1.013	4520	6277	1.021	58.3	0.76	4412	6568	1.069
	Opt				4213	7262	1.182	42.7	0.54			
Two-Sided	0.5	4137	6683	1.087	4078	6735	1.096	42.9	0.41	4059	7016	1.141
	0.6	4159	6512	1.059	4137	6518	1.060	45.9	0.44	4073	6875	1.119
	0.7	4222	6382	1.038	4220	6380	1.038	49.2	0.48	4109	6740	1.097
	0.8	4326	6287	1.023	4316	6295	1.024	52.6	0.51	4166	6653	1.082
	0.9	4473	6224	1.013	4414	6243	1.016	55.8	0.56	4244	6552	1.066
	1	4659	6185	1.006	4509	6211	1.010	58.8	0.61	4320	6479	1.054
	Opt				4059	7019	1.142	41.2	0.41			
Hybrid (Fut, Eff)	(0.5, 1)	4776	6742	1.097	4648	6714	1.092	57.8	1.20	5091	6166	1.003
	(0.6, 1)	4734	6565	1.068	4605	6577	1.070	57.6	1.19	4896	6175	1.005
	(0.7, 1)	4702	6430	1.046	4572	6467	1.052	57.8	1.20	4731	6185	1.006
	(0.8, 1)	4681	6333	1.030	4548	6380	1.038	57.6	1.24	4607	6195	1.008
	(0.9, 1)	4668	6267	1.020	4531	6322	1.029	58.2	1.30	4531	6206	1.010
	Opt				4508	6214	1.01	58.8	1.41			

"Fut" refers to the futility boundary and "Eff" refers to the efficacy boundary.
Equally spaced: The ASNs for GSDs with equally spaced analyses are evaluated for each specific P parameter.
By ^{P} : For each P, we obtained the GSD with the schedule of analyses that minimizes the ASN. Using this schedule of interim analyses based on the fixed P, we can find another GSD with some other parameter P' that may minimize this ASN^P further.
Opt: The optimal GSD with the best analysis schedule with the minimum ASN among $P > 0$.

The contour plot for the class of one-sided hybrid design with a fixed OBF efficacy parameter behaves relatively differently. In general, such a design constraining the efficacy

boundary using the OBF rule is optimal when the futility boundary is further away from 1. Within this specific class of designs, the best GSD has an interim analysis conducted at $\approx 58.8\%$ of the maximum statistical information with the futility parameter $P = 1.41$. For other choices of futility parameter P, the best schedule of interim analysis is generally conducted at $\approx 58\%$ of the maximum statistical information. Within the class of one-sided hybrid design having an OBF efficacy parameter and futility parameter of $P > 0.7$, there is generally mild inflation of the maximum statistical information relative to the FSD. This specific class of design has comparable ASN properties as the class of one-sided symmetric designs with $P \in (0.7, 1)$. In comparison, it may be worthwhile to consider exploring the smaller class of asymmetric designs, i.e., $\{P_{\text{futility}}, P_{\text{efficacy}}\} \in (0.7, 1) \times (0.7, 1)$ to potentially gain more efficiency in terms of ASN under the null while maintaining statistical power to detect the same design alternative.

3.3.2 Implications in the Time To Event Settings

We investigated the scenario of characterizing the optimal design using ASN as the optimality criterion while holding other aspects of the operating characteristics constant. By varying the spacing of the interim analyses, we observed the following:

- When we plan our study assuming equally spaced analyses based on some fixed parameter P, our ASN may not be best compared to other potential schedule of analyses under the alternative.

- The optimized schedule of interim analyses for a fixed P can however not be optimal with respect to another design with the same schedule of analyses.

- The difficulty of characterizing the best or even better optimal design is made more difficult as we expand the class of symmetric designs to incorporate asymmetric designs. Further complications include optimizing the schedule of analyses.

It is typical that in practice we seldom pick the design that minimizes the ASN since other scientific, logistical constraints may dictate the choice of the monitoring rule. However, it is important to note that when evaluating the operating characteristics of a GSD, one should

take into consideration other potential, competing designs so as to carefully compare various operating characteristics such as ASN. As illustrated, we see that some of the less known choices of the GSDs within the unified family, with a different schedule of analyses, can lead to improvements in ASN relative to common choices of sequential designs.

3.4 Summary

In adaptive designs where we may potentially switch monitoring rules as a consequence of adapting the maximum statistical information, we need to be cautious in understanding whether our ability to stop the trial early may change as a consequence of this switching. For example, when the test statistic attained 90% of the statistical information at 3/4 of the way through the study, we may have a tendency to stop the trial sooner when we have reliable safety data to establish the safety/futility of the treatment. On the other hand, at 3/4 of the way through the study, having only 30% of the statistical information may lead us to act differently. In such a situation, it is possible that we may want to explore a more conservative boundary to match our level of expectation. We now describe the advanced background issues in the time to event setting.

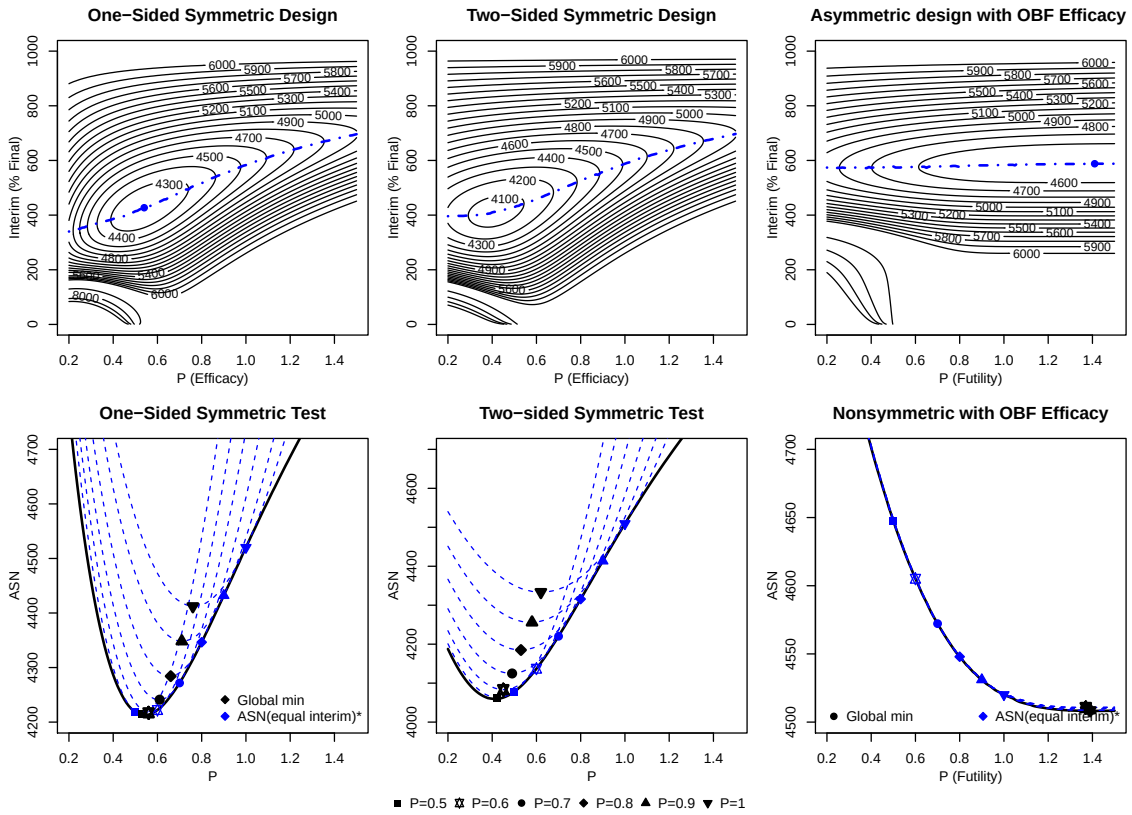

Figure 3.9: Top row: Contour plots of ASN for relative timing of interim analyses vs P for group sequential monitoring boundaries corresponding to one-sided symmetric, two-sided symmetric, and one-sided hybrid boundaries. Each of these designs has a total of two analyses, statistical power of 97.5% to detect the design alternative of 0.1, and known variance of 1. The blue dash-dotted line describes the ASN based on a equally spaced interim analyses for each fixed P. Bottom row: The blue dotted lines characterize the valley of local minimums of ASN (the minimum is represented by the black point) for a fixed level of early conservatism P based on the best schedule of interim analyses. The blue points correspond to the ASN with equal schedule of analyses which may not be the local minimum relative to the black point with the same symbol.

Chapter 4

Background: Advanced Issues in the Time To Event Setting

In this chapter, we describe some of the statistical themes and the less common statistical methods in the time to event setting used in this dissertation.

4.1 Censoring Distribution

We measure the study time of an individual on the trial from the time of randomization to the last follow-up (either observing the event of interest, lost-to-follow-up, or censored administratively at time of the interim analysis). Different time scales are of relevance in the time to event setting: study time, accrual time, calendar time, and the information time.

The study time is of scientific interest to the regulatory bodies and clinicians in determining whether sufficient evidence has been obtained about the treatment thought to improve aspects of the patients' well-being. This is typically measured from the time of randomization of an individual to the last observed time (either due to lost of follow-up, event of interest, or still alive at the calendar time). For example, a short study time in establishing the efficacy of a treatment may not be clinically important when the subjects in the study have only been followed up for an average of 3 months as compared to a longer study with at least 3 years of follow-up. Once all the outcomes have been observed, the study may be extended to obtain additional safety information that are less relevant to answering the primary question of interest. We are typically less interested in this additional follow-up for our primary objective once we acquire all number of events are reached or if the trial is

stopped early for efficacy, futility, or lack of benefit. However, we may be interested in this additional follow-up to collect more safety information.

The accrual time refers to the active period of time whereby subjects are enrolled into the study and randomized into the treatment groups. Recruitment typically occurs by enrolling patients at multiple sites. The rate of accrual depends on multiple factors such as the number of sites, the number of patients the sites can handle, as well as the disease prevalence/incidence. For instance, the accrual period in prevention trials may often be brief when participants can be recruited immediately. In rare disease settings, this period may take a longer time. Logistical constraints, however, can affect the number of participants a site can feasibly enroll. New sites may not be experienced with the recruitment of patients into the trial, but may over the course of time develop experience in identifying potential patients to be recruited over time. For example, in certain chronic disease settings, the early patients who are enrolled into the trial may more likely be the prevalent cases or patients who failed other treatment regimens. As these prevalent cases are exhausted over time, the recruitment rate may slow down as sites await for eligible incident and potentially healthier cases to be enrolled into the trial.

The information time provides a useful way to conceptualize the degree of early conservatism of the sequential monitoring rules at design phase. These stopping boundaries for the GSDs as seen in Chapter 2 are defined based on the information time, often characterized by the ratio of statistical information gathered at some interim analyses relative to the total statistical information at the final analysis. In adaptive designs, this information time, as seen based on the schizophrenia example in Chapter 3, is less well-defined as a consequence of the unknown maximum statistical information at design. A maximum duration study is conducted by allowing termination of the study when the pre-defined stopping time of the trial is met even when the amount of statistical information has not been gathered. When the study is terminated based on all the required number of events accumulated, we call this a maximum information study.

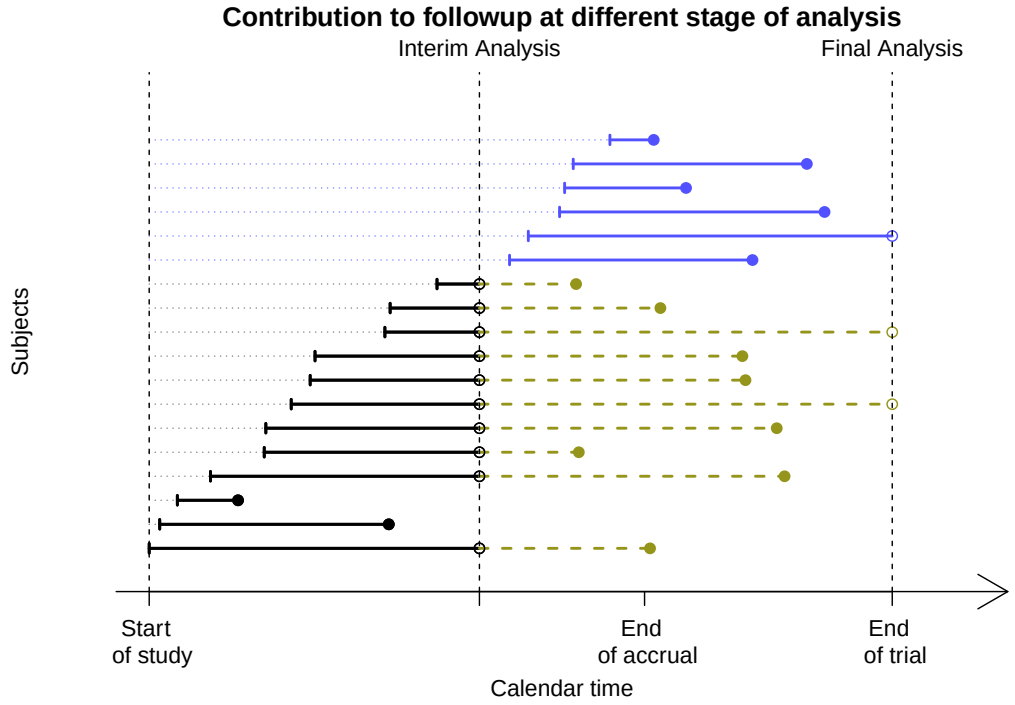

Figure 4.1: Patients are recruited in a staggered setting and followed until the disease of interest have been obtained. At some interim analyses when the accrual of subjects is incomplete, some participants (in black) may or may not have the event of interest. As they are followed further, those who are at risk at the interim analyses may develop the disease (in gold) as new participants are continuously recruited and randomized into treatment groups and followed (in blue).

Interim analyses are often conducted based on ethical, economic, and logistic concerns. It is very often the case that although the planning of GSDs or adaptive designs are based on the information scale, interim analyses are conducted based on the calendar time of the study. The calendar time most appropriately reflects the financial and logistical concerns of the sponsors or research institutes. It is viewed as some transformation of the information time and is of paramount importance in the application of sequential monitoring in settings with delayed ascertainment of outcomes.

The censoring distribution in a time to event analysis is defined to be some function of the accrual and survival distributions that can be described by the above time scales. With sequential monitoring, interim analyses truncate the survival distributions and create this "patient-wise" separation such that participants in the study may (1) be randomized and had the event of interest, (2) be randomized and are still currently followed but have not had the event of interest as observed in Figure 4.1. These different components are later seen to come indirectly into the estimation of the treatment effect in the most common time to event analyses.

4.2 Choice of Summary Statistic

In general, the summary statistic of choice in the time to event setting will depend upon assumptions about the actual distribution of the survival times in each group. At planning stage, it is of use to postulate plausible survival functionals that are clinically meaningful to the treatment of the disease. Often, the data gathered from prior trials are sufficient to enable us to establish the relevant alternative survival functionals of interest, and/or exclude functionals that do not reflect clinical/scientific importance but may plausibly surface during monitoring of a clinical trial. Having decided on the distributional functionals that capture our preferential ordering of treatments, and most appropriately reflect scientific importance and clinical benefit to patients, we now seek to choose summary measures that provide greatest statistical efficiency to quantify and distinguish/discriminate the functionals of interest.

Several summary statistics can be used to quantify/distinguish the survival functionals that represent most importance to us. We may choose to

1. Compare the vertical separation of the curves by considering the probability of surviving past any particular time point. *There are many ways that we may compare the difference in survival probabilities and the choice of a particular time point typically represents scientific/clinical importance.*

2. Compare the horizontal separation of the curves by considering an arbitrary quantile of the survival distribution. *The median survival is the most common quantile used for*

such comparison. Other potential quantiles may be used and the choice of the quantile is typically driven by scientific/clinical importance.

3. Compare the difference in area under the survival curves up to some specified time point/up to some specific survival quantile. *This can be interpreted as the average difference in the amount of time saved over a specified time period, which is often referred as restricted mean survival. Alternatively, we may compare the difference in area restricted to the quantiles of the survival distribution. This interpretation is thus more difficult to describe and additional restrictions may be required when the survival curves do not attain the required survival quantile.*

4. Compare the weighted average slopes of the survival curves relative to the survival probability at the particular time point. *The weights may be chosen based on statistical efficiency under some assumed (semi-)parametric model. The most common choice is the log rank test where the weights are efficient under Lehmann alternatives [Lehmann, 1953, Davies, 1971] and locally efficient when the hazards are proportional over time.*

Typically, when one survival curve is always greater than another (i.e., the distributions functions are stochastically ordered at all times) in the RCT setting, we have a clear idea which treatment we prefer. Any summary statistics based on the entire functionals that may be used as a basis for statistical inference will order the true survival curves accordingly to the preferred treatment. These summary statistics include any of the above choices [1-4], the medians, or other quantiles, survival probabilities at some pre-specified time, restricted means, or, with the censoring distribution commonly encountered in randomized clinical trials, a weighted hazard ratio as might be computed in Cox regression.

However, when the distributions of survival times are not stochastically ordered across treatment groups (i.e., when the true survival curves cross at some point in time), the definition of the "preferable" treatment is less clear. Factors that might have to be considered would include the time at which the survival curves cross, the survival probability at time of crossing, the relative advantage in mean survival for one treatment over the other prior to crossing (e.g., the area between the survival curves), the relative advantage in mean survival in the opposite direction after crossing (perhaps extrapolated into the future), and the patient population being treated (e.g., our interest in any extrapolation of treatment effect into the future might be more extensive for children than it would be for adults).

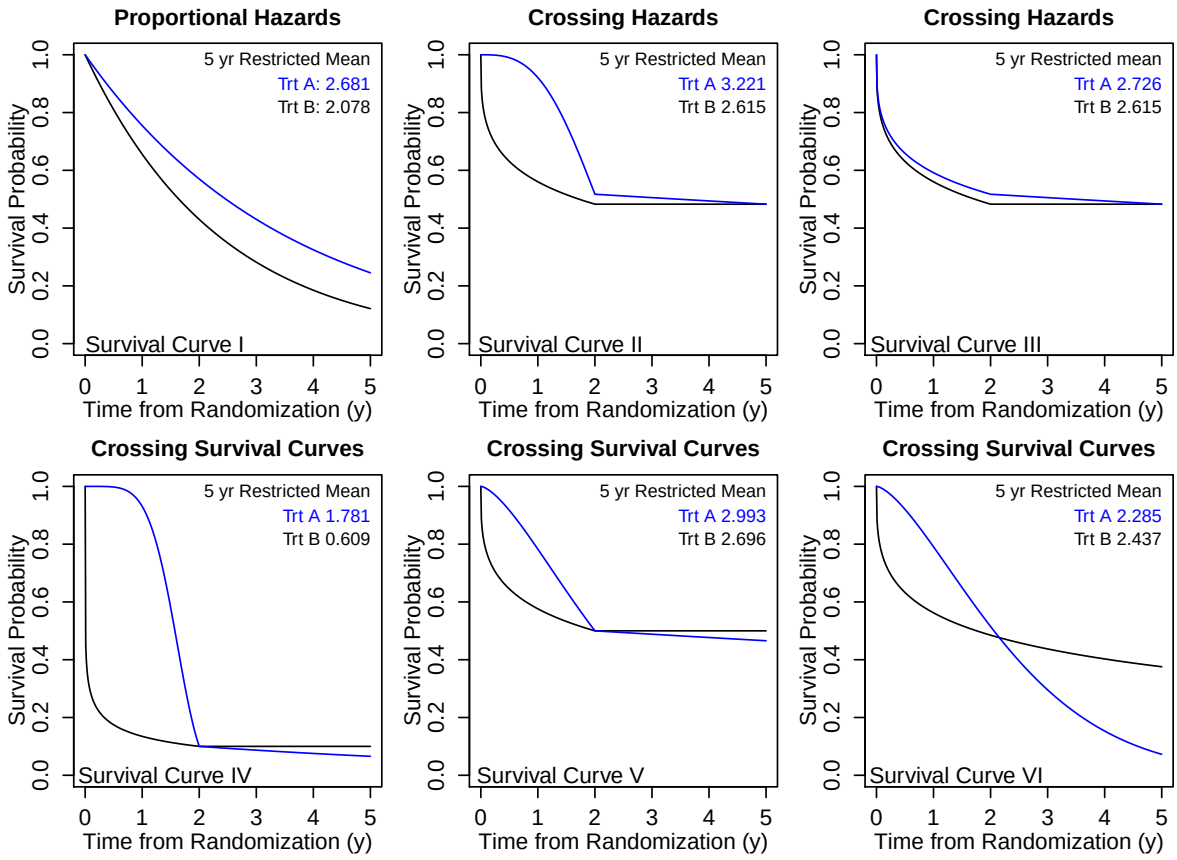

Figure 4.2: Curves (Truth) exhibiting the possibility of crossing hazards that may result in crossing survival.

To better illustrate how different survival functionals may affect our choice of summary statistics, we provided several non exhaustive alternative scenarios that represent different clinical benefit to patients in different disease settings as shown in Figure 4.2. The top row of the survival curves described the setting where we have stochastic ordering over the first 5 years. When the treatments' survival distributions are expected to exhibit proportional hazards as in the top left figure (and thus are stochastically ordered), all summary statistics such as the survival probability at any pre-specified time, any pre-specified quantile, restricted mean up to any pre-specified time will enable us to pick A as the "better" treat-

ment. The Cox proportional hazards model will tend to be most efficient in picking A to be the better treatment with the censoring distribution that commonly arise in RCT setting as a consequence of accrual.

In the other two settings, we may posit survival functionals that deviate from proportional hazards but ultimately exhibit stochastic ordering survival curves (as in the rest of the Figures on the top row) characterized by non proportional hazards with only "early differences" (as might arise with crossing hazards), or only "late differences" (as might arise with diverging hazards). Under these scenarios, treatment A will typically be identified as "better" if we choose to base comparisons on survival probabilities at any arbitrary time, on any arbitrary quantile, or the restricted mean up to any arbitrary time up to year 5.

With the log rank statistic, when accrual of patients over time naturally induce different censoring distributions (across interim analyses) as we analyze our data at different calendar times, the (unweighted) log rank statistic's weighting of the difference in hazard functions would tend to identify A as the same "better" treatment. Under censoring distributions that arise under "left entry" or with weights modified to emphasize later hazards, the conclusion drawn from the log rank test need not be consistent for the "better" treatment. Similarly, other weighted averages of the hazard functions could lead to inconsistent identification of the "better" treatment.

Crossing survival curves present both challenging scientific and statistical dilemmas. In the second row, it is not at all clear what is the preferred treatment when we have crossing survival curves. In practice, the decision in a RCT setting would at minimum depend on other quantitative measures such as

M1. The "degree of separation" between survival curves prior to crossing and after crossing

M2. The timing of the crossing

M3. The survival probability at which the curves cross

While additional information (such as toxicities, safety, etc) can be employed for decision making to quantify the better treatment, without loss of generality, these three quantitative

measures are sufficient for further elaboration on determining the choice of statistics for the primary endpoint.

This degree of separation between the survival curves prior to crossing might be judged simply by the vertical differences in survival curves. With minimal separation between the curves, there may be less dilemma in judging whether A or B is better. However, if a huge separation exists, then this may affect any tendency to base treatment decisions on the long term benefits, and many different decisions may be made depending on the disease setting.

The timing at which this crossing occurs may also have to be factored in to the treatment decision. To make a judgment based on any of those quantitative measures will depend upon the clinical setting and the objective of the trial. For instance, in childhood cancers, if this crossing occurs within the first year, we may gravitate more towards greater emphasis on the longer term survival than we might for older adults who may have similar survival curves.

Lastly, the survival probability at which this crossing occurs will affect the treatment decisions. In the survival curve IV of Figure 4.2, this crossing happened when the survival probability is 0.1, implying that the number of subjects receiving any benefit from the treatment with the better survival is thus fewer. Since this vertical separation is big, the disease setting and the age of the participants have to be factored in to determine whether this long term benefit is truly beneficial to patient survival. On the other hand, if this survival probability is relatively high, the decisions may be guided by any of the previous 2 points.

4.3 Consequences of Time Varying Treatment Effect

Consider a RCT comparing the use of a chemotherapy regimen vs a standard of care. Serious toxicity issues may affect the survival benefit earlier relative to the standard but if the patient recovers from the toxicity of the treatment, there may be some survival benefit as seen in Figure 4.3. In such situations, the unweighted logrank test may not be optimal to pick out the meaningful difference if we are interested in the survival benefit at year 5. The behavior of early treatment differences in such situations can be influenced by changes to the censoring distribution as might happen with interim analyses. Other time to event methods

may be preferred in presence of time varying treatment effect. However, the monitoring strategies chosen are often evaluated under the assumption of the strong null hypothesis of no treatment effect, thus presuming constant treatment effect across time. We must then try to evaluate the sensitivity of competing monitoring strategies to avoid picking out early/late differences that do not matter clinically when answering the scientific question. This is of particular importance in Chapter 7.

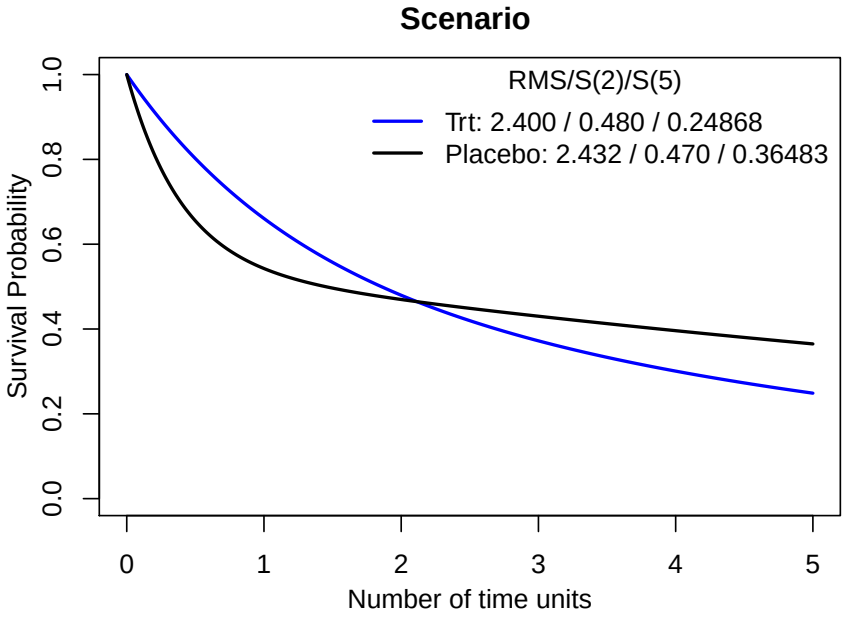

Figure 4.3: Crossing survival curves where prior to time 2, the treatment appear to be superior over the placebo. However, this effect wears off by time 2 and placebo is then superior over the treatment after time 2. The difficulty of picking the preferred treatment depends upon the clinical setting and personal preference of the patients.

The consequence of interim analyses induces a natural/administrative censoring of the follow-up time that further truncate what may be clinically relevant to the scientific question. Under such situations, when the survival curves appear to converge, clinical investigators may consider the plausibility that the early treatment effect has waned (i.e., survival curves start to converge), or that subsequent exposure to the treatment may be harmful (i.e., survival

curves may be expected to cross). We may thus be concerned with other analyses methods to better address the non PH settings as seen in the earlier examples (Figure 4.2) that all demonstrate some form of time-varying treatment effect (except for I).

4.4 Strong Null vs Weak Null

To facilitate investigation in the time to event setting, we set out to distinguish two different null hypotheses setting, namely, the strong null and the weak null hypothesis [Emerson, 2011]. We define testing the strong null hypothesis as $H_0^S : F(x) = G(x), \ \forall x$. By this definition, this would encompass equality of distribution for all moments such that the distance measure between F and G, $d(F, G) = 0$ [Rudser, 2007]. In the survival setting, we are interested in testing the exact equality of survival distributions which is, under this strong null hypothesis, naturally proportional hazards. The Kolmogorov-Smirnov test is an example of such consistent testing procedure of the strong null hypothesis.

The weak null hypothesis, H_0^W, can be heuristically defined such that the distance measure between $\widehat{d}(F, G) = 0$ but $F(x) \neq G(x)$ for some x. An example of such test of hypothesis can be seen when testing for difference in normal means. Let the density of F be $f \sim N(\mu_F, \sigma^2)$ and the density of G be $g \sim N(\mu_G, 2\sigma^2)$. Under testing of the weak hypothesis $H_0^W : \mu_F = \mu_G = \mu_0$, the means of the distribution of F, G are equivalent under the null hypothesis. However, the entire distribution F is not equivalent to G in the higher moments.

The notion of the weak null plays a pivotal role in enabling us to quantify the relative importance of the estimate of the treatment effect under the non proportional hazards settings. Under this definition, there is a rich collection of functionals where various test statistics characterize the weak null differently depending on the weighting scheme of these test statistics as well as the time frame of analysis.

In randomized clinical trials and many other studies, we characterize the operating characteristics of design under the (strong) null hypothesis of $H_0^S : S_A(t) = S_B(t), \forall t$ as analyzed by either the $G^{\rho,\gamma}$, weighted Kaplan-Meier, or Nelson-Aalen statistics. Our alternatives may be written as $H_A^S : S_A(t) \neq [S_B(t)]$, for some t, and $\theta \neq 1$. When $t \to \infty$, we obtain full

knowledge of the survival distribution, and thus all our test statistic are consistent tests of the strong null hypothesis with the correct size, α.

Most often, the above does not hold true when scientific questions are sufficiently addressed by a consistent test of the weak null hypothesis, i.e., $H_0^W : S_A(t) = S_B(t)$ as parametrized by using the average hazard ratio, or the average difference in survival truncated to some follow-up time of interest. In many clinical settings, incomplete follow-up as a consequence of limitations of resources or less emphasis on characterizing the full (later) survival distribution of the participants naturally truncate our test statistic up to some known maximum follow-up time of interest (i.e., we do not follow all patients in the clinical trial until all of them have the event of interest). Thus, we do not have full knowledge of the functional forms of the survival distributions [Emerson and Emerson, 2013].

4.5 Cox Proportional Hazards Regression/Log Rank Test

The hazard ratio is one of the most commonly used summary measures used to quantify this difference in hazards. It arises naturally based on the logrank test/Cox proportional hazards regression. Under the strong null and when presuming proportional hazards alternatives, the logrank statistic/Cox proportional hazards regression is the most efficient rank based test. In the presence of censoring, as defined either by incomplete accrual of subjects or interim analyses due to sequential monitoring, statistical information (the measure of the inverse of the variability of the log hazard ratio) is generally related to the number of events. Under the strong semi-parametric assumption of proportional hazards, adaptive modifications based on the primary endpoint with appropriate adjustments do not impact the scientific interpretation/credibility of the trial results.

4.5.1 Notation and Setup

For convenience, we describe the fixed sample setting by letting $J = 1$ and suppress the sequential notation for now. We also denote the analysis time to be defined on the calendar

time. Let E, T, C be the random variables corresponding to the calendar time of entry into the study, the study time for the event of interest, and the study time for loss-to-follow-up with the respective distribution functions H, F, and G.

Under the fixed sample setting with a total accrued sample of N subjects where the final analysis is conducted at calendar time τ, our data for the i^{th} subject can be represented in the form $(X_i,\, \Delta_i, Z_i)$ where $X_i = \max(\min(T_i, C_i, \tau - E_i), 0)$ is the observed time for individual i, $\Delta(X_i)$ is the indicator variable for an observed failure time if $X_i \leq \min(C_i, \tau - E_i)$ and 0 if $E_i > \tau$ and that loss of followup is only due to administrative censoring, and the randomized treatment assignment is

$$Z_i = \begin{cases} 0 & \text{if the } i^{\text{th}} \text{ individual belongs to treatment group 0} \\ 1 & \text{if the } i^{\text{th}} \text{ individual belongs to treatment group 1} \end{cases}$$

For notational simplicity later, we further let $\Delta_1(X_i) = \Delta(X_i)Z_i$ and $\Delta_0(X_i) = \Delta(X_i)(1-Z_i)$ where they are the indicators of failure for the i^{th} subject coming from group 1 and 0 respectively. Let $d_1(t) = \sum_{i=1}^{N} \Delta(X_i)Z_i 1_{[X_i \leq t]}$ denote the total number of events by time t where $1_{[.]}$ is the indicator function for treatment group 1. Thus, $d_0(t) = \sum_{i=1}^{N} \Delta(X_i)(1 - Z_i)1_{[X_i \leq t]}$ is the total number of events by time t for treatment group 0. We first describe the Score representation of the Cox proportional hazards regression model that is directly related to the logrank statistic.

The (partial) score statistic for the (unweighted) logrank statistic at some interim analyses written in both statistical/scientific interpretation, as evaluated at $\beta = 0$, is as follows

$$\mathcal{U}(\beta)|_{\beta=0} = \sum_{i=1}^{N} \Delta_i \underbrace{\left[Z_i - \frac{\sum_{l \in \mathcal{R}(X_i)} Z_l}{\sum_{l \in \mathcal{R}(X_i)}} \right]}_{\substack{\text{Statistical:} \\ \text{Estimating function}}} = \sum_{i=1}^{N} \frac{n_0(X_i) n_1(X_i)}{n_0(X_i) + n_1(X_i)} \underbrace{\left[\frac{\Delta_1(X_i)}{n_1(X_i)} - \frac{\Delta_0(X_i)}{n_0(X_i)} \right]}_{\substack{\text{Scientific: Weighted average} \\ \text{of difference in hazards}}}$$

where $\mathcal{R}(X_i) = \{l : X_l \geq X_i\}$ is the risk set at analyses time X_i, and $n_k(X_i)$ is the number at risk at analyses time X_i for the group k for $k = 0, 1$.

Under random censoring, this number at risk at some time X can be expressed as $n_k(X) \equiv N_k \Pr(T_k \geq X, C_k \geq X, \tau - E_k \geq X) = N_k(1 - F_k(X))(1 - G_k(X))H_k(\tau - X)$ is a function of the survival, censoring, as well as the entry times. N_k denotes the number initially at risk for group k. We note that under immediate accrual we observe the full survival time of the patients unless they are subjected to random censoring. With staggered accrual, this delay of accrual further induces a natural censoring that is common in most clinical trials. This dependence on both the censoring and survival later becomes an issue when we investigate the use of adaptive modifications to the design in the time to event setting.

A consistent estimator of the variance σ^2 for the above Score representation of the Cox regression can be written as

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} \frac{n_0(X_i)n_1(X_i)}{n_0(X_i) + n_1(X_i)} \left[1 - \frac{\Delta_1(X_i) + \Delta_0(X_i)}{n_0(X_i) + n_1(X_i)}\right] \frac{\Delta_1(X_i) + \Delta_0(X_i)}{n_0(X_i) + n_1(X_i)}$$

Under some local alternative β, the Cox regression model can be written as follows

$$\begin{aligned}
\mathcal{U}(\beta) &= \sum_{i=1}^{N} \Delta(X_i) \left[Z_i - \frac{n_1(X_i)\exp^{\beta}}{n_0(X_i) + n_1(X_i)\exp^{\beta}}\right] \\
&= \sum_{i=1}^{N} \left[\Delta_1(X_i) \left(\frac{n_0(X_i)}{n_0(X_i) + n_1(X_i)\exp^{\beta}}\right) - \Delta_0(X_i) \left(\frac{n_1(X_i)\exp^{\beta}}{n_0(X_i) + n_1(X_i)\exp^{\beta}}\right)\right] \\
&= \sum_{i=1}^{N} \frac{n_0(X_i)n_1(X_i)}{n_0(X_i) + n_1(X_i)\exp^{\beta}} \left[\frac{\Delta_1(X_i)}{n_1(X_i)} - \frac{\Delta_0(X_i)}{n_0(X_i)}\exp^{\beta}\right]
\end{aligned}$$

The statistical information can be represented as

$$\mathcal{I}(\beta) = \mathbb{E}\left[-\frac{\partial}{\partial\beta}\mathcal{U}(\beta)\right] = \sum_{i=1}^{N} \left\{\frac{\Delta(X_i)n_1(X_i)n_0(X_i)\exp^{-\beta}}{[(n_0(X_i)\exp^{-\beta} + n_1(X_i)]^2}\right\}$$

Our Score and Wald versions of the Cox PH model/logrank statistic are represented

below.

$$\begin{aligned} \text{Score Statistics:} \quad & \frac{\mathcal{U}(\beta)|_{\beta=0}}{\sqrt{\mathcal{I}(\beta)|_{\beta=0}}} \quad \xrightarrow{d} \quad \mathcal{N}(0,1) \\ \text{Wald Statistics :} \quad & (\widehat{\beta}-\beta)\sqrt{\mathcal{I}(\widehat{\beta})} \quad \xrightarrow{d} \quad \mathcal{N}(0,1) \end{aligned}$$

Asymptotically, the limiting distributions for both test statistics are standard normal. However, evaluating the variance of the Score/Wald statistics makes different assumptions about the hypotheses of interest. For the Score statistic, the variance of $\mathcal{U}$, $\mathcal{V}(\beta)$, is commonly evaluated under the null hypothesis by setting $\beta = 0$. This gives us the logrank test statistic. On the other hand, the variance of the Wald statistic is evaluated based on the (partial) MLE of the Cox PH model.

4.5.2 Sample Size Formula

We can denote $\theta(t) = -\beta(t)$ to be the log hazard ratio $-\log(\lambda_1(t)/\lambda_0(t))$ where $\lambda_1(t), \lambda_0(t)$ are the corresponding hazard function at time t for the treatment and placebo respectively. By this parametrization, under proportional hazards (we suppress the dependence on t under the PH setting), our hypotheses can then be interpreted such that positive values of θ denote superiority of the treatment over the placebo with $H_0 : \theta \leq 0$ vs $H_A : \theta > \theta_A$. At level α, and power β, our event size estimated based on a $r : 1$ randomization schedule can be obtained using $d = V(z_{1-\alpha} + z_\beta)^2/\theta_A^2$ where $V = (r + 1/r)^2$.

Under a $1 : 1$ randomization, $V = 4$. A reasonable method to determine the accrual size can be obtained via $N = 2d\left[\sum_{k=0,1}\left(1 - \frac{\exp(-\lambda_k(\tau-a))}{\lambda_k a} + \frac{\exp(-\lambda_k \tau)}{\lambda_k a}\right)\right]^{-1}$ [Schoenfeld, 1983]. This assumes uniform accrual of subjects over some time interval $(0, a)$, with the final analysis taking place at time $\tau \geq a$, and censoring of observations to occur only by continued survival at the time of analysis. Additionally, this formulation assumed exponential survival times for both placebo and treatment. Specifying these parameters requires making an educated guess using prior research. However, that guess may differ greatly from the conditions under

which the clinical trial is actually implemented.

Sample size revision strategy can be performed using the aggregate observed event rate during the course of the trial to revise the accrual of subjects when the hypothesized event rate may be lower than anticipated. Because under approximate proportional hazards and equal randomization, $N/V \approx d/4$ when using the logrank statistic. Such sample size revisions to increase accrual without changing the number of events do not affect the statistical power of the design because modification of the sample size does not affect the formula for computing the number of events. Alternatively, one may prespecify the threshold based on the lower quantile of the approximate number of events to be anticipated at the calendar time of analysis to facilitate sample size re-estimation. We describe more of this in Chapter 5. More complex "sample size" calculations (for e.g., piecewise survival, piecewise weibull, etc) can be performed using `RCTdesign` [Emerson, 2000].

4.5.3 Limitations of Cox Proportional Hazards Regression/Log Rank Statistic

The assumption of a common treatment effect across stages in a clinical trial may be reasonable under the (strong) null hypothesis (where we presume exact equality of survival distributions across all times), or under strong parametric, or semi-parametric assumptions (such as proportional hazards). In most situations, such departures may arise even when we are demonstrating superiority and do not preclude users from conventional test statistics such as the logrank test. Other comparisons of the experimental therapy with the standard of care may have different efficacy profiles across time. Thus, under such departures, we may no longer be concerned with the assumption of a constant treatment effect but may need to consider the design operating characteristics under a time varying treatment effect.

Amna Ibrahim, MD of the Food and Drug Administration and Center for Drug Evaluation and Research remarked during an Oncologic Drugs Advisory Committee meeting on September 13, 2005: *"Hazard ratios give only an incomplete picture. Hazard ratios may represent statistical significance, however, clinical relevance as the benefit provided to the patient is not captured. For example, hazard ratios will treat the improvement from three days to six*

days the same as improvement from three years to six years."

In the presence of non PHs, the hazard ratio may not be the best summary measure to make inferential decisions on the preferred treatment since it no longer provides consistent and relevant scientific interpretation. Because the risk sets are dependent on both the censoring and the underlying survival distribution as in 4.5.1, in presence of time varying treatment effects, we have an average hazard ratio interpretation [Xu and O'Quigley, 2000]. With sequential monitoring, the consequence of interim analyses thus imposes a differential weighting scheme that modifies the relative importance of the treatment effect across time, and/or unnecessarily induces trends unrelated to the scientific estimate of interest.

The average hazard ratio interpretation is also non transitive across trials [Gillen and Emerson, 2007]. This non transitivity has implications in the non inferiority settings. Since the estimates of the standard of care treatment effect are based on historical data from a placebo controlled RCT, those estimates may not be based on the same censoring distribution as the current active trial. As such, results comparing the placebo, standard of care, and experimental treatment may give rise to bizarre interpretations as discussed in Gillen and Emerson [2007]. Specific weighting schemes for the logrank statistics have been investigated to emphasize scientific importance rather than statistical efficiency, thereby preserving the transitivity of the test statistic in the group sequential setting [Gillen and Emerson, 2007]. However, the weighted estimates lack clear interpretable results.

4.6 Less Common Time To Event Analysis

We now introduce the less common time to event analyses methods that are often used in the setting of delayed ascertainment of outcomes.

4.6.1 Weighted Logrank Statistics/$G^{\rho,\gamma}$

Assume the notation from 4.5, let $w(t) = \frac{n_0(t)n_1(t)}{n_0(t)+n_1(t)}[\widehat{S}(t^-)]^\rho[1-\widehat{S}(t^-)]^\gamma$ where $\widehat{S}(t^-)$ is the pooled Kaplan Meier survival estimate. Fleming and Harrington [1991] introduced the above

flexible weight function in the log rank test statistic to accommodate comparison of a bigger class of survival curves. In this setup, at some time τ,

$$G^{\rho,\gamma} = \sqrt{\frac{N_0+N_1}{N_0N_1}} \sum_{i=1}^{N} w(X_i) \left[\frac{\Delta_1(X_i)}{n_1(X_i)} - \frac{\Delta_0(X_i)}{n_0(X_i)} \right]$$
$$W^{\rho,\gamma} = \sqrt{\frac{N_0N_1}{N_1+N_0}} G^{\rho,\gamma}$$

$W^{\rho,\gamma}$ has its equivalent representation as the weighted score statistics. Under the strong null hypothesis, $H_0 : S_0(t) = S_1(t),\ \forall t > 0$, a consistent estimator of the variance of the $G^{\rho,\gamma}$ statistic can be expressed as

$$\widehat{\sigma}^2 = \frac{N_0+N_1}{N_0N_1} \sum_{i=1}^{N} w^2(t) \left[\frac{1}{n_0(X_i)} + \frac{1}{n_1(X_i)} \right] \left[1 - \frac{\Delta_1(X_i)+\Delta_0(X_i)}{n_0(X_i)+n_1(X_i)} \right] \frac{\Delta_1(X_i)+\Delta_0(X_i)}{n_0(X_i)+n_1(X_i)}.$$

The number of patients at risk in each comparison group will depend on the analyses time under staggered accrual. Since the $G^{\rho,\gamma}$ statistics or the weighted logrank statistics can be re-expressed as a weighted sum of statistics, $U_{\tau_j}(\beta)$, based on accumulated data up to calendar time τ_j (the calendar time at the j^{th} interim analysis), we can represent $V(\tau_j)$ in the form of $\widehat{\sigma}_j^2$ and estimate the information growth at the j^{th} interim analysis, Π_j, via

$$\Pi_j = V(\tau_j)/V(\tau_J) = \left[\left(\frac{N_{0,j}N_{1,j}}{N_{0,j}+N_{1,j}} \right) \widehat{\sigma}_j^2 \right] \Big/ \left[\left(\frac{N_{0,J}N_{1,J}}{N_{0,J}+N_{1,J}} \right) \widehat{\sigma}_J^2 \right].$$

where $N_{k,j}$ denotes the number initially at risk for treatment group k at the j^{th} interim analysis.

Under the strong null hypothesis, our weighted test statistic, U_{τ_j}, is approximately normal with mean 0 and variance/Fisher's information, $V(\tau_j)$. U_{τ_j} has the asymptotic properties of the Brownian motion with uncorrelated increments in the covariance structure with $\text{Cov}(U_{\tau_{j+1}}, U_{\tau_j}) = V(\tau_j)$ for $j = 1, \cdots, J-1$ under the strong null hypothesis [Tsiatis, 1982, Gu and Lai, 1991, Bilias et al., 1997].

4.6.2 Nelson Aalen Statistic

Let E, T, and C be the random variables corresponding to the calendar time of entry into the study, the study time for the event of interest, and the study time for loss-to-follow-up with respective distribution functions H, F, and G. Let x be the study time of interest. At calendar time τ, a total of $N = N_0 + N_1$ independent subjects have entered the study, and are randomized to treatment groups 0 and 1. The data for the i^{th} subject can be represented as $(X_{ik}(\tau), \Delta_{ik}(\tau), k)$ where $X_{ik}(\tau) = \max(\min(T_{ik}, C_{ik}, \tau - E_{ik}), 0)$ is the observed time for individual i, $\Delta_{ik}(\tau)$ is the indicator variable for an observed failure time if $X_{ik}(\tau) \leq \min(C_{ik}, \tau - E_{ik})$ and 0 if $E_{ik} - \tau > 0$ and that loss of followup is only due to administrative censoring, and k is the randomized treatment assignment taking values 0 or 1.

At calendar time τ, the Nelson-Aalen estimator of the cumulative hazard function $\Lambda_k(x) = \int_0^x \lambda(u)du$ at some time x computed for the k^{th} treatment group is thus

$$\widehat{\Lambda}_k(x;\tau) = \sum_{\{i:X_{ik}(\tau)\leq x\}} \frac{\Delta_{ik}(\tau)}{n_{ik}(\tau)}$$

where $n_{ik}(\tau) = \sum_{i=1}^{N} 1\{X_{jk}(\tau) \geq X_{ik}(\tau)\}$, $\Delta_{i0}(\tau) = \Delta_i(\tau)(1 - Z_i)$, and $\Delta_{i1}(\tau) = \Delta_i(\tau)(Z_i)$.

The estimate of the variance $\widehat{\sigma}_k^2$ for group w is thus

$$\widehat{\sigma}_k^2(x;\tau) = \sum_{\{i:X_{ik}(\tau)\leq x\}} \frac{\Delta_{ik}(\tau)}{n_{ik}^2(\tau)}$$

Formally, we can compare the difference in the survival curves by testing the following hypothesis, $\mathrm{H}_0^{NA} : \mathrm{S}_1(x) = \mathrm{S}_0(x)$. Since $\mathrm{S}(x) = \exp(-\Lambda(x))$, the above hypothesis can be represented based on the cumulative hazard function. We may thus write the hypothesis as $\mathrm{H}_0^{NA} : \Lambda_1(x) = \Lambda_0(x)$. At some calendar time, τ, a total of N_0 and N_1 subjects from group

0 and 1 respectively are initially at risk, our asymptotic distribution

$$\sqrt{N}\left[(\widehat{\Lambda}_1(x;\tau)-\widehat{\Lambda}_0(x;\tau))-(\Lambda_1(x)-\Lambda_0(x))\right]\to\mathcal{N}(0,\mathcal{W})$$

with the variance $\mathcal{W}$ of the Gaussian process defined as $\rho^{-1}\sigma_0^2(\tau)+(1-\rho)^{-1}\sigma_1^2$ where $\rho=\lim_{\min\{N_0,N_1\}\to\infty}N_0/(N_1+N_0)$ in Lin et al. [1996].

Denote $i=1,\cdots,N_0$, and $l=1,\cdots,N_1$ to be the indices for the total number of subjects in group 0 and 1 respectively. Then our estimator can be expressed as follows:

$$\widehat{\Lambda}(x;\tau)=\widehat{\Lambda}_1(x;\tau)-\widehat{\Lambda}_0(x;\tau)=\sum_{\{l:X_{l1}(\tau)\le x\}}\frac{\Delta_{l1}(\tau)}{n_{l1}(\tau)}-\sum_{\{i:X_{i0}(\tau)\le x\}}\frac{\Delta_{i0}(\tau)}{n_{i0}(\tau)}$$

Our variance estimator of $\mathcal{W}$ is

$$\begin{aligned}\widehat{\mathcal{W}}&=\rho^{-1}\widehat{\sigma}_0^2(\tau)+(1-\rho)^{-1}\widehat{\sigma}_1^2=N\left(\widehat{\mathrm{var}}\{\widehat{\Lambda}_1(\tau)\}+\widehat{\mathrm{var}}\{\widehat{\Lambda}_0(\tau)\}\right)\\&=N\left(\sum_{\{i:X_{i1}\le x\}}\frac{\Delta_{i1}(\tau)}{n_{i1}^2(\tau)}+\sum_{\{i:X_{i0}\le x\}}\frac{\Delta_{i0}(\tau)}{n_{i0}^2(\tau)}\right)\\&=N\widehat{V}(\tau)=\overline{\mathcal{I}}(\tau)^{-1}\end{aligned}$$

The inverse of the statistical information, $\overline{\mathcal{I}}(\tau)^{-1}$, can be estimated using $N\widehat{V}(\tau)$.

At calendar time τ, our test statistic is then represented by

$$Z(\tau)=\frac{\widehat{\Lambda}(\tau)-\Lambda_{\mathbb{H}_0}(\tau)}{\sqrt{\widehat{V}(\tau)}}$$

where $Z(\tau)$ is the Wald test statistic with $\overline{\mathcal{I}}(\tau)\equiv\mathcal{I}_1(\tau)$. Note that $\overline{\mathcal{I}}_N=N\overline{\mathcal{I}}(\tau)$ and can be estimated by $1/\widehat{V}(\tau)$. Under the null, the Score representation of the test statistic is

$$\mathcal{U}_N(\tau)=\sqrt{N\mathcal{I}_1(\tau)}Z(\tau)=\sqrt{\mathcal{I}_N(\tau)}Z(\tau)=Z(\tau)/\sqrt{\widehat{V}(\tau)}$$

We later consider the above Nelson-Aalen test statistic to be used jointly with the logrank

test statistic to obtain the composite statistics that will be investigated in Chapter 7.

4.6.3 Weighted Kaplan Meier Statistic

The restricted mean or the Weighted Kaplan Meier statistic [Pepe and Fleming, 1989] can be used to compare the area under the survival curves. In particular, we are interested in comparing this area under the two survival curves at some calendar time τ while restricting the comparison to a study time of x. This is formulated by

$$\widehat{\mathrm{WKM}}_{1 \text{ vs } 2}(x) = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \int_0^x \widehat{w}(t) \left[\widehat{S}_1(t) - \widehat{S}_2(t)\right] \partial t$$

where $\widehat{w}(t)$ denotes the weight function that can be chosen to place importance on parts of the survival curves, $\widehat{S}_k(t), \widehat{C}_k(t^-)$ are the Kaplan Meier estimate of survival distribution up to time t and the probability of not being censored before time t^- in group k, and N_k is the total number of patients initially at risk in group k for $k = \{1, 2\}$. Choosing $\widehat{w}(t) = 1$ reduces to the restricted mean statistic.

Typically, let $x = \sup\{t : \widehat{C}_1(t) \wedge \widehat{C}_2(t) > 0\}$ where $\widehat{C}(t)_k$ is the Kaplan Meier estimator of the censoring survival function for each group. In practice, we replace x by $x - t^*$ where t^* is some number of time units chosen to avoid integrating over a support of 0 for stability reasons.

The unpooled and pooled variance estimator are defined as below

$$\widehat{\sigma}^2_{\text{Unpooled}} = -\sum_{l=1}^{2} \widehat{p}_{3-w} \frac{N_l}{N_l - 1} \int_0^x \frac{[\int_0^x w(u)\widehat{S}_w(u)du]^2}{\widehat{S}_w(t)\widehat{C}_w(t)\widehat{S}_w^-(t)} d\widehat{S}(t)$$

$$\widehat{\sigma}^2_{\text{Pooled}} = -\int_0^x \frac{[\int_0^x w(u)\widehat{S}(u)du]^2}{\widehat{S}(t)\widehat{S}^-(t)} \left(\frac{\widehat{p}_1\widehat{C}_1(t) + \widehat{p}_2\widehat{C}_2(t)}{\widehat{C}_1(t)\widehat{C}_2(t)}\right) d\widehat{S}(t)$$

where $\widehat{S}(t)$ is the Kaplan-Meier estimator calculated from the pooled sample, $\widehat{p}_1 = N_1/N$, $\widehat{p}_2 = N_2/N$, and N is the total number at risk at the beginning of the trial. Under $\mathbb{H}_0$, both estimators are consistent [Pepe and Fleming, 1991]. One may choose the weight function,

$w(t) = 1(t \geq 0)$, or the optimal choice described by Pepe and Fleming [1989] as

$$w_c(t) = \left(\frac{p_1}{\widehat{C}_2(t)} + \frac{p_2}{\widehat{C}_1(t)} \right)^{-1}$$

4.7 Summary

This chapter serves mainly as a background to the common statistical methods that are employed as the analysis tool of choice in the time to event setting during the design and conduct of clinical trials. There are many recent statistical methods such as targeted learning [Van Der Laan, 2011] that are gaining popularity. Other approaches also may allow covariate adjustments in analyzing the primary endpoint in the clinical trial setting. These are not discussed in this dissertation since we choose to focus on using these common approaches to portray some of the statistical concerns with the use of GSDs and adaptive designs.

We now investigate the use of adaptive designs under the PH setting assuming the use of the efficient logrank test statistics in the next chapter.

Chapter 5

Adaptive Monitoring of HIV Prevention Trials in the Presence of Extreme Treatment Effect

In recent years, there has been considerable interest in the possibility of incorporating adaptive features in clinical trials of new prevention strategies. Recent HIV prevention trials such as HPTN052 and Partners Pre-Exposure Prophylaxis dealt with operational issues related to the observation of extremely low event rates at blinded analyses. In such a setting, trial investigators may need to consider increasing accrual rates, prolonging calendar time of follow-up, and/or accepting a lower number of events. The best strategy from among these options may depend on whether such an observation arises primarily from a low "background" event rate, or from an extremely beneficial treatment effect, or from a combination of both. In this research, we compare our ability to preserve study precision solely through the use of blinded adaptation within a pre-specified group sequential design versus the use of an unblinded adaptation that might better distinguish between low background event rates and extreme treatment effects. In particular, we consider the constraints that calendar time might impose on the scientific interpretation and statistical credibility of the chosen strategies.

5.1 Introduction

HIV prevention trials typically involve randomizing a large number of participants and following them for a period of time until sufficient information has been obtained about the incidence of sero-conversion. Because statistical information in a time-to-event setting is

presumed to correspond to the number of events, determination of the sample size requires additional information about the event rate and patterns of accrual into the study. An event rate markedly lower than planned can lead to an undesirably long calendar time before the planned number of events can be observed. Fortunately, recent experiences of some randomized controlled trials (RCTs) have shown that extreme treatment effects can mitigate problems of unanticipated low event rates, but knowledge that such has occurred requires access to unblinded interim trial results that may bias the operational aspects of RCT. It is useful to consider the experience of two such recent RCTs:

HPTN052 was declared as one of the scientific breakthroughs of the year in 2011 [Cohen, 2011]. The primary objective in HPTN052 was to determine whether the early use of combination anti-retroviral therapy (ART) in infected patients among serodiscordant couples is effective in the prevention of HIV-1 transmission to uninfected partners [Cohen et al., 2011]. The trial was designed to provide at least 87% power to detect at least 39% reduction in the primary endpoint of HIV incidence. Based on various logistical constraints (18 months accrual, projected completion of follow-up at 6.5 years), an estimated accrual size of 1750 participants was computed assuming average 5-year placebo (13.2%) and treatment (8.3%) event rates, and an anticipated 188 events. Six years after HPTN052 started, blinded analysis during a planned formal interim review showed 39 HIV infections among the 1,763 enrolled couples (877 on delayed vs 886 on early) with 28 of them being linked transmissions. Unblinded analysis showed that 27 of the linked transmissions arose on the delayed ART arm while only one came from the early ART arm yielding a hazard ratio of 0.04 (95% CI: 0.01 - 0.27; p-value < 0.0001). On the basis of this analysis, the DSMB recommended stopping further follow-up in the RCT due to demonstrated efficacy of the experimental treatment.

Partners PrEP(Pre-Exposure Prophylaxis) is a Phase III, randomized control, double-blind, three arm trial of daily oral tenofovir (TDF), or emtricitabine/tenofovir (FTC/TDF) for the prevention of HIV transmission as their primary endpoint among HIV serodiscordant partners [Baeten et al., 2012]. Based on a placebo event rate of 2.75 infections per 100 PY, 4,747 HIV serodiscordant couples, randomized with equal probability to the three arms and

followed for 36 months, would be expected to provide the necessary number of events. Using a GSD with up to a maximum of four planned interim analyses, the trial was stopped early at the third interim analysis due to crossing the efficacy boundary. The observed placebo event rate was much smaller than what was used in planning, with the observed treatment effects to be more extreme than had been anticipated.

5.1.1 Scientific Issues

These two trials illustrate two main factors that can complicate the design and conduct of vaccine/preventive strategies and influence the calendar time of the trial greatly: lower than anticipated control event rates (30% lower than planned in Partners PrEP, and more than 60% lower than planned in HPTN052) and stronger than anticipated observed treatment effects (strikingly so in HPTN052). Other time to event trials that face problems of low event rates during monitoring have also been observed previously in other disease settings such as the Look AHEAD [Group, 2003] study and the National Lung Screening Trial [Aberle et al., 2011].

Poorly characterized event rates at design stage in a time to event trial may not balance the ethics of randomizing exposure of prevention strategy with unknown efficacy profile to participants. An under-estimated sample size can limit the total number of events accumulated by the constraints of the calendar time, thereby preventing one from distinguishing between the scientific hypotheses of interest. Likewise, this can result in unnecessary extension of a trial that may not balance the overall goal of science, ethics, and efficiency. In this research, we focus on the setting when the true treatment effect is extremely more effective than hypothesized. This presents the additional challenge as observed in the trials above, whereby during a blinded interim analysis, trial investigators may be posed with the dilemma of choosing between unnecessarily prolonging the study to obtain the events, to increase accrual, to accept a lower number of events, or some combination of the above.

Following the promising results from HPTN052 and Partners PrEP, the introduction of non-vaccine preventive methods (NVPM) to the public will likely lead to a decrease in the

incidence of the HIV infections in the coming years [Janes et al., 2013]. If the overwhelmingly positive results of these trials translate directly to effectiveness, future prevention trials may require a larger accrual size in order to detect meaningful effects within some three to five years of follow-up. Alternatively, the usual five years of follow-up data that current prevention trials have been using may be insufficient to address the effectiveness of preventive strategy as the rate of HIV infections becomes even lower. Hence, it is important to anticipate potential misspecification in incidence rates of HIV during trial planning, and explore the implications on statistical power to detecting the hypothesized treatment effects, the impact on calendar time, and the total number of patients required.

Estimating the incidence of HIV seroconversion over time for future trials is more complex in a time to event setting. Since the estimate of the incidence is an average over various at risk populations that can differ greatly in terms of risk, attitudes, and behavior, this may present challenges to appropriately characterizing the incidence over time. The ease of modern preventive methods as compared to traditional preventive strategies (condom use) can modify individual's behavior and attitudes towards sexual practices over time. In addition, new efficacious interventions have yet to demonstrate effectiveness in a public setting, thus many unanswered scientific questions on whether poor adherence to such strategies will result in resistant strains of HIV or that the increase risks of exposure with a waning prevention strategy have to be determined in subsequent trials to assess the impact on the participants. Potential changes in sexual behaviors, poor adherence in modern preventive strategies, in combination with emerging resistant HIV strains may introduce some form of waning treatment effect over time. As such, extension of the HIV prevention trial may change the scientific objective from effectiveness to durability of the prevention strategy.

5.1.2 Statistical Issues

Group sequential designs are the current gold standard used to balance scientific, ethical and efficiency concerns in clinical trials. Typically, a monitoring rule is pre-specified at design stage to determine the maximum statistical information $\mathcal{I}_J$ required where J is the final

analysis. At periodic intervals during the course of the trial, a test statistic is computed and compared with the pre-specified monitoring rule to determine whether the trial should stop early with a maximum of J such analyses being allowed.

The growing literature of adaptive strategies and FDA's recent guidance on adaptive designs [FDA, 2010] have led many researchers to consider whether adaptive RCT designs could protect against greatly prolonged (at best), or largely non-informative (at worst) studies. Following guidelines of FDA [2010], it is thus of use to compare adaptive designs with the class of designs that are "well-understood" (fixed sample design, blinded sample size revision, or group sequential designs with prespecified futility and efficacy boundaries). One of the most common form of adaptation includes extending the trial beyond some previously planned maximum stopping time in the immediate settings, such as cure of infections within 10 days of treatment initiation [Bauer and Köhne, 1994, Proschan and Hunsberger, 1995, Cui et al., 1999, Schäfer and Müller, 2001, Chen et al., 2004].

Levin et al. [2013] evaluated this class of pre-specified adaptive designs with GSDs using the average sample size as their optimality criterion under the immediate setting. They found that the best GSD have only very slightly higher ASN over the best prespecified adaptive design. In addition, when the pre-specification was relaxed, GSD is almost fully efficient compared to the best analogous adaptive design with ad-hoc unplanned adaptations. Other authors [Tsiatis and Mehta, 2003, Jennison and Turnbull, 2003, 2006a] also found little efficiency gains in adaptive designs over the use of standard GSDs. However, such explorations have typically been limited to the setting when the sample size is a surrogate measure of statistical information.

In the time to event setting, the overall cost of the trial is generally related to the total number of subjects as well as the calendar time of obtaining all relevant events. Since the calendar time is related to the financial and logistical burdens of the sponsor or research institute, there is thus an appeal to enable possible modification of the accrual of subjects depending on trends in treatment effect to decrease the overall cost of the trial. Emerson et al. [2011a] investigated a limited spectrum of pre-specified adaptive designs where the best

GSD averaged a higher sample size over pre-specified adaptive designs without compromising the trial duration in the censored setting. Their exploration did not consider sub-optimal adaptive rules commonly favored in the adaptive literature [Schäfer and Müller, 2001, Shen and Cai, 2003] which re-weights the test statistics to preserve the overall Type 1 error. Under low "background" rates, there is still a degree of uncertainty on whether the use of pre-specified adaptive design can buy the trialists added efficiency and protection when the timing of interim analysis do not align with the expected schedule.

A low "background" rate during a blinded interim analysis in the trial can often be due primarily to an extremely effective treatment, or primarily due to low event rate, or a combination of both. As determination of the sample size typically requires additional information about the event rate and patterns of patient accrual, an event rate markedly lower than planned can lead to an undesirably long calendar time before the planned number of events can be observed. In the extreme situations, such as HPTN052, where the treatment effect is more extreme than hypothesized, misspecification of the event rates at planning phase can also lead to insufficient sample size as a consequence of events only accruing on the placebo arm. As such, trials may either have to consider the possibility of both extending the calendar time as well as increasing the accrual size to account for these misspecifications, or instead terminate a trial for futility.

In situations when blinded increase in accrual is allowed, the timing during which the additional accrual is performed can be crucial. An adaptation too early during the trial may benefit the sponsor in terms of continued accrual, but the lack of precision of the event rate may unnecessarily expose relatively more patients to treatments with unproven benefit-to-risk profile. Likewise, an adaptation made too late in the trial may provide better precision of the overall "background" rate, but it has the downside of having to restart accrual which is logistically difficult. Additionally, ethicists may argue against the use of such strategies in extreme settings, such as HPTN052, when there is potentially strong evidence already accumulated regarding the efficacy of the treatment.

There are strong ethical reasons to favor the use of unblinded interim data to better guide

decision making to efficiently increase accrual only when necessary. An interim treatment effect that is close to statistical significance with the placebo event rate as expected, may not require as much an accrual in sample size as compared to a situation when the event rate is clearly underestimated. In settings with poorly characterized incidence rates regarding a serious, life-threatening rare disease, there may be a stronger rationale to using unblinded interim results to better characterize the event rate during the conduct of the trial. The accuracy of the decision making to increase accrual may, similar to the blinded setting, depend on whether this adaptation should be conducted during an interim analysis when patient accrual is still open as opposed to after the accrual of subjects have stopped.

5.1.3 Organization

Our goal is to investigate the use of unblinded interim analyses in the setting of potentially low event rates/extreme efficacy to determine whether we can address the "right" scientific question with statistical credibility as compared to the use of "well-understood" designs. We restrict attention to comparing between group sequential designs vs pre-specified adaptive designs using common monitoring rules that are used in practice. Within the scope of GSDs, we introduce a modification to incorporate the maximum calendar time as an additional stopping criteria together with the design of choice. We further incorporate a blinded revision of sample size with the hopes of increasing the overall event rate. We then examine the operating characteristics of such a GSD, which is equipped with fully blinded adaptations and pre-specified calendar time of stopping, under possible low event rate settings.

In the immediate setting, adaptive designs have not been shown to be markedly more efficient than GSD in the best case. This is consistent with theoretical arguments that these statistics violate the minimal sufficiency principle, thus leading to an inefficient weighting scheme of the data collected prior to and after a design modification based on unplanned, unblinded interim results [Tsiatis and Mehta, 2003, Jennison and Turnbull, 2003, 2006a, Emerson, 2006]. However, many statistical papers considered such inefficient adaptations that might include a decision at the penultimate analysis to more than double the maximal

sample size [Cui et al., 1999, Chen et al., 2004, Gao et al., 2008, Mehta and Pocock, 2011]. It is thus unclear whether an optimal flexible adaptive strategy can be selected to maximize the operating characteristics and beat GSDs.

To distinguish whether such a flexible adaptive strategy can provide any advantage over fully blinded adaptations based on GSDs, we must first separate out design issues that are related to choosing a good/bad sampling scheme (i.e., the timing of the unplanned adaptation) vs the use of weighted and inefficient statistics. We model this flexible adaptive strategy by considering the worst-case scenario when an adaptation was prespecified but may fail to provide the sponsor an edge to make desired adaptations. This is particularly important in the time to event setting when planned event rates/treatment effect may deviate from the expected calendar time of analyses. As such, we consider finding the "best" sampling scheme when making adaptations based on sufficient, but not necessary minimal, statistics.

This approach differs from the methods described in Emerson et al. [2011a] and Levin et al. [2013] where the best sampling scheme is obtained based on finding the design that uses minimal sufficient statistics. By selecting this "best" inefficient rule and pre-specifying it at design stage, this "flexible adaptive rule" would now be based on a minimal sufficient statistics among the class of sampling schemes that employs the use of weighted statistics. Henceforth, we are evaluating the class of fully flexible adaptive designs in the best possible light. This allows us to contrast how much gain/loss of overall power we can obtain had the adaptation be specified in advance. We note that these flexibilities in the fully adaptive design also mean that there are limited statistical inference techniques that can be used to account for the sequential stopping rules and early stopping.

5.2 Conventional Statistical Designs at Planning Stage

We motivate the notation under the FSD setting based on the use of the logrank analysis or Cox proportional hazards regression. Let $\theta(t)$ denote our hazard ratio, $\lambda_1(t)/\lambda_0(t)$, where $\lambda_1(t), \lambda_0(t)$ are the corresponding hazards at time t for the treatment (1) and control arm (0) respectively. Under PH alternatives, $\lambda_1 = \theta\lambda_0$, with $\theta(t) = \theta, \ \forall t$. When testing a new

treatment with respect to an existing treatment or placebo, our hypotheses can be expressed as $H_0 : \log(\theta) \geq 0$ vs $H_A : \log(\theta) \leq \log(\theta_A)$, where $\theta < 1$ represents the efficacy of treatment over placebo. Using a one-sided level α, and power β, our event size can be estimated using $d = 4(z_{1-\alpha} + z_\beta)^2/[(\log(\theta))^2]$ where z_p denotes the p^{th} quantile of a standard normal distribution for $p \in (0, 1)$. A reasonable method to determine the accrual size assuming a 1:1 randomization scheme can be computed via $N = 2d\left[\sum_{k=0,1}\left(1 - \frac{\exp(-\lambda_k(\tau-a))}{\lambda_k a} + \frac{\exp(-\lambda_i \tau)}{\lambda_k a}\right)\right]^{-1}$ [Schoenfeld, 1983]. This assumes uniform accrual of subjects over some time interval $(0, a)$, with the final analysis taking place at time $\tau \geq a$, censoring of observations to occur only by continued survival at the time of analyses, and assuming exponential survival times for both placebo and treatment. Specifying these parameters require making educated guesses based on prior research that may differ vastly from current conditions.

Without loss of generality, the accrual size at planning stage can be obtained based on either λ_0 and θ_A, or λ_0 and λ_1. The latter takes into account treatment estimates obtained from earlier trials. The former approach makes the PH assumption and exponential survival rates. Since λ_1 can be obtained via $\theta\lambda_0$ under PH, this approach only requires guessing the estimate of λ_0 from prior studies and hypothesizing a clinically meaningful θ. Using our design alternative of $\theta = \theta_A$ to estimate λ_1 is similar to blinded repowering strategies when the overall event rates are lower than anticipated. When the average event rate, $\widehat{\lambda}$, is provided, one may revise the trials' accrual size by directly invoking the accrual size formula again. If only $\widehat{\lambda}_0$ is provided, one may plug in the design alternative θ_A to obtain $\widehat{\theta}_1$ without unblinding the treatment effect to re-estimate the new accrual size.

In the immediate setting, the sample size is a direct surrogate to the statistical information. In a time to event setting, under PH and presuming the use of the log rank analysis, the maximum statistical information is directly proportional to the number of events. The number of subjects to accrue depends then on the accrual rate, accrual time, and the distribution of the survival time. Blinded revision of the background rates to increase the accrual size in the time to event setting, when baseline rates are "incorrect", do not materially change the overall power of the trial at design stage.

5.2.1 Notation for Group Sequential Design

A naïve FSD in a time-to-event setting may not appropriately address and balance scientific, ethical, or efficiency concerns. GSD is the current gold standard for clinical trial design. The ultimate goal of sequential sampling in group sequential design is to only proceed to the maximal sample/event size when the treatment benefit/risk is uncertain at interim analyses, or when there is potential that additional results of the trial can change the current public health or clinical practice.

Consider a GSD with continuation sets $\mathcal{C}_j \equiv \{(a_j, b_j] \cup [c_j, d_j)\}$ such that $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$ for $j = 1, \cdots J$ analyses [Kittelson and Emerson, 1999]. At each analysis, we compute the normalized score statistic, $Z_j = U_j/\sqrt{V(t_j)}$, and define our proportion of information at the analysis, j, to be of the form, $\Pi_j = V(t_j)/V(t_J)$, where U_j is the (cumulative) score statistic at j^{th} analysis, and $V(t_j) = \text{Var}[U_j]$ is the variance of the score statistics or Fisher's information. Π_j refers to the fraction of the statistical information available from all patients at the time of the j^{th} interim analysis relative to the maximum statistical information defined at design stage.

Sequential analysis sampling schemes usually consist of stopping sets which are linked across interim analysis via smooth parametric functions of the proportion of statistical information Π_j. Wang and Tsiatis [1987] and Kittelson and Emerson [1999] are some examples of such classes of stopping boundaries. The unified family in Kittelson and Emerson [1999], as described in section 2.2.1, includes many of these commonly used sequential sampling schemes which we shall use for construction of the group sequential boundaries for a test of $H_0 : \log(\theta) \geq 0$ against $\mathbb{H}_A : \log(\theta) \leq \log(\theta_A)$. Different parametrizations of the boundary shape function will provide different levels of early conservatism at each interim analysis.

In order to stop at a particular analysis, scientific, statistical, and ethical aspects have to be taken into consideration. An appropriate stopping rule would provide operating characteristics to provide statistical credibility of the study when the trial is stopped. The choice of a monitoring rule serves as a reference for the DSMBs to make decisions on stopping the

trial early for safety, efficacy, or futility reasons. Often, decisions undertaken by DSMBs in sequential monitoring are a complex balance of whether sufficient evidence has been established for them to act on the monitoring rules, i.e., to make decisions on early stopping, or extend follow-up to better understand the safety profile of the therapy. In other situations, decisions may be straightforward if the safety profile of the participants are unfavorable.

Several monitoring boundaries can be considered to achieve the desired Type 1 error and power as well as balance ethical and scientific considerations. The O'Brien Fleming boundary, the most commonly implemented stopping boundary used in clinical trials, is well-known to be relatively conservative at early stages of the trial and tends to only stop trials when extreme treatment effects are highly statistically significant [O'Brien and Fleming, 1979]. The Pocock boundary assumes a constant critical boundary on the Z statistics scale across interim analyses and tends to be more efficient in terms of ASN under the same hypotheses tested [Pocock, 1977]. Other choices can be a hybrid of O'Brien Fleming efficacy boundary and a moderate futility boundary as described in Kittelson and Emerson [1999].

In many time to event trials, conduct of interim analyses are often defined on the calendar time basis. This may be conducted on an annual or biannual basis. With sequential monitoring, increasing the number of analyses while keeping to constraints of the alternative and maintaining the same power can result in a increase in required maximal statistical information. With certain monitoring boundaries, this can inflate the maximum statistical information slightly while keeping other operating characteristics competitive.

In planning a GSD in the time to event setting for this research, we assume that the interim analyses are conducted biannually. Because the use of interim analyses can increase both financial and logistical cost of the trial, and it is often the case that sponsors commit financial and logistical resources to the conduct of the study on the calendar time scale. Thus, interim analyses that are planned on the calendar time relates directly to the sponsors' logistical and financial concerns. A monitoring boundary can be appropriately adjusted based on the accumulated statistical information using "well-understood" procedures [Lan and DeMets, 1983, Burington and Emerson, 2003], so long as the decision to eliminate any

future interim analyses is independent of the current estimated treatment effect.

5.2.2 Specification of a Group Sequential Design

Following Emerson et al. [2007], we motivate the planning and design of a clinical trial using a FSD before incorporating a monitoring rule. We consider a fixed sample design (FSD) with 92.16% power to detect at least a 36.6% reduction in risk of infection using a one-sided $\alpha = 0.025$, thus requiring 220 events. As previously described, additional constraints such as accrual rate, accrual time, amount of follow-up after the subjects are accrued, total study duration, and the distribution of the survival curves have to be factored in to determine the total number of subjects. With these constraints, one has to further choose the appropriate study duration to enable answering the relevant scientific question as closely as possible. Planning the study under the null hypothesis or alternative hypothesis can result in different sample sizes.

With censoring, and different choices of λ_0 and λ_1, the maximum calendar time of analysis will result in a different estimated sample size to maintain the statistical power for the alternative of interest, θ_A. The planning of the design is preferred under the alternative as such a maximal sample size will ensure stopping at the desired (average) calendar time when the trial parameters are approximately correct [DeMets and Lan, 1995].

We consider a modification of the HPTN052 study as an example to illustrate how the different assumptions can lead to various sample sizes. First, we impose the logistical constraint of uniform accrual of 18 months and restrict the total duration of the trial to be 78 months. The minimum amount of follow-up can range from as short as 60 months to a maximum of 78 months, thus giving an average follow-up of 69 months. With this consideration, Table 5.1 presents the plausible average accrual size when we plan the study under either the null or alternative, and using different combinations of accrual time and total study time.

For example, in Table 5.1, assume that the trial duration is 78 months with 18 months of accrual. If our event rates were $\lambda_0 = 0.002395, \lambda_1 = \lambda_0\theta = 0.001519$, and we plan our accrual size under the null hypothesis, on average, 1450 subjects are necessary to provide on

average 220 events by 78 months. On the other hand, if our alternative is true, then with 1750 subjects, we tend to stop on average by 78 months while possibly stopping earlier on average (at approximately 65 months) if the null hypothesis is correct. In that table, for other plausible values of event rates, we get a completely different total number of subjects to be accrued into the study. One can thus expect the total number of subjects to almost double had the event rates been halved or even lower.

Table 5.1: Fixed sample design powered based on average exponential null rate or alternative rate of (1.555e-3) to ensure stopping at 78 months with a sample size of 1750 assuming 18 months accrual (in bold) obtained via numerical integration. The remaining sample size is computed based on the average event rate under null or alternative.

	Calendar Time: 78 months				Calendar Time: 156 months			
	H_0	H_A	H_0	H_A	H_0	H_A	H_0	H_A
Design Parameters (x10^{-3})	Accrual of 18 months		Accrual of 36 months		Accrual of 18 months		Accrual of 36 months	
$\lambda_0 = 1.948, \lambda_1 = 1.235$	**1750**	2120	2000	2420	884	1060	934	1120
$\lambda_0 = 2.392, \lambda_1 = 1.517$	1450	**1750**	1650	2000	742	887	783	937

For the remainder of this chapter, we considered a sequential design with 10 equally spaced analyses based on a one-sided symmetric O'Brien Fleming stopping boundary, a Type 1 error of $\alpha = 0.025$, and 90% power to detect the alternative $\theta_A \leq 0.6343$ (*GSDOBF*). As such, the maximal event size of 220 is required if we do not stop early for overwhelming efficacy or futility, and that the interim estimates permit continuation to the final analysis.

We assumed our baseline placebo event rate is $\lambda_0 = 0.002395$, with accrual parameters and study time as described in the FSD example. By planning under the alternative, when the new treatment presents no benefit in efficacy, i.e., the null hypothesis of no treatment effect, the average calendar time of stopping is 65 months. On the other hand, under the alternative, the average calendar time of stopping will be 78 months. We consider alternative monitoring rules using a hybrid design (*GSDHYB*) that comprises of an O'Brien

Fleming efficacy boundary, and unified family boundary shape parameter $P = 0.8$ for the non-efficacy/futility boundary. We constrain the *GSDHYB* with the same design assumptions. The hybrid design has a less aggressive futility boundary with 89.2% power to detect the same alternative θ_A.

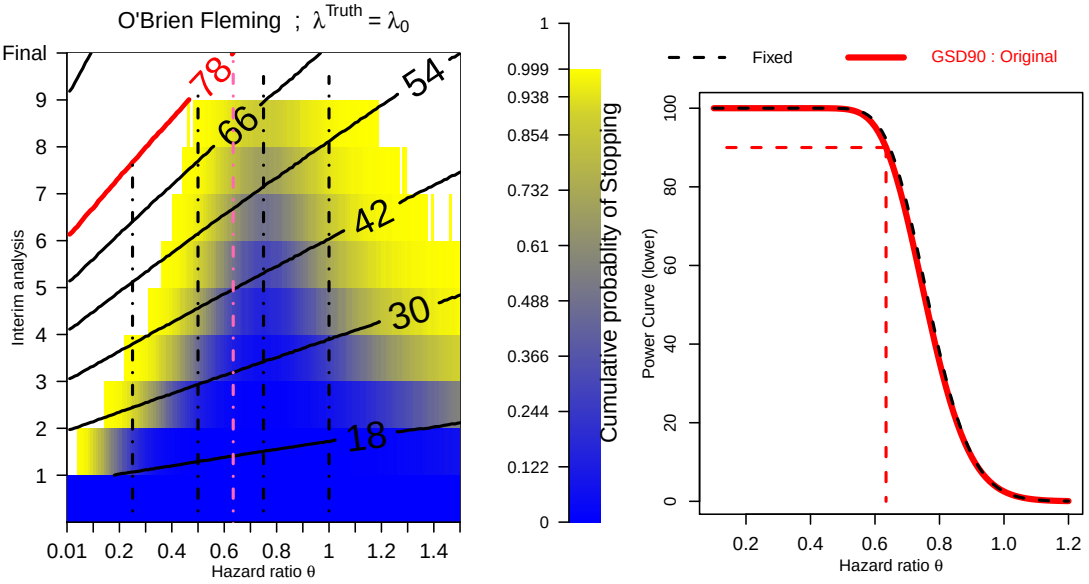

Figure 5.1: Heatmap of the cumulative probability of stopping at the interim analyses overlaid with the contours of the average calendar time of stopping for various hazard ratios θ is shown on the left. The average stopping time under the alternative is 78 months and shown as the red contour line in the heatmap. The pink vertical dash-dotted line represents the design alternative. The power curve under the planned event rate is presented on the right.

To investigate the impact of how incorrect baseline rates can affect the operating characteristics of the design, we also consider additional baseline rates $\lambda \in \lambda_0\{1/8, 1/4, 1/2, 3/4\}$ with λ_0 as our preplanned design rate. We evaluate the operating characteristics (power, average sample size, average event size) for the set of $\theta \in (1.2, 1, 0.75, \theta_A, 0.5, 0.25, 0.1, 0.04)$ in combination with the various event rates of consideration. At interim/final analyses, the log rank statistic was fit to compute the Z statistic.

Using the monitoring rules, we can evaluate the operating characteristics of the design

under misspecification of the event rates and compute the average calendar time at each interim analysis. When the baseline rates are as expected, the cumulative stopping probability for the various θs' (ranging from 0.01 to 1.5) at each interim analysis can be computed. Similarly, we can also compute the expected calendar time at each interim analysis for these θs'. These results are shown in Figure 5.1. Using the OBF rule, there is high probability(> 0.999) that the trial can stop prior to 78 months under extreme treatment effects ($\theta < 0.5$) in conclusion of efficacy of the treatment. At values of $\theta > \theta_A$, the trial would proceed to stop on average at a later calendar time.

5.2.3 Consequence of Low Event Rate

In general, the expected calendar time of attaining the maximal statistical information under the design alternative is greatly prolonged. Consider the *GSDOBF* design in Figure 5.2 (left column when the true event rates are halved or quartered of what was planned). Under the design alternative and presuming the use of an OBF boundary, our calendar time is on average doubled. The contour lines are shifted downwards with the red line indicating the expected maximum trial duration at planning stage. By planning for multiple interim analyses, there is a high probability of allowing the trial to stop early when θ is (sufficiently) extreme. At lower event rates, as the cumulative probability decreases at extreme treatment effects, we are somewhat protected with the ability to stop early by planning these multiple interim looks at design stage. If we assume that the calendar time of stopping is at 78 months, then there is high probability of stopping a trial at some earlier interim than planned at various values of θ. At values of θ that are moderately effective, i.e., $\theta \in (0.5, \theta_A)$, there is a lower cumulative probability of stopping by 78 months which directly impact the overall power of the trial. Similar effect is also seen with the hybrid monitoring boundary.

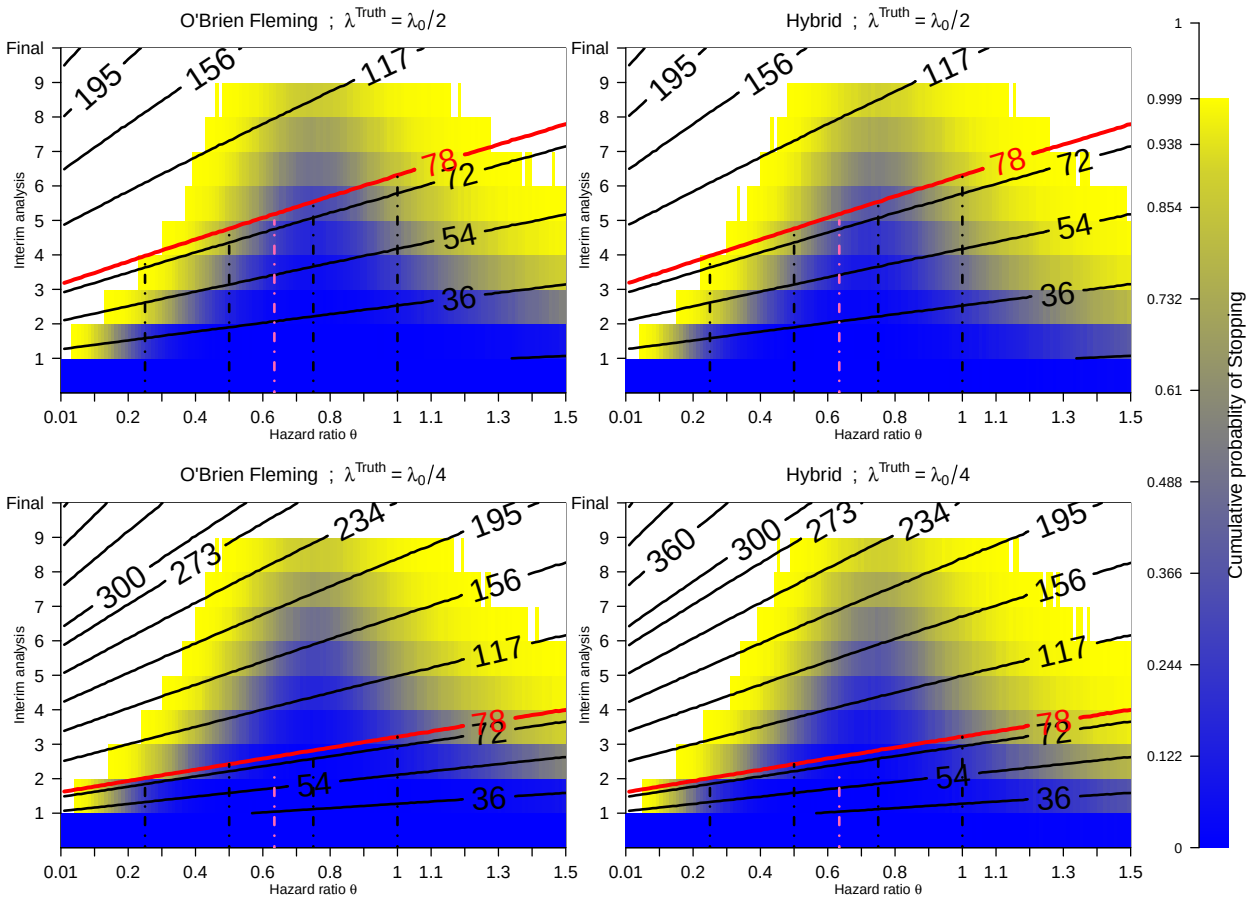

Figure 5.2: Cumulative stopping probability for OBF and hybrid designs overlaid with the average calendar time at various lower baseline event rates. The left column describes this cumulative stopping probability for the *GSDOBF* designs while the right column corresponds to the *GSDHYB* design under plausible low event rate settings. There is a slightly higher cumulative stopping probability for futility for the *GSDHYB* when the hazard ratio θ is in the wrong direction.

In order to avoid having a more prolonged trial than anticipated, a natural strategy that we choose here is to pre-specify the maximum calendar time of stopping. This strategy allows us to incorporate an "escape clause" into the group sequential design. With a sequential monitoring rule, we describe in the next section how to implement such a procedure with the use of error spending functions. Such an approach while avoiding the possibility of a prolonged follow-up, does not overcome the issue of loss of power. To increase the overall power, we later combine the strategy of using blinded sample size revision with pre-specification of the censoring distribution to increase the overall event rates. We describe the logistical concerns with the use of such a fully blinded strategy.

5.2.4 Prespecified Calendar Time of Stopping

Using the group sequential monitoring rule, we can prespecify the maximum duration of the trial, τ^*, to enable us to terminate the trial in presence of low event rates. This prespecified choice of τ^* must be chosen appropriately to enable us to have sufficient follow up data to try to address the primary scientific question on the effectiveness/efficacy of the prevention strategy. The use of the "escape clause" thus addresses the operational constraints and protects against having an unreasonably prolonged follow-up of participants as a direct consequence of extreme treatment effect.

The future analyses of the GSD are dropped when we need to terminate the trial early in presence of low event rates and operational constraints. With a prespecified monitoring boundary, the overall Type 1 error can be adequately controlled with the use of pre-defined error spending functions [Lan and DeMets, 1983] or "well-understood" procedures [Pampallona et al., 1995, Burington and Emerson, 2003] even when these future analyses are dropped [Lan and DeMets, 1983, Proschan et al., 1992]. In essence, when the trial does not attain the maximal number of events by τ^*, and has not been terminated earlier in favor of efficacy or inefficacy, the "escape clause" can be applied by spending the remaining unused α at τ^*.

The Lan-DeMets error spending function [Lan and DeMets, 1983] or constrained boundaries approach [Burington and Emerson, 2003] are some flexible implementations that can be

used during the course of trial monitoring. The error spending function, $\alpha[\Pi_j]$, is a function of Π, defined by the amount of statistical information with respect to the maximum planned statistical information at some interim analyses. This function $\alpha[\Pi_j]$ is a monotonically, non-decreasing function of the amount of statistical information such that at the beginning of the trial, $\alpha(0) = 0$, and when all statistical information is obtained in an event driven trial, $\alpha(1) = \alpha$.

The monitoring procedure is as follows: When the trial does not attain the maximal number of events by the prespecified calendar time, and has not been terminated earlier for efficacy or non-efficacy, we invoke the "escape clause" by spending the remaining $\alpha - \alpha(\Pi_{J^*-1})$ at the maximum calendar time, τ^*, where $J^*(\leq J)$ is now our final analysis. This means that $\alpha(\Pi_{J^*}) = 1$ instead of $\alpha(\Pi_J) = 1$, with Π_{J^*} defined to be our new maximum statistical information at the stopping time.

Figure 5.3 describes some of the revised *GSDOBF* monitoring boundaries under different event rates on the left and its respective power curves shown on the right. The use of the pre-specified maximum calendar time of stopping means that we are stopping a trial much earlier than expected. This induces some loss of power under our hypothesized alternative and can be regarded as a disadvantage.

For illustration, we describe the impact on the overall power when the true placebo rate is half of the original planned rate (Figure 5.4). With a monitoring rule, the trial is protected by the high probability of stopping by the maximum planned calendar time when there is an extreme treatment effect, i.e., $\theta < 0.5$. These contour lines that represent the average calendar time are observed to have doubled for each interim analysis under the design proportional hazards alternative of θ_A. In order to obtain all statistical information under the design alternative, the average calendar time has to double. When the treatment effect is extremely more efficacious than planned $\theta < 0.5$, there is sufficiently high cumulative probability of stopping the trial (as indicated by the bright yellow and white regions) by 78 months. Since we have imposed frequent interim looks in this design, we are protected even with the escape clause, translating to high power in presence of extreme efficacy. However, when our

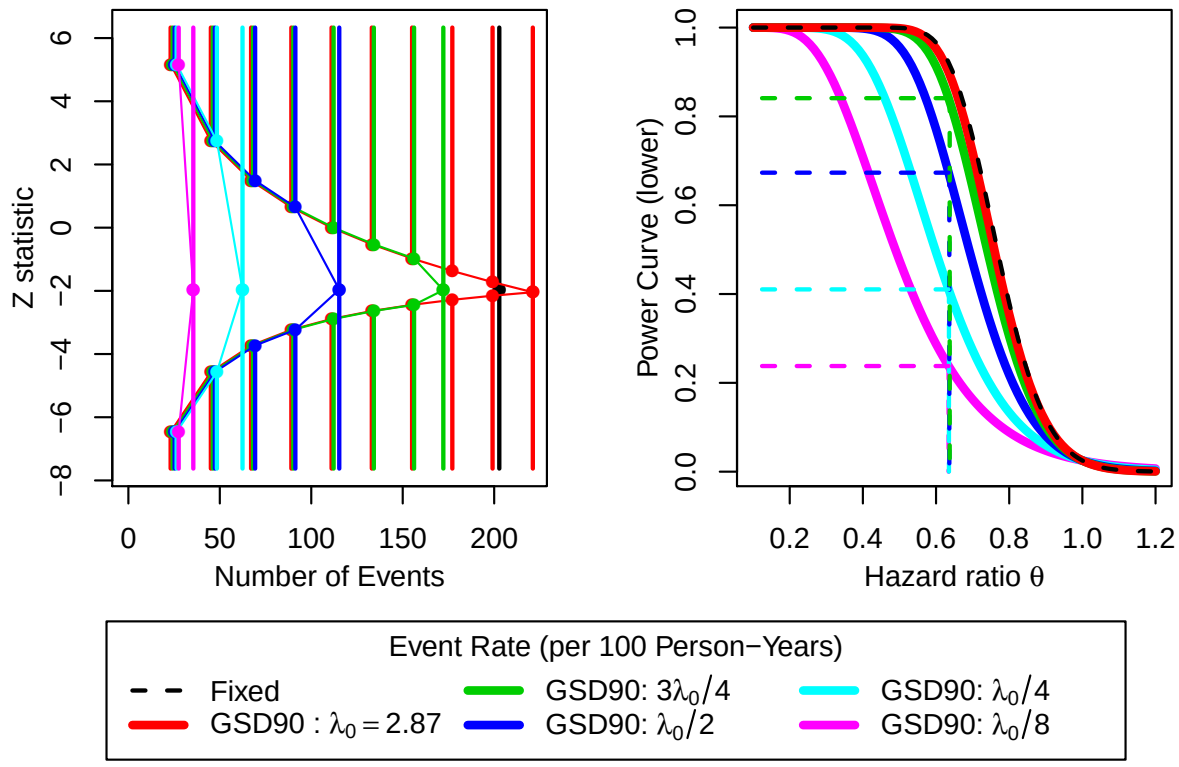

Figure 5.3: Left: Revised boundaries for the various baseline event rate considered for the *OBF* design. Right: Revised power curves for the respective GSDs with the revised boundaries.

treatment effect is moderately effective, i.e., $\theta \in (0.5, \theta_A)$, this cumulative probability of stopping prior to 78 months decreases (as indicated by the blue regions), and consequently we lose power when applying the "escape clause".

5.2.5 Incorporate Blinded Revision of Sample Size

Blinded sample size revision strategies can be employed to increase the overall event rate by either increasing accrual of subjects into the study, and/or extending the calendar time of the trial. The use of this strategy requires an approximate "estimation" of the optimal

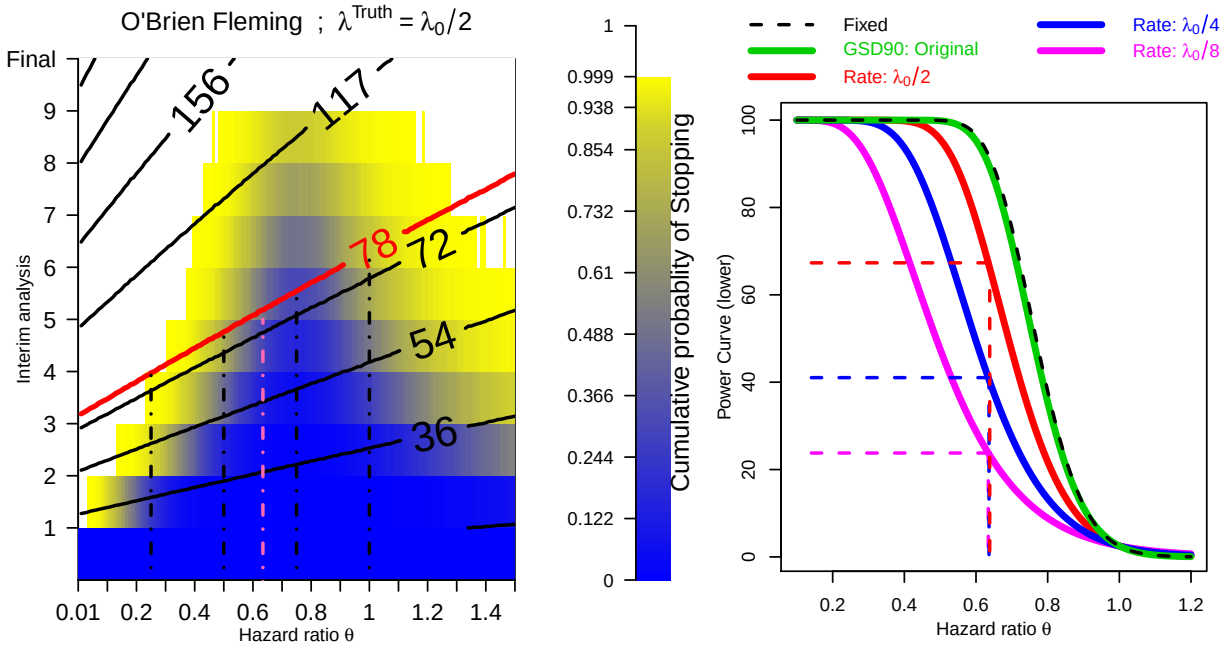

Figure 5.4: Heatmap of the cumulative probability of stopping by the interim analysis overlaid with the contours of the calendar time of stopping (left) and the power curve under the 1/2 the event rate is presented on the right. On average, under the alternative, the average stopping time is twice that of 78 months. The use of frequent interim looks protects us by allowing more opportunities for early stopping prior to 78 months in presence of extreme efficacy even when our hypothesized event rate is incorrect.

time to increase accrual. In practice, the availability of subjects is dependent on the disease setting. From a logistical perspective, performing blinded strategies too early in the trial may be practical but may raise ethical concerns of exposing more subjects to treatments with unknown safety profile. However, choosing to restart accrual much later during the study may unnecessarily encourage speculation of potentially poor treatment response and induce operational bias. On the other hand, when the treatment effect is extreme and compounded with an extremely incorrect event rate, it is arguable that repowering the trial at any point in time is useful.

Many approaches to revise the sample size in a blinded fashion during the course of trial

monitoring have been proposed. Gould and Pecore [1982] and Gould [1992] explored approaches in the immediate setting to revise the sample size for the trial based on interim estimate of the variance when the variance hypothesized at design stage may be incorrect. Since the variance estimate is ancillary to the estimate of the treatment mean, such procedures do not inflate the overall Type 1 error and are regarded as "well-understood" by FDA. Other authors [Whitehead et al., 2001, Mehta and Tsiatis, 2001] have also described approaches to revise the sample size in the time to event setting by maintaining the statistical information in GSDs which were elaborated in section 2.4. Recall the "sample size" formula based on proportional hazards assumption,

$$\begin{aligned} \text{Sample size formula:} \quad & D = \frac{(z_{1-\alpha} + z_{\beta})^2}{(\log(\theta_A))^2} V \\ \text{Maintain statistical information:} \quad & \frac{D}{V} = \frac{(z_{1-\alpha} + z_{\beta})^2}{(\log(\theta_A))^2} \end{aligned} \tag{5.1}$$

Using a $r : 1$ randomization, $V = (r+1)(1/r+1)$. With equal allocation, $V \approx 4$. This formulation under the time to event setting naturally indicates that one can choose to maintain a prespecified maximal statistical information D/V by holding our level α and power β fixed to detect our design alternative $\log(\theta)$. Whitehead et al. [2001] described that such revisions can be based on the estimate of the aggregate event rate $\widehat{\lambda}$ to preserve blinding. This is computed dividing the aggregate number of events with the estimated amount of follow-up based on the expected accrual rate and the current number of patients randomized. Assuming PH, based on the estimated $\widehat{\lambda}$, we can solve for λ_0 using $\lambda = \lambda_1 + \theta\lambda_0$ and $\lambda_1 = \lambda_0\theta$ to determine the additional subjects required.

We consider a modification of the above strategy to increase the overall event rate by increasing the accrual when this number of events falls below some prespecified quantile. Based on our planned baseline event rate of $\lambda_0 = 0.0023$, at 18 months, we expect to increase accrual with some prespecified accrual rate when this accumulated number of events is < 22 events (or < 111 events at 48 months respectively). We consider the timing of the blinded sample size revision to be conducted at 18 months to reflect the strategy of continuing

accrual. Thus, the strategy of restarting accrual at 48 months is thought to be considered a late accrual size revision in the study. Upon conducting this predefined interim analysis, we either drop the subsequent analysis or skip the interim analysis that is less than 3 months prior to the prespecified examination of accrual rate. Following an increase in accrual size, we do not distinguish between subjects who had entered earlier but only contribute events after this added accrual. We assumed that there is no surrogate information available to predict the survival times [Bauer and Posch, 2004]. This is an important assumption made when we are evaluating both blinded and prespecified adaptive design that was discussed in section 2.7.6.

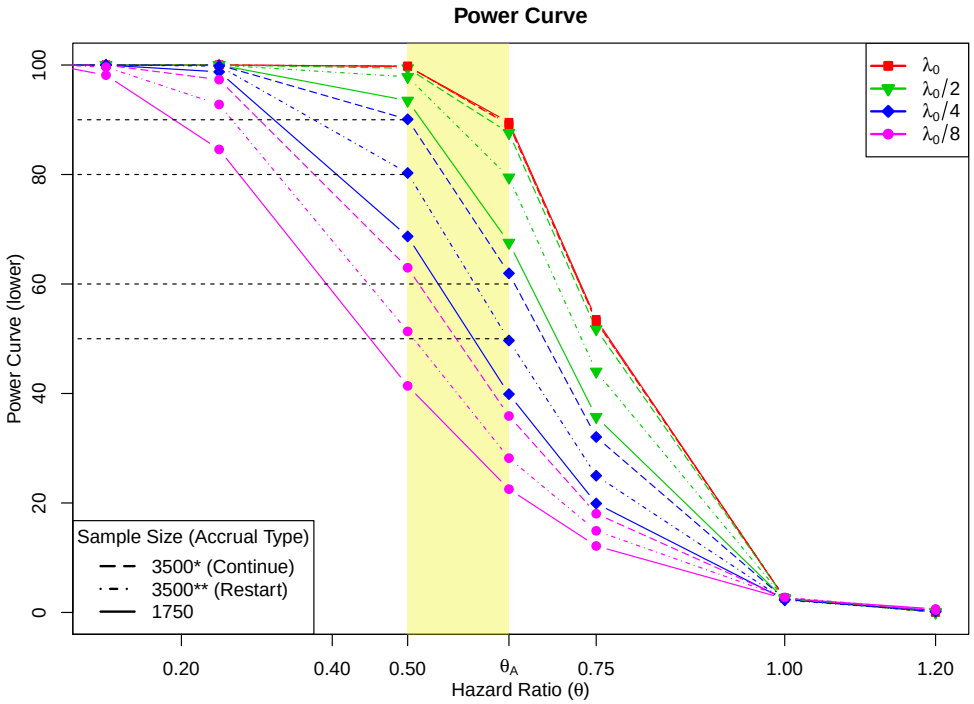

Figure 5.5: Power curves for design *GSDMod* that include blinded adaptation and maximum calendar time of stopping under various true baseline event rates. While the overall power is generally improved under different low baseline rates with the use of blinded adaptations, the yellow region represents moderate efficacy that is of potential clinical concern. In particular, when we have moderate efficacy, i.e., $\theta \in (0.5, \theta_A)$, there is substantial loss of power under markedly low event rates $\lambda_{\text{Truth}} < \lambda_0/2$ even after blinded sample size revision.

5.2.6 Group Sequential Design + "Escape Clause" + Blinded Revision of Sample Size

This comprehensive combined strategy is formulated within the "well-understood" setting with the objective of preserving the study integrity through the use of blinded adaptations using a group sequential design. Using 10,000 simulations, we evaluate the operating characteristics of this comprehensive strategy that include the use of GSD, equipped with blinded revision of sample size when aggregate events are low at a predefined calendar time, and assuming the use of the "escape clause" strategy.

In general, there is an improvement in overall power with this strategy under low event rates (Figure 5.5). The power of the original design, as indicated by the solid blue lines, is approximately 40% under the design alternative at a quarter of the planned rate. An early, one time blinded increase in accrual size at 18 months greatly improves the power from 40% to approximately 60%, while restarting accrual at 48 months boosts the power to only 50%. The disadvantage in continuing accrual is that a much larger accrual size is needed in both the setting of low event rate as well as extreme treatment effect.

Restarting accrual later during the course of the trial can enable one to decrease the need for additional accrual particularly when this observed low "background" rate is a consequence of a stronger treatment effect than hypothesized. However, it is often viewed to be less feasible and difficult in practice. The shaded yellow region is of particular interest to us later since this corresponds to a moderately effective treatment effect. In this research, we now consider whether the use of unblinded interim analysis can help us better identify this source that is either a consequence of low event rate, extreme treatment effect, or a combination of both.

5.2.7 Issues with Fully Blinded Adaptations

Consider the combined strategy in the previous section where a prespecified blinded revision is planned at a calendar time of 48 months. In this example, the first two interim analyses are conducted at 39 and 48 months with the respective total number of events being 22 and

31. At 48 months, this total number of events is lower than what was prespecified (and anticipated), an increase in sample size should be made based on the specified protocol to increase the event rates. We assume no extension of the calendar time is made in this setting. Figure 5.6 shows the original O'Brien Fleming design boundaries (grey) on the Z statistic scale without the sequential path for now. At this blinded interim analysis, our Z statistic is neither too high or too low to enable early stopping.

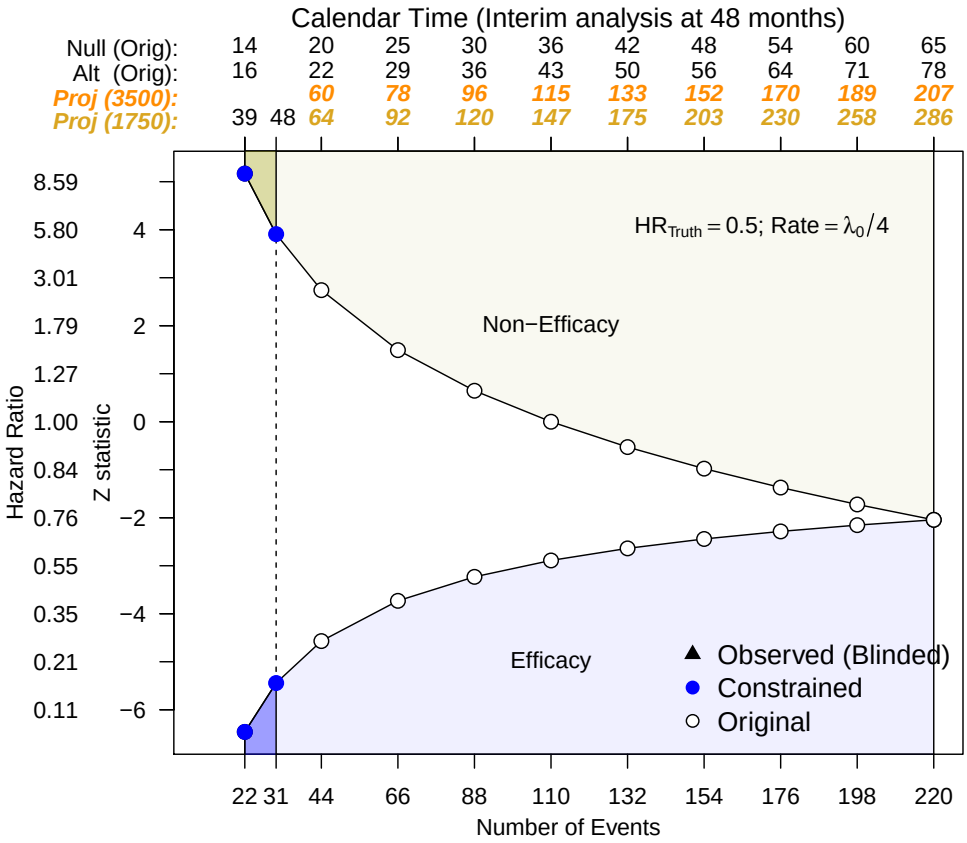

Figure 5.6: Sequential design for a simulated realization of a clinical trial based on true hazard ratio of 0.5 and baseline rate $\lambda_0/4$ (sequential path is not shown) where an interim analysis is conducted at 48 months to potentially increase accrual. Both the original schedule and projected schedule of analyses are shown to illustrate how far off one is in this particular realization of a clinical trial. The projected schedule of analyses can be estimated based on the total number of events and the aggregate follow-up based on initial trial assumptions.

Table 5.2: Table of efficacy and futility boundary based on OBF monitoring rule under the common scales Z, sample mean θ, 1-sided fixed P-value scale, and the error spending scale E. The expected calendar time under specific $\theta = \{0.5, \theta_A, 1\}$ are presented. Comparing the expected number of events to be obtained at 48 months with the example in Figure 5.6, there is clear evidence of low event rate.

Analysis		Efficacy				Mean Time			Futility			
No	Events	Z	θ	P	E	0.5	θ_A	θ_0	Z	θ	P	E
1	22	-6.461	0.06	1.0000	0.0000	16	15	14	5.169	9.06	0.0000	0.0000
2	44	-4.569	0.25	1.0000	0.0001	23	22	19	2.741	2.29	0.0031	0.0001
3	66	-3.731	0.40	0.9999	0.0039	31	29	25	1.492	1.44	0.0678	0.0039
4	88	-3.231	0.50	0.9994	0.0261	38	36	30	0.646	1.15	0.2591	0.0261
5	110	-2.890	0.58	0.9981	0.0861	45	42	36	0.000	1.00	0.5000	0.0861
6	132	-2.638	0.63	0.9958	0.1961	53	49	42	-0.528	0.91	0.7011	0.1961
7	154	-2.442	0.67	0.9927	0.3590	61	56	47	-0.977	0.85	0.8357	0.3590
8	176	-2.284	0.71	0.9888	0.5707	69	63	53	-1.371	0.81	0.9148	0.5707
9	198	-2.154	0.74	0.9844	0.8136	77	71	59	-1.723	0.78	0.9576	0.8136
10	220	-2.043	0.76	0.9795	1.0000	85	78	65	-2.043	0.76	0.9795	1.0000

Highlighted in yellow are the interim analyses for potential hazard ratios where the calendar time is approximately 48 months.

Based on the total number of events and the total amount of follow-up thus far, we can project the future calendar time of analyses assuming (a) no increase in accrual is made, or (b) an additional 1,750 subjects are enrolled into the trial. This projected schedule, as shown in Figure 5.6, suggests that it is more feasible to increase accrual to stop on average by 78 months, with an average of 66 events, as opposed to concluding the trial with a much lower number of events. Compared to Table 5.2 as well as the original schedule defined on the top row of Figure 5.6, we may suspect our event rate to be a quarter of what was planned if our design alternative is true.

There are arguments against the use of fully blinded adaptations in the low event rate setting in the above example. Even though the calendar time of this adaptation is considered rather late in the study, it is otherwise early on the information time scale ($\approx$15% of the total amount of statistical information). Additionally, at this interim analysis, our continuation

region is rather uninformative since it ranges from less than -5 to approximately 4 on the Z scale. If our current Z statistic is close to the efficacy/futility boundary, i.e., the event rate is low and the absolute Z statistic was large (with either low HR or high HR), it may no longer be ethical to randomize more subjects into the trial since the future trend of the data may allow early termination by the calendar time of interest. However, if this current Z statistic is further away from the monitoring boundaries, and our event rates are truly low, then this may necessitate increasing accrual to gather more information about the treatment effect.

Had our estimated treatment effect been between the null and alternative, it may seem reasonable to increase accrual since we can potentially rule out extreme treatment effect. Alternatively, had our estimated treatment effect appear sufficiently strong such as, $\widehat{\theta} = 0.5$, we may presumably stop within some reasonable calendar time if the trend persists. Thus, a potential adaptive decision is to stay the course. One may then argue that if the event rate is truly a quarter of what was planned, and the treatment effect may be moderately effective, then it may be reasonable to increase accrual to obtain a more reliable estimate of the treatment effect with better precision by some maximum calendar time.

Thus, in presence of low "background" rate, there may be a stronger ethical rationale to decide the adaptation on the basis of the unblinded interim data to hopefully distinguish between low event rate, or extreme treatment effect, or both. If one can precisely identify the reason for this low "background" rate, and make the appropriate adjustments to either stay the course or increase accrual, it seems plausible that an adaptive design with a carefully planned adaptive rule should potentially beat the fully blinded strategy using a GSD.

To some extent, treatment effects that are less effective than our hypothesized alternative are of less important in a setting where current strategies already confer benefit. This leads us to place emphasis in the shaded region in Figure 5.5 that will be of interest when we further consider flexible adaptive designs under these scenarios. To the extent that the best comprehensive GSD strategy does not further improve our overall power under moderately effective treatment effect and markedly lower event rates, we investigate whether the use of unblinded interim results with more flexible adaptations can potentially help us better

distinguish between low event rate vs extreme treatment effect.

5.3 Adaptive Design

In recent years, there has been considerable interest in the possibility of incorporating adaptive features by unblinding interim results to make design changes. Several authors have found minimal gains in efficiency with the use of adaptive designs in comparison with group sequential designs in the settings where the sample size is a surrogate for statistical information [Mehta and Tsiatis, 2001, Jennison and Turnbull, 2006a, Levin et al., 2013] . The advantage of the use of pre-specified adaptive rules is that they are based on minimal sufficient statistics and thus mitigate issues of inefficiency where there is a need to adjust for the unblinded adaptations [Levin et al., 2013, 2014]. Additionally, while frequentist inference can be conducted in presence of early stopping with these prespecified adaptive designs, fully flexible adaptive designs have limited inferential procedures to account for this early stopping [Brannath et al., 2009, Gao et al., 2013, Levin et al., 2014].

In the survival/time to event setting, there has been even less comprehensive evaluation of adaptive procedures in the low event rate setting. Prior research by Emerson et al. [2011a] saw some benefit in limited censored time to event settings. However, lacking in their evaluation was whether this benefit still persists when using weighted statistics that require further adjustment to control the overall Type 1 error as often explored in earlier adaptive literature [Schäfer and Müller, 2001, Shen and Cai, 2003, Togo and Iwasaki, 2011]. These adaptive sampling strategies are also obfuscated by poor choices of analysis schedule or sub-optimal rules, and thus lead to poor understanding of the benefits/limitations of adaptive designs. Particularly, adaptive rules are often recommended to be made late into the study. Under low event rates, this may no longer be justified or appropriate. In section 3.2, we explored this consequence in the FSD to gain understanding on how the timing of adaptation can affect the degree of overall power loss when comparing adaptive designs that use minimum sufficient statistics vs weighted statistics that require further adjustments.

Our goal now is to investigate the use of group sequential rules together with the use of

unblinded interim treatment effect to determine whether these unblinded adaptations can better guide us when confronted with low "background" rate at some interim analyses. In settings of unanticipated extreme treatment effect or low event rates, we want to compare the best strategy the adaptive trialists may consider that can lead to increasing accrual vs the best GSD to appropriately characterize the benefit/loss in overall power while holding other operating characteristics as constant as possible. Since the calendar time dictates the duration of the trial and thus the overall cost of the trial, it is important to consider the trade-offs in calendar time, power, total sample size, and the event size when accepting a smaller event size with any such adaptations. In practice, the statistician who is in favor of adaptive strategies may choose to extend the calendar time of the study. However, as pointed out by several authors [Bauer and Posch, 2004, Jenkins et al., 2011, Magirr et al., 2014, 2016], there are major statistical issues that need to be carefully addressed when incorporating extension of follow-up into the analysis.

5.3.1 Prespecified Adaptive Design

There is ethical rationale in favoring an adaptive design over the GSD in the low "background" rate setting. In prevention settings, disease incidence rates can vary from one region/country to another. Even though such knowledge can be revised from prior Phase 2 studies or burn-in pilot Phase within the Phase 3 settings, they may not be reliable when the trial is later conducted in a bigger setting.

Within the framework of unblinded adaptations, adaptive strategies can be classified based on using either minimal sufficient statistics or weighted statistics. The use of a pre-specified unblinding, or flexible strategy to better understand the cause of the low event rate in these settings may thus seem more appropriate in a clinical setting. However, it is necessary that such a decision rule be envisioned and specified to refine the design by adaptively increasing accrual or accepting a smaller event size, draw sensible conclusions while preventing operational bias from having knowledge of the unblinded study results. Emerson et al. [2011a] and Levin [2013] have shown that it is entirely possible that an adaptive

sampling scheme can be completely pre-specified so that minimum sufficient statistics can be used at the stopping time. However, many proposed adaptive sampling schemes are not based on minimal sufficient statistics or are chosen in some sub-optimal manner that lack comprehensive evaluation with other competing designs.

In this exploration, we focus on monitoring rules that are commonly chosen to address competing scientific and ethical concerns. Since sponsors often value calendar time as one of the limiting constraints in time to event settings, conservative boundaries such as OBF allow eliminating non efficacious treatments early and re-investment of public resources to other potential clinical trials with promising agents. These boundaries are often not optimal in terms of ASN but address many competing goals in the design of a clinical study. Thus, for this exposition, we limit exploration to common boundary choices such as the OBF and/or the hybrid designs to form the basis of our comparison. In Chapter 3, we highlighted some of the difficulties in comparing several GSDs and this holds true even more in the class of adaptive design.

5.3.2 Statistical Issues with Unblinded Interim Analysis during Monitoring

In this section, we described the "flexible" version of the GSD design that has a prespecified, prospectively planned opportunity to use unblinded interim results to possibly (1) decide on modifying aspects of the accrual size when event rates are low, or (2) decide on modifying aspects of the accrual size with pre-specification on the use of both estimated treatment effect and overall event rates. These fully adaptive sampling schemes are based on weighted statistics via the use of some form of conditional error functions [Proschan and Hunsberger, 1995, Schäfer and Müller, 2001].

Any form of unplanned, unblinded adaptations during the conduct of clinical trial have been shown to substantially inflate the overall Type 1 error [Proschan and Hunsberger, 1995]. To protect against inflation of the overall Type 1 error, several authors [Bauer and Köhne, 1994, Proschan and Hunsberger, 1995, Cui et al., 1999, Lehmacher and Wassmer, 1999, Schäfer and Müller, 2001, Jennison and Turnbull, 2003, Chen et al., 2004, Müller and

Schäfer, 2004] have proposed similar methods to specify control of the conditional error under the null hypothesis. The conditional rejection principle (CRP) in Müller and Schäfer [2004] is used to control for the overall Type 1 error by preserving the conditional error in a GSD. That is, the probability of incorrectly rejecting the null hypothesis conditional on the current interim estimate.

When the design is "unchanged", i.e., after we have/have not adapted, the future monitoring boundaries should be monitored on the conditional error scale. More specifically, with a fully flexible adaptive design, any form of unblinded look leading to an adaptation or staying the course would be considered an adaptation if this ultimately lead to an early stopping at some maximum calendar time of stopping. As such, it is necessary to make statistical adjustments after an unblinded interim analysis.

5.3.3 Notation

Recall in section 2.6, we introduced briefly the conditional rejection principle (CRP) approach by Müller and Schäfer [2001] that allows for the control of the overall Type 1 error with the use of sequential monitoring if the future course of the analysis schedule is altered based on unplanned unblinded adaptations. The application of the CRP under the low event rate setting can be applied when the maximum calendar time is used as a stopping criterion after an unblinded adaptation that may or may not lead to changing aspects of the design.

To ensure comparability with GSD, we define our continuation regions in the form of group sequential monitoring boundaries by considering the approach in section 2.5. These continuation regions are decision rules that are based on some joint function of the average event rate and sufficient statistics usually defined based on the interim estimated treatment effect. The estimated treatment effect can then either be the difference in event rates $\widehat{\lambda}_1 - \widehat{\lambda}_0$, or the estimated hazard ratio $\widehat{\theta}$. At the h^{th} interim analysis (either at 18 months or 48 months), accounting for the possibility of low event rate, our continuation region can then be partitioned into $\mathcal{C}_h = \mathcal{C}_h^1 \cup \mathcal{C}_h^2$, where $\mathcal{C}_h^1 = [A, D]$ and $\mathcal{C}_h^2 = (a_h, A] \cup [D, d_h)$.

Under low event rates, we can base these continuation regions using some function of θ^h

and λ^h, where λ^h is our average event rate that is defined by the total number of events divided by the total follow-up at the interim analysis for all participants accrued thus far. The sampling plan at the unblinded interim analysis is defined as such:

- If $\widehat{\theta} \in \mathcal{C}_h^1$ and $\widehat{\lambda} \leq \lambda^h$, we increase accrual for the next 18 months using the same accrual rate based on the first N subjects.
- If $\widehat{\theta} \in \mathcal{C}_h^1$ and $\widehat{\lambda} > \lambda^h$, we stay the course.
- If $\widehat{\theta} \in \mathcal{C}_h^2$, we stay the course.

When our sampling strategy involves the difference in event rates $\lambda_{\text{Diff}} = \lambda_0 - \lambda_1$, the sampling plan is not a direct function of our continuation regions. We can either define some regions based on λ_{Diff}^h such that values of $\widehat{\theta}$ do not lead to an early stopping at the h^{th} interim analysis or apply proportional hazards assumption and equate $\lambda_{\text{Diff}} \approx \lambda_0 - \lambda_0\theta$. Under PH assumption, the pair of values, λ^h and λ_{Diff}^h, can be used to characterize the adaptive rules so that we can then map them back to the continuation regions that is defined on the hazard ratios. For simplicity, we assume our adaptive rules is defined based on $\lambda_{\text{Diff}}^h, \lambda^h$ in our simulation study. The adaptive sampling plan based on λ_{Diff} and λ is as follows:

- If $f(\widehat{\lambda}_{\text{Diff}}, \widehat{\lambda}) \in \mathcal{C}_h^1$, we increase accrual for the next 18 months at the same accrual rate.
- If $f(\widehat{\lambda}_{\text{Diff}}, \widehat{\lambda}) \notin \mathcal{C}_h^1$, we stay the course.

5.3.4 Optimization Procedure

Using the 10-look GSD, we hold fixed the prespecified interim analysis where an additional accrual is made based on the blinded data, and proceed to find the best fully adaptive design. We further constrain all adaptations to have some known probability of p of increasing accrual. We then use a grid search to find the best rule with this known probability p of adapting to a bigger accrual size that requires further adjustments. We call this rule to be "sup-optimal" since this adaptation is presumed to first be unplanned and thus based on the sufficient statistic, such as estimated hazard ratio, or difference in event rates, which

is selected among the class of fully flexible adaptive designs that require further statistical adjustments.

Among this larger class of designs (bigger than the class of pre-specified adaptive designs), we can then find the "best sub-optimal" rule that maximizes this overall adjusted power. Using this best "sub-optimal" rule that maximizes the overall power of the fully flexible adaptive design, we can then pre-specify this rule so that this is now based on minimal sufficient statistics, and thus can be evaluated as if this was a prespecified adaptive strategy following Emerson et al. [2011a] and Levin [2013]. Our approach of optimizing this reweighted rule eliminates comparison with some non-optimal, inefficient sampling scheme, while holding fix the interim analyses for which the unblinding is to be made. By doing so, we can then evaluate adaptive designs in the best possible light.

5.3.5 Conditional Monitoring Example

We illustrate the differences in trial conclusions with the use of the CRP procedure vs the usual monitoring procedure based on the earlier example shown in Figure 5.7. In this unblinded approach, the same interim analysis is conducted at 48 months and an adaptive decision is made to decide whether an increase in accrual is necessary. The CRP procedure is applied to preserve the overall Type 1 error to account for this unplanned decision.

Under the fully blinded adaptation, an adaptation to increase accrual is made and the full sequential path is shown in Figure 5.8. After an interim analysis is conducted at 48 months, subsequent analyses are conducted at 66 months (with 44 events, and $Z = -2.465$) and then at 78 months (with 63 events and $Z = -2.456$). In the blinded setting, the group sequential boundaries are revised appropriately based on the error spending approach. By 78 months, the calendar time is up and since all 220 events are not attained, we apply the error spending approach to spend the remaining unused error and treat this calendar time as our final analysis when there is a total of 63 events. The final revised critical value is -1.96, leading to an efficacy decision upon comparison with our Z statistic ($Z = -2.456$) based on the data collected by the end of 78 months.

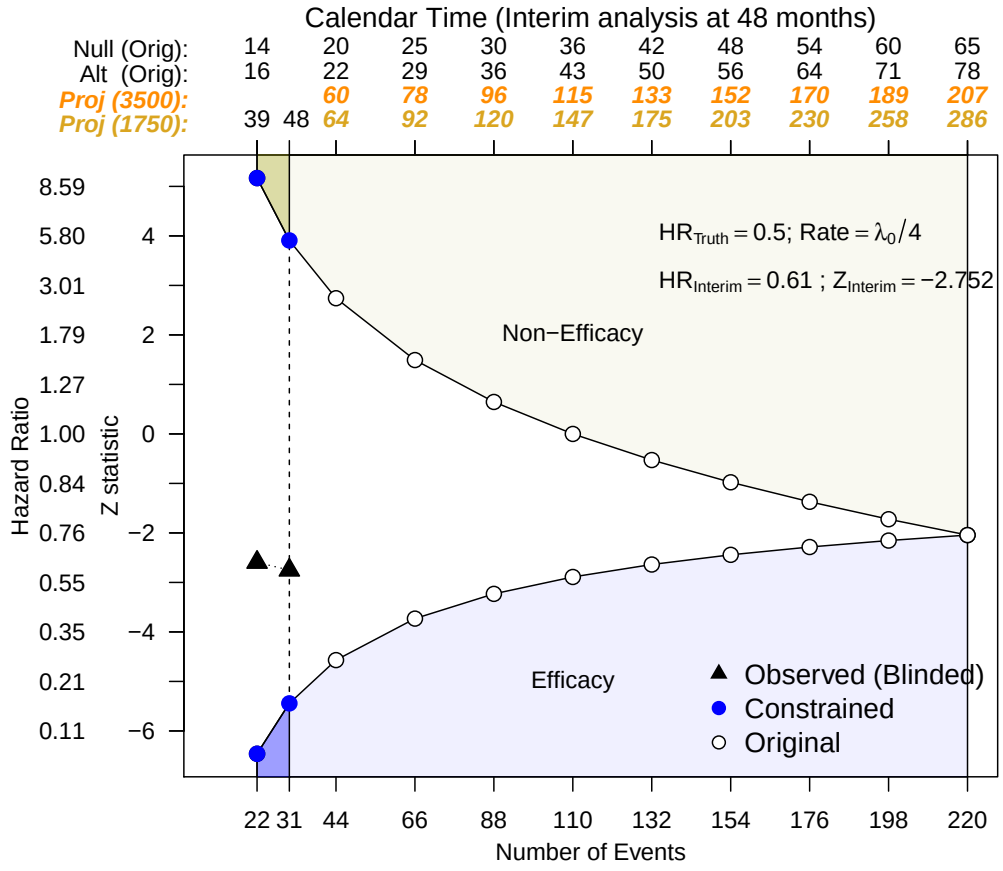

Figure 5.7: Sequential path for a simulated realization of a clinical trial where an interim analysis is conducted at 48 months to potentially increase accrual. We describe two potential strategies whereby (a) blinded revision of sample size is performed if the event rate is too low, or (b) consider unblinded revision of sample size with the treatment effect to determine whether the low event rate is due to treatment effect or background rate.

For the unblinded adaptation, the future monitoring rule has to be reweighted accordingly using CRP to preserve the overall Type 1 error rate. This means that the first 31 events is weighted as 14% of the total data (relative to 220 events), leaving the remaining 86% of the weight to be redistributed among the future number of events one can get between 48 and 78 months. The conditional Type 1 error on the basis of the observed data, $Z = -2.7515$ is 0.1961. Conditional on 48 months (31 events), the future boundary at 66 months is

unadjusted since the event size of 44 was planned for this interim analysis.

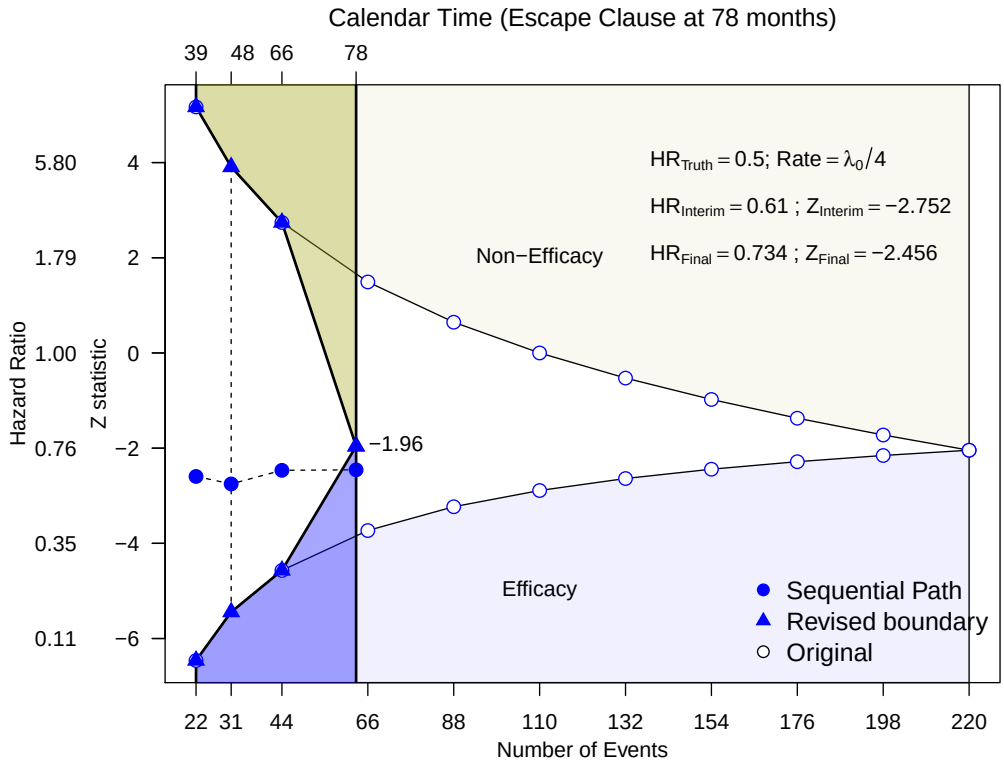

Figure 5.8: Sequential path (extended from Figure 5.7 for the same simulated realization) of a clinical trial where an interim analysis is conducted at 48 months to increase accrual. Blinded adaptation leads to increasing accrual of more subjects. At the final analysis, based on the strategy described in section 5.2.6 (i.e., using constrained boundaries monitoring, with a prespecified fully blinded adaptation, and application of the "escape clause"), the trial terminates with a conclusion for efficacy.

By applying the CRP, our "new" design starts at a calendar time of 48 months with $\Pi_0^C(0)$, and the maximum statistical information defined by 189 events if there was no change to the design. At 66 months, relative to this (conditional) maximum statistical information of 189 events, the information fraction is 6% for this "new GSD" that has a conditional Type 1 error rate of 0.1961. Our design is nonetheless "unchanged" at this point and our original maximum statistical information is still 220 events.

When we stopped at the maximum calendar time, the CRP approach also has to spend the remaining unused conditional error at this final analysis. Based on the original design, had this been a fully blinded adaptation, each of these newly accrued 32 events would be weighted similarly as each of the first 31 events. With early termination using the calendar time, these 63 events in the fully blinded GSD setting are weighted equally.

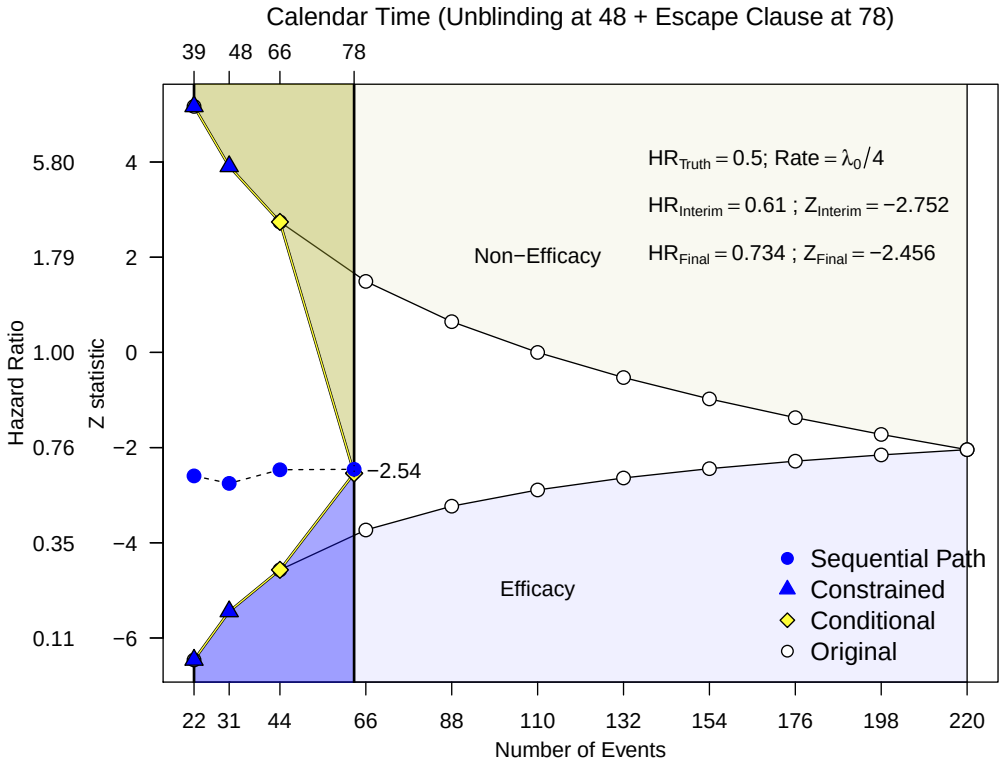

Figure 5.9: Sequential path for the same simulated realization as in Figure 5.8 of a clinical trial where an interim analysis is conducted at 48 months to increase accrual. For the same simulated sequential path, an unblinded interim analysis leads to the same decision to increase accrual of subjects. The flexibility in the design requires conditional monitoring to control the conditional Type 1 error. At the final analysis, when the trial terminates early with the same sample path, a futility conclusion is obtained as opposed to an efficacy decision.

With an unplanned unblinded design, the weighting scheme now differs when we apply

the CRP which weights the events non linearly using the conditional error function. We provide a simple explanation by first considering the remaining unused information fraction (86% relative to the original design) assuming the naïve Z statistic is used at the end of the trial. If we assume that these 63 events correspond to the original maximum statistical information, then the new 32 events should correspond to 51% of the total weight rather than 86% of the weight assuming 220 events as the original event size. So, in order to control for the overall Type 1 error, the CRP approach must adaptively re-weight these events. This results in a more extreme final critical value of -2.5399, thus leading to a conclusion for non efficacy (shown in Figure 5.9). Although this example is a simulated realization, we can anticipate the impact on the overall power when applying the use of this flexible adaptation during the course of trial monitoring.

5.4 Simulation Study Comparing Fully Adaptive Designs, Pre-specified Adaptive Designs, and GSDs

In the prevention setting, it is of particular concern whether extension of the study, by increasing the follow-up of the subjects, can affect the scientific hypothesis of interest. For example, in the setting of HPTN052, the study requires having at least 60 months of follow-up to avoid the possibility that "any short-term interruption of the transmission of HIV virus may be a direct consequence of delaying infection due to the potential for more resistant variants" as defined in the protocol of HPTN052 in the supplementary material of Cohen et al. [2011].

Thus, any evaluation of HIV transmission over the defined follow up should provide sufficient long term information about the effectiveness and public health utility of this therapy. If the trial was extended in a manner that participants are followed for longer than the average specified follow up, the scientific question would be changed and the results of the trial may no longer be addressing the primary objective. As such, we focus on describing our results in the setting where the maximum calendar time is not extended (A2 which is in bold). In our exploration, we also considered various scenarios as such:

A) No extension of study calendar time (Possibly answering the same scientific question)

   1. No extension of accrual size
   2. **Allow extension of accrual size (at most twice the original sample size while adhering to accrual specifications)**

B) Allow extension of study calendar time (by most 50%) (May no longer answer the same scientific question)

   1. No extension of accrual size
   2. Allow extension of accrual size (at most twice the original sample size while adhering to accrual specifications)

Briefly, we conducted 10,000 simulations to evaluate comparisons between the best GSD, the best prespecified adaptive designs, and the best fully adaptive designs. We simulated patient survival data assuming an exponential distribution with baseline rate parameter $\lambda_0 = 0.002395$, with uniform accrual over a period of 18 months. We considered the following combination of $\theta \in \{0.04, 0.025, 0.5, 0.6343, 0.75, 1.0, 1.2\}$ and $\lambda_{\text{Truth}} = \{1/8, 1/4, 1/2, 3/4, 1\}\lambda_0$. We then analyze patients as according to the sequential monitoring rules and use the logrank test statistics at each interim analyses. In our simulation study, when we consider more extreme event rates (such as $\lambda_0/8$), it is possible that there are no events observed at the predefined interim analysis used to revise the accrual size when this accrual is made early. As such, we chose to increase the accrual size and maintain the accrual rate and only analyze the trial at the prespecified maximum calendar time.

When imposing the calendar time as a constraint, we may refer to the fixed sample design described earlier with 18 months of accrual as *FSD078* or *FSD117* where 78 and 117 represent the total duration of the trial on the calendar time. Several ideal designs can be constructed to serve as a reference as we try to quantify the potential gain or loss in power across the blinded vs unblinded strategies. In other words, we can construct the optimal strategy of having planned the design with an accrual size of $2N$. We can (a) accrue subjects uniformly over the first 36 months, or (b) accrue the first N subjects over the first 18 months and then restarting accrual of the remaining N subjects at 48 months. In both

settings, we hold the accrual rate similar. These designs with different accrual patterns thus reflect the ideal situation when we plan for a bigger accrual size of $2N$ while holding the prespecified maximum calendar time of stopping at either 78 months (*GSD3500A*), or 117 months (*GSD3500B*). This allows us to understand the maximum attainable power possible had we always decided to accrue patients uniformly over the first 36 months.

The above two strategies have relevance to our design choices later. The first corresponds to having planned the optimal strategy with an accrual size of $2N$ with subject enrollment performed uniformly over the first 36 months. This strategy is similar to continuing accrual when invoking blinded adaptations at early interim analysis. The second describes the optimal strategy where accrual is conducted twice, the first enrollment is conducted within the first 18 months while the remaining accrual is conducted uniformly between 48 months and 66 months. This represents the setting of restarting accrual later in the study. These designs thus allow us to assess whether our overall power for the various θs are maximized. Hence, any additional gain in power should ideally be due to appropriate adaptations.

We summarize our main results for setting A2 in the next section. We refer the interested reader to Appendix D for the other results.

5.4.1 Results for Setting A2

We shall denote the GSD with the "escape clause" strategy, incorporating a possible blinded revision of sample size as *GSDMod*. Later, this *GSDMod* has the dual interpretation as a fully prespecified adaptive design. What this means is that if one had completely prespecify all the potential adaptation one would want to make during the conduct of the trial, as well as having prespecify the procedure when there is low event rate, then we can use the minimum sufficient statistics at the end of the study. Thus, the results for the fully blinded strategy (*GSDMod*) is exactly the same as that of a prespecified adaptive design.

If such prespecified unblinding does not have a properly defined adaptive rule, and that no adjustments to the design have been made, the conservative critic can suspect any of the following potential adaptations were contemplated and evaluated:

1. They saw something that was sufficiently close to statistical significance and so decided not to increase accrual with the hope for statistical significance at the calendar time of stopping, or

2. They saw that the event rate was low and the results were far from statistical significance and choose to increase accrual with the hope that at the maximum calendar time, one may obtain statistical significance.

Thus, we pay a price for allowing this flexibility as seen in the adjusted analysis (denoted as *GSDCond*) where we apply the CRP approach to account for this unplanned opportunity. The results are shown in Table 5.3. Compared to *GSDMod*, we pay some penalty for allowing such flexibility in the protocol with a loss in overall power (<5% relative to the *GSDMod*) when the true event rates are markedly lower ($\lambda > \lambda_0/4$) and moderately effective ($\theta \in (0.5, \theta_A)$). This observation holds true regardless of whether this adaptation was made early or late. There is however a substantial loss in overall power (> 10%) with the use of fully flexible adaptive design when this event rate is more extreme, i.e., $\lambda \leq \lambda_0/4$, relative to *GSDMod*.

One can claim that the above adaptation is unfair since we did not incorporate any rule to attain the best possible adaptation or provide the adaptive strategy the best sampling scheme to make the adaptation. The proportion of adaptations to a larger accrual size is consistently determined by the ad-hoc rule based on the overall number of events rather than the estimated treatment effect. In particular, under the alternative, when the true baseline rate is markedly lower ($\leq \lambda_0/2$), we increase accrual 100% of the time. Thus, it is unfair since we have not allow the adaptive design to make more realistic adaptations based on interim treatment results. As described earlier, we thus have to decrease this probability of adaptation to some $p < 1$ so as to realistically compare GSDs with adaptive designs in a fair manner.

Table 5.3: Table of power for *GSDMod* (One-sided symmetric OBF design with 90% power, allowing for blinded adaptations, and use of "escape clause") vs *GSDCond* and GSD Ref. There is generally a loss of power when applying the CRP as a consequence of early termination of the trial at the maximum calendar time.

		78 Months						117 Months					
		Continue			Restart			Continue			Restart		
		Blinded		Adaptive	Blinded		Adaptive	Blinded		Adaptive	Blinded		Adaptive
		GSD Ref	*GSDMod*	*GSDCond*	GSD Ref	*GSDMod*	*GSDCond*	GSD Ref	*GSDMod*	*GSDCond*	GSD Ref	*GSDMod*	*GSDCond*
			$N_{Max} = 3500$			$N_{Max} = 3500$			$N_{Max} = 3500$			$N_{Max} = 3500$	
$\theta = 0.04$	$\lambda_0/8$	100	100	100	99.96	99.96	99.75	100	100	100	100	100	100
	$\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.1$	$\lambda_0/8$	99.99	99.99	99.97	99.63	99.63	98.46	100	100	100	100	100	100
	$\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.25$	$\lambda_0/8$	97.33	97.33	97.19	92.8	92.8	86.84	99.94	99.71	99.7	99.68	99.68	99.49
	$\lambda_0/4$	100	100	100	99.8	99.8	99.54	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.5$	$\lambda_0/8$	62.97	62.97	62.01	51.34	51.34	44.65	83.89	83.08	82.84	77.29	77.29	75.61
	$\lambda_0/4$	90.08	90.08	89.72	80.27	80.27	76.88	98.75	98.76	98.75	96.95	96.95	96.94
	$\lambda_0/2$	99.49	99.41	99.41	97.84	97.84	97.78	99.77	99.69	99.69	99.79	99.79	99.79
	$3\lambda_0/4$	99.79	99.47	99.46	99.72	99.73	99.73	99.79	99.47	99.46	99.84	99.85	99.85
	λ_0	99.82	99.8	99.8	99.81	99.81	99.81	99.82	99.8	99.8	99.81	99.81	99.81
$\theta = \theta_A$	$\lambda_0/8$	35.89	35.89	35.33	28.19	28.19	24.8	53.45	52.99	52.72	46.85	46.85	45.75
	$\lambda_0/4$	61.95	61.95	61.83	49.68	49.68	47.37	82.82	82.82	82.8	76.38	76.39	76.4
	$\lambda_0/2$	88.4	87.61	87.6	79.47	79.47	79.51	90.22	89.34	89.32	90.41	90.41	90.41
	$3\lambda_0/4$	89.83	86.6	86.62	90.06	89.76	89.75	89.83	86.6	86.62	90.4	90.08	90.08
	λ_0	89.96	89.05	89.05	89.88	89.17	89.17	89.96	89.05	89.05	89.88	89.17	89.17
$\theta = 0.75$	$\lambda_0/8$	18.04	18.04	18.09	14.92	14.92	13.21	27.34	27.19	27.3	23.57	23.57	23.02
	$\lambda_0/4$	32.05	32.05	31.92	24.98	24.98	23.64	48.15	48.14	48.15	42.04	42.06	42.24
	$\lambda_0/2$	53.08	51.75	51.73	44	44.01	43.93	54.15	52.88	52.87	54.37	54.39	54.39
	$3\lambda_0/4$	53.85	50.9	50.85	53.89	52.9	52.88	53.85	50.9	50.85	53.94	52.95	52.93
	λ_0	54.03	53.16	53.16	53.57	53.17	53.16	54.03	53.16	53.16	53.57	53.17	53.16
$\theta = 1$	$\lambda_0/8$	2.6	2.6	2.49	2.75	2.75	2.46	2.67	2.69	2.66	2.59	2.59	2.65
	$\lambda_0/4$	2.42	2.42	2.38	2.25	2.25	2.29	2.73	2.73	2.74	2.71	2.71	2.76
	$\lambda_0/2$	2.75	2.63	2.62	2.48	2.49	2.5	2.76	2.64	2.63	2.69	2.7	2.7
	$3\lambda_0/4$	2.75	2.56	2.56	2.54	2.58	2.57	2.75	2.56	2.56	2.54	2.58	2.57
	λ_0	2.84	2.57	2.57	2.55	2.69	2.64	2.84	2.57	2.57	2.55	2.69	2.64

GSD Ref design corresponds to the setting when an accrual size of 3500 was planned right from the start with the continuous, uniform accrual over 36 months (Continue) or uniform accrual over first 18 months for 1750 subjects and restarting later at 48 months to accrue uniformly 1750 subjects over another 18 months (Restart).

GSDMod can be interpreted as a pre-specified adaptive design that makes the same adaptation as blinded repowering.

GSDCond is the fully adaptive design with conditional monitoring.

5.4.1.1 Adaptive Rule Based on A Hazard Ratio of 0.5 and Baseline Event Rate of $\lambda_0/4$

We describe results based on the OBF boundaries. We arbitrary set p to be 80% in the adaptive design and investigate the setting under moderate efficacy $\theta \in (0.5, \theta_A)$ and markedly lower event rate, i.e., $\lambda \in (\lambda_0/4, \lambda_0/2)$ to determine the best adaptive rule. To appropriately compare results obtained from the adaptive design with a lower rate of adapting under moderate efficacy, we find the average event rate such that the GSD only makes a blinded increase in accrual 80% of the time. This then allows us to match all other operating characteristics (average accrual size, average calendar time, and average event size) so as to appropriately compare the best adaptive design with the GSD under low event rate and extreme treatment effect setting. By further holding the timing of this adaptation fixed, we can then prespecify the best sampling strategy obtained from the inefficient weighting scheme and compare the results in a fair manner with the best GSDs to evaluate the potential for benefit.

At $\theta = 0.5, \lambda_{\text{Truth}} = \lambda_0/4$, our average accrual size, event size, and calendar time of trial completion is similar. Reducing the proportion of blinded adaptations by 20%, our overall power decreases by approximately 5% (relative to adapting fully using the ad-hoc rule) regardless of when the added accrual is conducted.

Table 5.4 shows the summary of the results based on finding the best adaptive rule when either continuing accrual or restarting accrual. We interpret the results for restarting accrual when $\theta = 0.5$. Using the fully blinded *GSDMod*, the best achievable power is 80.27%, which is similar to the optimal setting when we start off planning the trial with 3500 patients in mind and restarting accrual. This optimal strategy can be thought to be starting a second site when funding opportunities or approval is later provided to conduct the study in that setting. When we choose to decrease the probability of adaption to 80%, the power for *GSDMod* decreases to 78.27% as a consequence of 20% fewer adaptations.

Using the best ("sub-optimal") rule that adapts 80% of the time and having pre-specified these rules in advance, the overall power of the best adaptive design is nearly efficient to the

Table 5.4: Summary of overall power for various designs when making adaptations at 80% of the time. The results under Cond refers to the strategy of applying CRP while Pres corresponds to using minimal sufficient statistics at the trial termination. The adaptive designs (*Adapt*) make an adaptation either based on hazard ratio (HR) or rate difference (Rate Diff) at 80% of the time.

	HR=0.5; $\lambda_0/4$				HR=θ_A; $\lambda_0/2$			
	Continue		Restart		Continue		Restart	
	Pres	Cond	Pres	Cond	Pres	Cond	Pres	Cond
FSD078 (N=1750)	68.69	-	68.69	-	67.55	-	67.55	-
Ref	90.08	-	80.27	-	88.40	-	79.47	-
GSDMod (100%)†	90.08	89.72	80.27	76.88	87.61	87.60	79.47	79.51
GSDMod (80%)‡	86.33	85.74	78.27	73.91	84.63	84.59	77.55	77.36
Adapt: Rate Diff (80%)	88.09	86.52	80.27	75.25	86.21	85.69	79.31	78.84
Adapt: HR (80%)	87.55	86.31	80.10	75.07	86.10	85.58	79.35	78.77

FSD078: GSD design based on 1750 subjects, and terminating at 78 months
Ref: Optimal strategy (*GSD3500A*) when this accrual size of 3500 is always enrolled depending on the accrual patterns.
†: GSD with "escape clause" and blinded adaptations at 100% of the time.
‡: GSD with "escape clause" and blinded adaptations at 80% of the time.

the reference design (Ref or *GSD3500A*). This best adaptive design also beats the design *GSDMod* that adapts 80% of the time. There are slight, negligible differences in the overall power depending on whether the rate difference or estimated hazard ratio is used. However, when these adaptive rules are not prespecified in advance and requires further statistical adjustments, we lose a substantial amount of power. A much later adaptation results in up to 5% loss of power if this rule was not prespecified prior to the conduct of the trial.

Similar results are observed when we choose the best adaptive rule under θ_A. When we continue accrual, i.e., an adaptation is made very early during the study, this loss of power is considerably lesser and is consistent with the results observed in section 3.2.

5.4.1.2 Application of the Best Rule Based on $\lambda_0/4$ and $\theta_A = 0.5$ to Other Settings

It is harder to compare how the adaptive rule can compare with each other when applied to other values of θ since we are no longer holding other operating characteristics similar. This

is as seen in Table D.10 in Appendix D. Because the probability of adaptation is different, this is generally harder to interpret. It is often the case, however, that a higher probability of adaptation increases the overall power of the design. When the adaptive rule is selected based on late adaptations, at other values of θ, there is basically little gain in power as compared to *GSDMod*. The adaptive rule that is selected based on early adaptations generally results in more adaptations across other values of θ. However, such adaptations in accrual also increases the average patient size to be recruited into the study at the cost of little gain in overall power.

5.5 Discussion

In vaccine efficacy trials, sponsors may be concerned about the possibility of non-proportional hazards with a beneficial vaccine efficacy that may wane over time. A beneficial vaccine treatment may potentially wane over time, thus requiring booster shots to maintain immunity. Using a fixed sample design, it is entirely possible that one may detect an average hazard ratio greater than the hypothesized effect and thus miss an opportunity to find an effective vaccine. With multiple interim monitoring, one can detect this non proportionality and the possibility of a vaccine that wanes over time.

Suppose again for illustration, our vaccine treatment may be efficacious over the first 4 years with a low hazard ratio close to 0.1. However, after 5 years, the vaccine becomes less effective in conferring protection. If the vaccine treatment has an average hazard ratio $\theta = 0.3$, interim monitoring can enable one to detect the earliest effect at roughly 42 - 54 months with high probability above 0.999. Since this stopping probability is almost certain at this interim analysis, with a monitoring rule in place, the DSMB can act to "extend" the trial without unblinding the results until 78 months if there are speculations that the treatment may modify how public health practice or lead to future studies on how booster shots may be implemented.

For discussion purposes, we consider additional scenarios whereby the first three scenarios present high protection up to 78 months but differing in protection past 78 months. Survival

curve 1 corresponds to the use of placebo in the public health setting. In scenario 2, we postulate a piecewise constant hazard where the vaccine may no longer be effective by 78 months, i.e., a booster shot may be required to maintain the levels of protective antibody. Such survival curve has a hazard ratio of 0.1, 0.75, 1.6, and 1 over the first 22.5, next 27 months, next 30 months, and past 79.5 months with respect to the placebo. The last scenario postulates the setting where the vaccine/prevention effects wear off totally after 104 months with signs of increasing hazard after 66 months. A piecewise constant hazard of such setting can have a hazard ratio of 0.1, 0.75, 1.6, and 1 over the first 30, next 36 months, next 40 months, and past 104 months respectively relative to the placebo survival curve. This reflects the possibility that the levels of protective antibody are no longer sufficient to confer protection or the potential changes in behaviors that increase the risks of infection.

While the survival curves are stochastically ordered, the use of the log rank test may not be most efficient under such scenarios to pick the better treatment. Instead, other weighted versions of the logrank test statistics may be preferred to gain efficiency when the above treatment scenarios are highly plausible. For example, the use of $G^{1,0}$ or the Peto-Peto Prentice Wilcoxon statistics may be used to emphasize early differences in vaccine efficacy that may wane over time. Even then, it is conventional to evaluate the operating characteristics of a trial under the strong null of proportional hazards. In situations whereby waning treatment effect is likely, then it is useful to consider evaluating such alternatives which will be discussed in later chapters.

5.6 Conclusions

We investigate the use of GSD to adaptively revise the study design based on the pre-specified use of "escape clause" in situations when misspecification of event rates or presence of extreme treatment effect is plausible. While such a strategy results in reasonable loss of power under the hypothesized alternative, there is high power in presence of extreme treatment effect under various combinations of lower event rates than anticipated. When the study design is not limited to constraints of calendar time or patient size, extension of

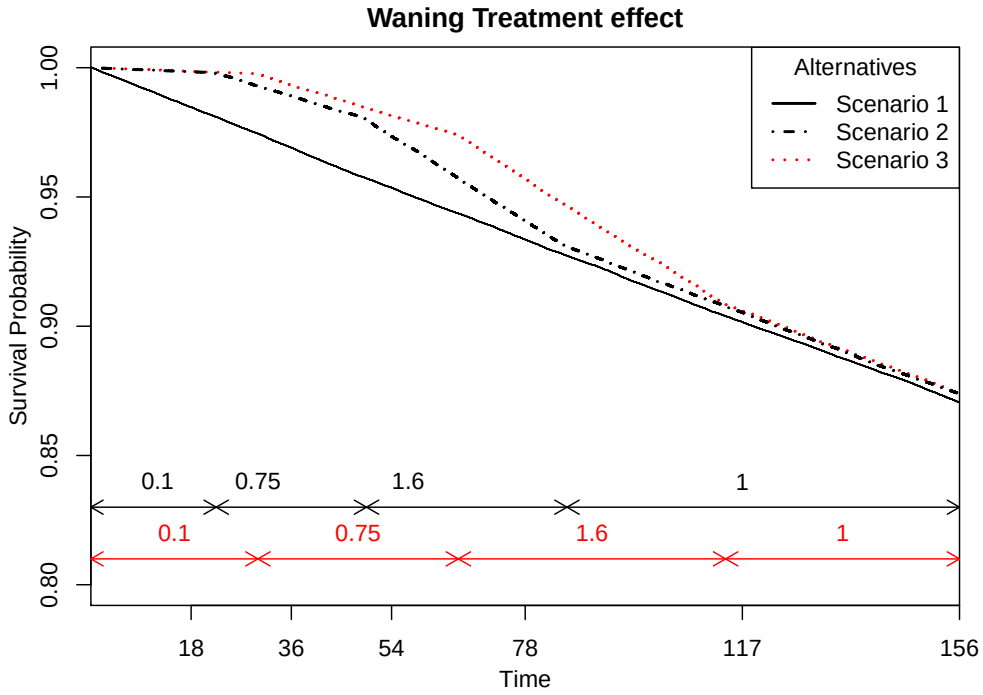

Figure 5.10: Different non proportional hazards scenarios with time varying treatment effect. Sequential analysis with the use of the logrank statistic estimates a different functional since interim analyses affect the risk sets as used by the logrank statistic. Consequently, a different scientific question is addressed at each interim analysis. The logrank statistic is less efficient under this stochastically ordered, non proportional hazards setting. Despite that, the logrank statistic has sufficient power to detect the difference in hazards at early interim analyses.

the study with blinded revision of sample size to either continue or restart accrual generally improved the overall power under our hypothesized alternative.

When this calendar time is of interest and the patient size can be feasibly enlarged, fully blinded strategies through the use of GSD is more than sufficient. In situations when the treatment benefit may not be overwhelmingly optimistic, simple adaptive rules may not provide clear benefit in terms of overall power but a slight increase in patient size and longer study period. The fully blinded GSD strategy has the additional advantage of a clear interpretation at the end of the trial.

In limited scenarios corresponding to moderate efficacy and markedly lower event rate, we see potential benefit with the use of adaptive designs. However, at either extreme efficacy or lower than anticipated event rate, GSDs are shown to be nearly efficient and generally protect us without the complications such as potential operational bias induced as a consequence of the use of unblinded adaptations. Our results also indicate that if one cannot justify the potential operational bias and require further statistical adjustments through the use of weighted statistics, there is a general loss of overall power. This points out the difficulty as well as the added complications when planning an adaptive design since more thought has to be judiciously placed in the selection of not only the monitoring rules but also sensible, fully specified, adaptive rules in order to avoid having to pay the price with the use of weighted statistics.

Our exploration of the adaptive strategies is limited to several monitoring boundaries. We did not find the best adaptive strategy using the optimal GSD that is efficient at the interim analysis to better evaluate the utility of adaptive strategies. The practicality of selecting an optimal GSD may not be best in addressing other constraints such as minimal increase in maximum statistical information or scientific, logistical, or ethical concerns. Here, we want to demonstrate that rather than making flexible adaptations, one should carefully evaluate all potential competing strategies and evaluate the potential for misspecification of design assumptions so as to better understand the robustness of their chosen design. The general principle that we have chosen in evaluating both GSDs and adaptive design as described in this chapter can be applied to general settings when planning a clinical trial.

In the next chapter, we consider the use of weighted log rank statistics to help us gain power under the plausibility of waning treatment effect as often hypothesized in many clinical settings other than vaccine trials. We consider the implications of naïvely presuming the number of events as a general rule to quantifying statistical information in the time to event setting and the perils when considering unblinded adaptations with the use of such "less well-understood" survival analyses methods.

Chapter 6

Information Growth for the Weighted Logrank Statistics

Common approaches to analyzing time to event data consider the use of the logrank statistic or Cox PH regression to compare treatment groups. The logrank statistic or Cox PH regression is semi-parametric efficient under the strong null hypothesis when the hazards for the comparison groups are proportional and similar over time or under proportional hazards alternatives. However, when there is a priori evidence of time varying treatment effect as characterized by the difference in hazard function, weighted forms of the logrank statistics ($G^{\rho,\gamma}$ family) may be used to gain power under these hypothesized alternatives. Several Phase III confirmatory/prevention trials such as Women's Health Initiative [The Women's Health Initiative Study Group], XINLAY trial [Carducci et al., 2007], or National Lung Screening Trial [Aberle et al., 2011] have employed the use of weighted logrank statistics to analyze the primary endpoint of interest. For example, the National Lung Screening Trial [Team et al., 2011] was conducted to determine whether the use of spiral CT as compared to chest X-rays was effective as a screening tool in the prevention of mortality from lung cancer. In this trial, a weighted version of the log-rank statistics was used to specifically down-weight earlier lung cancer deaths to account for the delayed effects of the intervention strategy.

Typically, in an event driven time to event setting, interim analyses are performed when some number of events have been accumulated. While the "sample size" or statistical information is directly proportional to the number of events when analyzing censored data

using logrank statistic under the strong null, this may no longer hold true when using the weighted forms of logrank statistics. Information growth in the weighted versions of the logrank statistics ($G^{\rho,\gamma}$) has been shown to be a function the censoring (that includes the patient accrual distribution) and the underlying survival distribution [Gillen and Emerson, 2005]. In the sequential setting, Gillen and Emerson [2005] characterized the nonlinear behavior of the true information growth vs the proportionate events (the fraction of the number of events accumulated relative to the pre-defined maximum number of events). Using these "less well-understood" weighted versions of the logrank statistics, the presumption of a linear trend between information growth and proportionate events no longer holds true.

In this chapter, we explore some of these issues that may arise in the use of adaptive designs while making the naïve presumption of using the number of events accumulated at interim analysis as a direct surrogate to measure statistical information based on less commonly used analyses techniques in survival. The $G^{\rho,\gamma}$ family of weighted logrank statistics is such an example as described in section 4.6 whereby a different weighting scheme is chosen to gain efficiency under non proportional hazards alternatives to place emphasis on either early or late differences in survival. To demonstrate that this naïve presumption is no longer true when using these "less well-understood" survival analyses methods, we evaluate this assumption by assessing the degree of overall Type 1 error control when we modify the censoring distribution, as measured by the accrual patterns, that may commonly arise in many clinical settings.

We investigate this under the setting where this modification is first made based on fully blinded adaptations as described in section 6.2, and then expand this to the adaptive setting whereby modification is made based on unblinded interim results (section 6.3). Then, we describe how an adaptive modification to the censoring distribution when using these weighted statistics have inherently modified the information growth in section 6.4. We investigate the use of Cui et al. [1999] to control for this inflation of overall Type 1 error and the degree to which we must accurately characterize information growth when applying these "less well-understood" adaptive procedures with these "less well-understood" survival

analyses methods in section 6.5. We discuss the potential for benefit as well as the risks of applying these adaptive procedures in the clinical trial setting when dealing with censored data.

6.1 Sequential Analysis with $G^{\rho,\gamma}$ Family

We briefly review some of the notation in section 4.6 necessary for this chapter.

6.1.1 Notation

Consider a GSD with continuation sets $\mathcal{C}_j \equiv \{(a_j, b_j] \cup [c_j, d_j)\}$ such that $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$ with $j = 1, \cdots J$ analyses [Kittelson and Emerson, 1999]. Let t_j be the information time at which the interim analysis is conducted where we define $t_J = 1$ at the final analysis. At the j^{th} analysis, we compute the normalized score statistic $Z_j = U_j/\sqrt{V(t_j)}$ where U_j is the (cumulative) score statistic, and $V(t_j) = \text{Var}[U_j]$ is the variance of the score statistics, or Fisher's information at the j^{th} analysis.

Define our proportion of information at analysis j to be of the form $\Pi_j = V(t_j)/V(t_J)$ where Π_j is the fraction of total statistical information available from all patients at the time of interim analysis relative to the maximum planned statistical information at the end of the trial. Under the null hypothesis, our test statistic U_j is approximately normal with mean 0 and variance $V(t_j)$. With efficient estimators, Scharfstein et al. [1997] and Jennison and Turnbull [1997] showed that U_j has the independent increments covariance structure such that $\text{Cov}(U_{j+1}, U_j) = V(t_j)$ for $j = 1, \cdots, J$. Under the design alternative, U_j can be approximated using the normal distribution with mean $\theta V(t_j)$ and variance $V(t_j)$ where θ is the parameter of interest. In this section, we focus on the two stage design with no interim stopping for efficacy or futility similar to the intent to cheat example in Proschan and Hunsberger [1995].

6.1.2 Weighted Logrank Statistics/$G^{\rho,\gamma}$

Consider the comparison of the survival experience between two treatment groups, $k = 0, 1$ in a GSD where a total of J analyses are conducted at the pre-defined calendar times $\tau_1, \cdots, \tau_J$. Our hypothesis of interest can thus be written as $\mathbb{H}_0 : S_1(t) = S_0(t)\ \forall\ t$ vs $\mathbb{H}_A : S_1(t) \neq S_0(t)$ for some t. Under the strong null hypothesis (and thus proportional hazards alternative), we presume exact equality of survival and our null hypotheses are in the form $\mathbb{H}_0 : S_1(t) = S_0(t)^\theta \equiv \theta = 1\ \forall\ t$, vs $\mathbb{H}_A : S_1(t) \neq S_0(t) \equiv \theta \neq 1$ for some t. For the remainder of this chapter, our hypothesis of interest is set up such that we are interested in the superiority of treatment (1) vs the placebo (0). Thus, $\mathbb{H}_0 : \theta \leq 1$ vs $\mathbb{H}_A^{\text{PH}} : \theta \geq \theta_A$ where θ_A is our design proportional hazards alternative. We parameterize $\log(\theta_A) = \log(h_0(t)/h_1(t))$ such that large values of $\log(\theta_A)$ indicate superiority of the treatment.

Let E, T, C be the random variables corresponding to the calendar time of entry into the study, the study time for the event of interest, and the study time for loss-to-follow-up with the respective distribution functions H, F, and G. At the j^{th} interim analysis that is conducted at some calendar time τ_j, where a total of N_j subjects have been accrued, the i^{th} subject has data of the form $(X_{i,j}, \Delta_{i,j}, Z_i)$ where $X_{i,j} = \max(\min(T_{i,j}, C_{i,j}, \tau_j - E_{i,j}), 0)$ is the observed time for individual i, $\Delta(X_{i,j})$ is the indicator variable for an observed failure time if $X_{i,j} \leq \min(C_{i,j}, \tau - E_{i,j})$ and 0 if $E_{i,j} > \tau$ and that loss of followup is only due to administrative censoring, and the randomized treatment assignment is

$$Z_i = \begin{cases} 0 & \text{if the } i^{\text{th}} \text{ individual belongs to treatment group 0} \\ 1 & \text{if the } i^{\text{th}} \text{ individual belongs to treatment group 1.} \end{cases}$$

For notational simplicity later, we further let $\Delta_1(X_{i,j}) = \Delta(X_{i,j})Z_i$ and $\Delta_0(X_{i,j}) = \Delta(X_{i,j})(1 - Z_i)$ where they are the indicators of an observed failure for the i^{th} subject coming from group 1 and 0 respectively at the j^{th} interim analysis. Recall from section 4.3, our (partial) score statistic based on the data collected at the j^{th} interim analysis that is

conducted at some calendar time τ_j can be expressed as

$$\begin{aligned}\mathcal{U}(\beta^j)|_{\beta^{(j)}=0} &= \sum_{i=1}^{N_j} \Delta_{i,j} \left[Z_i - \frac{\sum_{l \in \mathcal{R}_j(X_{i,j})} Z_l}{\sum_{l \in \mathcal{R}_j(X_{i,j})}} \right] \\ &= \sum_{i=1}^{N_j} \frac{n_{0,j}(X_{i,j}) n_{1,j}(X_{i,j})}{n_{0,j}(X_{i,j}) + n_{1,j}(X_{i,j})} \left[\frac{\Delta_1(X_{i,j})}{n_{1,j}(X_{i,j})} - \frac{\Delta_0(X_{i,j})}{n_{0,j}(X_{i,j})} \right]\end{aligned}$$

where $\mathcal{R}_j(X_{i,j}) = \{l : X_{l,j} \geq X_{i,j}\}$ is the risk set at analyses time $X_{i,j}$ that is based on the data collected at the j^{th} interim analysis, N_j is the total number initially at risk at the j^{th} interim analysis, $N_{k,j}$ is the number initially at risk for group k at the j^{th} interim analysis, $N_j = N_{0,j} + N_{1,j}$, and $n_{k,j}(X_{i,j})$ is the number at risk at analysis time $X_{i,j}$ for the k^{th} treatment group for $k = 0, 1$.

Let $w(t) = \frac{n_{0,j}(t) n_{1,j}(t)}{n_{0,j}(t) + n_{1,j}(t)} [\widehat{S}(t^-)]^\rho [1 - \widehat{S}(t^-)]^\gamma$ where $\widehat{S}(t^-)$ is the pooled Kaplan Meier survival estimate. Fleming and Harrington [1991] introduced this flexible weight function within the log rank test statistic to allow comparison of a bigger class of survival curves. By this setup, at some time τ, the general form of the $G^{\rho,\gamma}$ statistics is

$$G^{\rho,\gamma} = \sqrt{K} \sum_{i=1}^{N_j} w(X_{i,j}) \left[\frac{\Delta_1(X_{i,j})}{n_{1,j}(X_{i,j})} - \frac{\Delta_0(X_{i,j})}{n_{0,j}(X_{i,j})} \right].$$

where $K_j = \frac{N_{0,j} + N_{1,j}}{N_{0,j} N_{1,j}}$.

By this general representation, when ρ, γ are both 0, we obtain the Cox regression/log rank test for comparison of the survival experience between two groups. Under the strong null hypothesis, $H_0 : S_0(t) = S_1(t)\ \forall t > 0$, a consistent estimator of the variance of the $G^{\rho,\gamma}$ statistic can be expressed as follows:

$$\widehat{\sigma}^2 = K \sum_{i=1}^{N_j} w^2(X_{i,j}) \left[\frac{1}{n_{0,j}(X_{i,j})} + \frac{1}{n_{1,j}(X_{i,j})} \right] \left[1 - \frac{\Delta_1(X_{i,j}) + \Delta_0(X_{i,j})}{n_{0,j}(X_{i,j}) + n_{1,j}(X_{i,j})} \right] \frac{\Delta_1(X_{i,j}) + \Delta_0(X_{i,j})}{n_{0,j}(X_{i,j}) + n_{1,j}(X_{i,j})}$$

We can further re-write the $G^{\rho,\gamma}$ statistic in the sequential testing framework as

$$W_j^{\rho,\gamma} = \frac{1}{\sqrt{K_j}} G^{\rho,\gamma} \equiv U^*_{\tau_j}$$

such that $W^{\rho,\gamma}$ can be represented as some form of weighted score statistics $U^*_{\tau_j}$.

Define our proportion of information at analysis j to be of the form $\Pi_j = V(\tau_j)/V(\tau_J)$ where Π_j is the fraction of the statistical information available from all patients at the time of interim analysis τ_j relative to the maximum statistical information at the end of the trial conducted at τ_J. This gives us the information fraction at the j^{th} interim analysis as

$$\Pi_j = \left[\left(\frac{N_{0,j}N_{1,j}}{N_{0,j}+N_{1,j}}\right)\widehat{\sigma}_j^2\right] \Big/ \left[\left(\frac{N_{0,J}N_{1,J}}{N_{0,J}+N_{1,J}}\right)\widehat{\sigma}_J^2\right]$$

6.1.3 Censoring Distribution

The censoring distribution, as characterized by the accrual distribution of patients, plays an integral role to estimating the information growth of the test statistic. We consider a flexible parametric accrual distribution used in Gillen and Emerson [2005] to simulate the accrual patterns. Let E be the random variable describing the accrual distribution. The cumulative accrual distribution is of the form,

$$F_E(t) = \left(\frac{t}{A}\right)^q \quad , \quad A > 0, q > 0 \text{ and } 0 < t \leq A \tag{6.1}$$

where A controls the period of accrual. Note that the random variable $E^{'}$ that has cumulative distribution $F_{E'}(t) = (1 - t/A)^q$ generates another potential family of accrual distribution.

When A is non-zero, patients are accrued over the time period A. The parameter q controls the rate of patient accrual into the clinical trial. When $q = 1$ and $A > 0$, patients are accrued uniformly over the time period between 0 and A. As $q \to \infty$, patients enter slowly at the beginning of the trial and more rapidly when we are closer to time A. As $q \to 0$, patients are accrued in rapidly at the beginning of the trial.

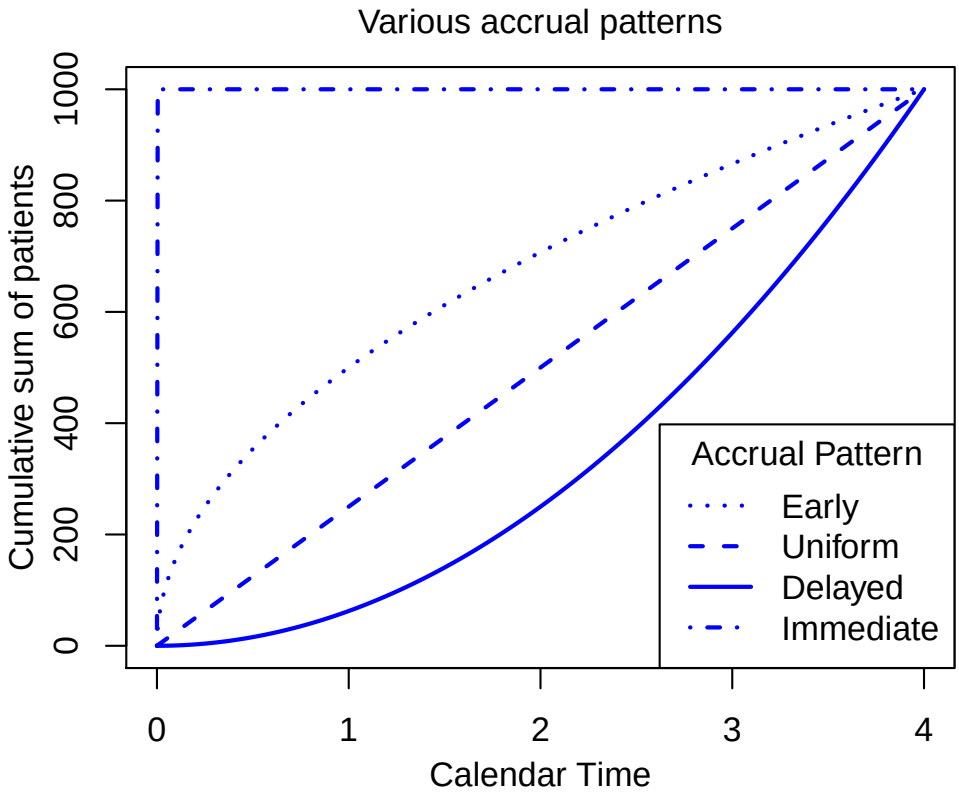

Figure 6.1: Varying plausible patterns of accrual in a clinical trial setting based on the parametric formulation of equation 6.1. This formulation gives us a variety of accrual patterns that can be used to investigate the impact of censoring in a clinical setting.

For example, suppose we consider accruing 1,000 subjects into a clinical trial over a period of four years, with $A = 4$ as according to Equation 6.1. Several plausible accrual scenarios based on Equation 6.1 can be obtained. For example, when $q = 1$, we have uniform accrual of subjects as shown via the dotted bold line in Figure 6.1. $q = 2$ represents a particular instance of delayed accrual as shown via the solid blue line. $q = 0.5$ describes a particular form of early accrual as shown by the thin dotted line. If we let $q = 0$, and for any arbitrary positive, non-zero A, this mimics the immediate accrual setting shown by the dash-dotted line where the entire patient size is recruited at calendar time 0. The above parametric family of accrual distributions allow us to simulate various accrual patterns for our exploration later.

6.2 Blinded Accrual Size Adaptations using $G^{\rho,\gamma}$ Statistics

In Chapter 5, we explored the use of blinded sample size re-estimation to decrease accrual using the logrank statistic. We previously consider the strategy of allowing both extension of accrual size as well as calendar time, and/or keeping calendar time fixed as an "escape clause". We investigate the impact of increasing accrual on the control of overall Type 1 error when using the $G^{\rho,\gamma}$ statistics while holding the total number of events fixed. The $G^{0,0}$ test corresponds to the unweighted logrank statistic. We consider four specific forms of the weighted statistics for the remainder of this chapter to illustrate the extremes for these combinations of ρ and γ parameters. Namely, the logrank statistic ($G^{0,0}$), $G^{1,0}$ (commonly referred to as Peto-Peto and Prentice Wilcoxon statistic), $G^{0,1}$, and $G^{1,1}$ are of interest. We explored the impact of blinded adaptations on the control of the overall Type 1 error under the strong null setting.

6.2.1 Simulation Setup for Blinded Adaptations

To investigate the above, we suppose a FSD is planned with a (minimum) sample size commitment of 1,000 subjects, and that the final analysis is conducted when a total of 765 events are attained. In this FSD, we assume an administrative look is made part way through the study to modify accrual, with the objective of increasing the overall event rates based on blinded data. We interpreted this as some version of "start small, ask for more later" concept whereby trials may commit a minimum number of subjects to be recruited at the beginning of the trial. When there is more funding, or logistically more feasible to open up new sites, sponsors may increase accrual part way through the study without increasing the total number of "events".

We conducted 100,000 simulations under the strong null setting where the survival distributions for $S_0(t)$ and $S_1(t)$ are drawn from the Weibull distribution with shape parameter of 0.5, and "rate" parameter of 120.1 such that $S(6) = 0.8$ (with the definition of the Weibull distribution consistent with the parametrization in Shorack [2010]). At level $\alpha = 0.05$, the

unweighted logrank statistic has 93.1% power under no censoring (other than administrative censoring) to detect a proportional hazards design alternative, $\theta_A \approx 1.2538$. Based on θ_A, the $G^{1,0}$, $G^{0,1}$, and $G^{1,1}$ statistic has 90%, 90%, and 84.8% power respectively to detect this design alternative. Thus, under θ_A, a total of 1,000 subjects are thought to be required to obtain approximately 765 events at the end of the study. Note that we did not choose specific alternative distributions that are most efficient for these weighted logrank statistics due to the difficulty of characterizing the appropriate survival functionals particularly for the weighting schemes based on $G^{0,1}$ and $G^{1,1}$.

Various accrual patterns are considered. Based on Equation 6.1, we set our accrual period A to be 4. We investigated the setting of uniform accrual ($q = 1$), delayed accrual ($q = 2$ based on $F_{E'}$ rather than F_E), and early accrual ($q = 1/2$). The immediate accrual setting is considered to provide a limiting case of what would happen in absence of censoring. We assume further that there is no loss to follow up and administrative censoring to take place at the time of analysis.

An interim analysis is conducted at either 1/3, 1/2, or 2/3 (255, 382, and 510 respectively) of the total number of events to increase the accrual size. At this interim analysis, we do not allow early stopping for futility or efficacy. Additionally, we presume this blinded adaptation is made without any other knowledge of secondary results. We consider expanding our total accrual size from 1,000 to either 1500, 2000, 3000, and 5000. Following this interim analysis, all remaining subjects who have not been enrolled (based on the 1,000), together with the additional subjects to be accrued, are then enrolled uniformly at double the original accrual rate, i.e., at 500 subjects per year.

We apply the above blinded procedure consistently at this interim analysis and continue follow-up of all subjects until a total of 765 events are accumulated. We analyzed the trial by computing the Z statistic for each simulation and each analysis method ($G^{0,0}, G^{1,0}, G^{0,1}$, and $G^{1,1}$). These Z statistics are parameterized such that large values denote the superiority of treatment (1) over placebo (0), small values are consistent with superiority of placebo over the treatment. At level α, we compute the total number of significant trials where

$Z > \Phi^{-1}(1-\alpha)$.

Realistically speaking, these forms of accrual adaptations in accrual size may not be optimal for the various $G^{\rho,\gamma}$ statistics where we do not prolong the total study duration and/or increase the total number of events. For example, the weighting scheme of the $G^{1,0}$ statistic places emphasis on early differences in survival. Thus, increasing accrual part way through the study only improves upon the power of the test statistic when these early differences are of scientific interest. In the context of $G^{0,1}$ or $G^{1,1}$, increasing accrual rapidly while not allowing for additional follow-up can potentially not improve the overall power since the weighting scheme places emphasis on mid to late differences in survival. In other words, any early events based on the later accrual is given less emphasis by the weighting scheme. Nonetheless, our objective in this chapter is basically to demonstrate what are the consequences of making adaptations in terms of accrual size while maintaining the maximal number of events. That is, we attempt to demonstrate how incorrect assumptions on the part of clinical trialists (i.e., that information growth is proportional to the number of events) might be particularly deleterious with adaptive clinical trials..

6.2.2 Results for Blinded Accrual Size Adaptations

Results for $\alpha = 5\%$ and 2.5% are similar. We refer the interested reader to Appendix E.1 for additional results when blinded adaptations are conducted at other interim analyses (Table E.1, E.2, E.3, E.4, and E.5). In summary, the overall Type 1 error rate is protected when a blinded increase in sample size is made at the interim analyses defined above. The results are somewhat intuitive. Since we are only conditioning on our knowledge of the presumed "statistical information" $d(t)$, which is the number of events, this is ancillary of the estimated treatment effect. Thus, under the strong null hypothesis, blinded adaptations of the accrual size should not inflate our overall Type 1 error.

Under the PH alternative, we see that the overall power of the unweighted logrank statistic remains unaffected (compared to the power based on a accrual size of 1000). In summary, any form of accrual size increase will improve the overall power for the $G^{1,0}$ statistic slightly

(2% relative to 90%). For $G^{0,1}$ and $G^{1,1}$ statistics, any form of blinded adaptation to increase accrual size results in considerable loss of power. For the $G^{0,1}$ statistic, this can decreasing the overall power from 90% to 80%. For the $G^{1,1}$ statistic, this can result in the overall power to decrease from 85% to 80%. In particular, the overall power decreases (for $G^{0,1}$ and $G^{1,1}$) as more subjects are added into the trial following interim analysis without either extending follow-up and/or the total number of events.

Despite having characterized and evaluated the overall Type 1 error under the strong null hypothesis which is coincidentally proportional hazards, we did not evaluate the above strategy based on the alternatives that are most efficient for these weighted statistics. This is generally difficult especially for the $G^{0,1}$ or $G^{1,1}$ statistic where the alternative functionals are harder to parametrize. An alternative evaluation for the $G^{0,1}$ or $G^{1,1}$ statistic is to decrease the rate of accrual by half so that the remaining events take a longer time to accumulate. However, such an approach no longer provides benefit due to economic and financial concerns with a larger trial and prolonged follow-up. Although the PH alternatives may not be most efficient for $G^{1,0}$ since it emphasizes early differences, blinded increase in accrual can maintain/improve the overall power of the $G^{1,0}$ statistic since this generally allows better characterization of the survival curves earlier on. Since we did not consider extending the follow-up time, increase in accrual should improve only the precision of the estimate of the early survival curve.

6.3 Intent To Cheat Sensitivity Analysis

We investigate the impact of sample size adaptation based on the use of unblinded interim results. We term this as the intent-to-cheat sensitivity analysis since at interim, the future accrual of the subjects is now dependent on the estimated treatment effect we observed at interim. Our objective is to determine the degree of inflation of the overall Type 1 error when *holding our total number of events fixed*.

Thus, at some interim analyses, we condition on the unblinded treatment effect to determine whether we should increase accrual. Since we are no longer conditioning on an ancillary

statistic, but rather our sufficient statistic that is a function of our estimated treatment effect, this sensitivity procedure is similar to the worse case inflation described by Proschan and Hunsberger [1995] with the caveat that we now presume holding constant our "sample size" as defined using the total number of events.

Consider the clinical trial scenario in the context of a DMC meeting involving the statistician in the DMC as well as the independent statistical center presenting and analyzing the results for the DMC. When results are first presented in a blinded fashion, the DMC may potentially recognize that the event rate may be lower than anticipated and if not presented any form of unblinded safety/adverse events data, recommendations can be provided to the sponsor or investigators to decide whether additional accrual is required. However, once the DMC are unblinded to the interim data, the DMC are possibly biased to making such suggestions unless they are pre-specified in the protocol. Additionally, the DMC or the independent statistical center are also unable to make recommendations to the sponsors or investigators to increase accrual unless the sponsors have appropriately stipulated the rule for which an increase in accrual is deemed reasonable subject to pre-specified definition of low event rate in the protocol prior to the start of the trial, or possibly before unblinding of any (interim) trial results.

6.3.1 Unblinded Accrual Size Adaptations

Using the same simulation set up as in section 6.2.1, we choose to now use the unblinded interim estimate of the treatment effect to adapt our future accrual of subjects into the study. Let $\widehat{Z}_{\text{Interim}} \in \Re$ be our Z statistic computed based on the data accrued at the interim analysis when a total of d_{Interim} events are obtained. Following the approach of Chen et al. [2004], Gao et al. [2008] and subsequently Mehta and Pocock [2011], we partition this sample space of plausible Z's at interim analysis into the three zones: "Unfavorable", "Promising", and "Favorable". Let Z_{Lower} be the value that divides "Unfavorable" and "Promising" zones, and Z_{Upper} be the value that divides the "Promising" and "Favorable" zones. The three zones can be described as follows:

1. *Unfavorable*: When the $\widehat{Z}_{\text{Interim}} < Z_{\text{Lower}}$, we modify accrual from N to N_1. Since the interim results may be disappointing, an investigator may choose to modify accrual. Mehta and Tsiatis [2001] describe such a region to have sufficiently low conditional power that is generally not worth continuing the trial further.

2. *Promising*: When the $\widehat{Z}_{\text{Interim}} \in [Z_{\text{Lower}}, Z_{\text{Upper}}]$, we modify accrual from N to N_2. This region is characterized as the promising zone in the adaptive literature since the results computed based on conditional power are "promising" enough to warrant modification of the sample size to repower the study to a treatment effect θ_A^* where $\theta_A^* \in (1, \theta_A)$.

3. *Favorable*: When the $\widehat{Z}_{\text{Interim}} > Z_{\text{Upper}}$, we modify accrual from N to N_3. In this region, the results have high conditional power of being statistically significant at some future analyses. Often, increase in accrual in these regions are not truly warranted.

The summarized algorithm is shown in Figure 6.2.

Based on a particular fixed value of Z_{Lower} and Z_{Upper}, we can determine the future sample path for all the simulations when the interim result falls into one of the above zones. We can then compute the significance of this trial based on the final statistic obtained after such an adaptation. This is then computed for all the simulations to obtain the overall Type 1 error associated with an adaptation based on this fixed value of Z_{Lower} and Z_{Upper}. In order to determine the maximum Type 1 error, we have to evaluate all potential placement of Z_{Lower} and Z_{Upper}, with the added condition that $Z_{\text{Lower}} < Z_{\text{Upper}}$, based on a grid search. We describe this grid search procedure next.

Let the interim analysis be conducted at some fraction k of the total number of events that is fixed at 765, for $k = 1/3, 1/2, 2/3$. These values of k correspond to 255, 382, and 510 events respectively. For each of the 100,000 simulations, we compute the interim Z statistics for the analyses method of choice when $765k$ events have been gathered. We allow Z_{Lower} and Z_{Upper} to take possible values of $\Phi(Z)$ such that $p = \Phi(Z) \in (0.05, 1-\alpha)$ subject to the constraint that $Z_{\text{Lower}} < Z_{\text{Upper}} \leq \Phi^{-1}(1-\alpha)$ where $\Phi(.)$ is the cumulative distribution function of the

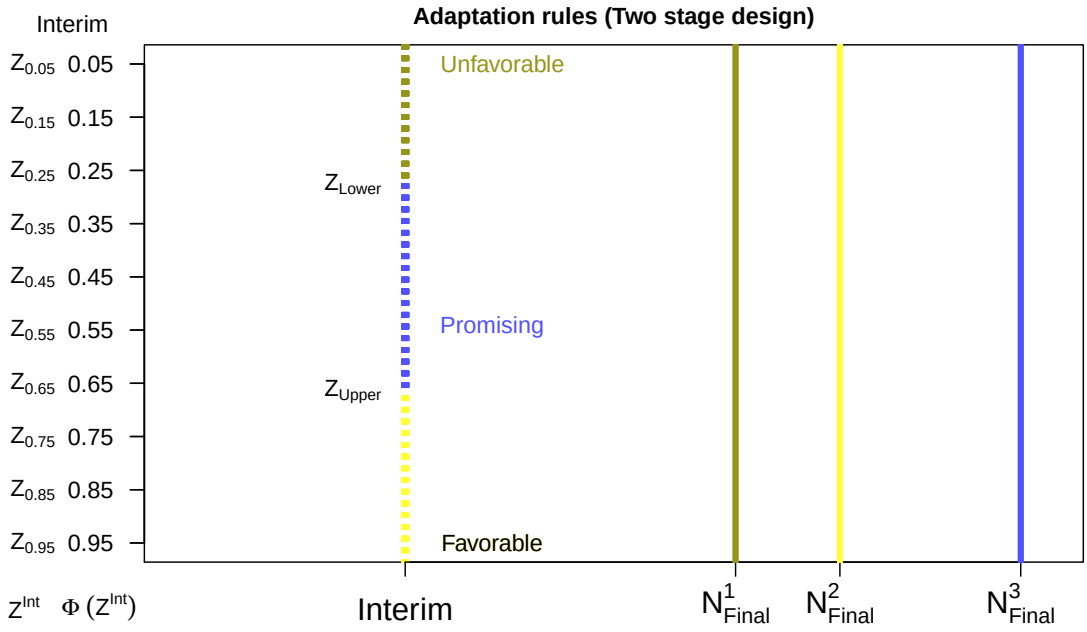

Figure 6.2: The intent to cheat sensitivity analysis is shown with the x-axis to reflect the total number of subjects at the end of the study. The beige, blue, and yellow dotted line describes the continuation regions corresponds to the "Unfavorable" (allowing adaptation in accrual up to N_1^{Final}), "Promising" (allowing adaptation in accrual up to N_2^{Final}), and "Favorable" (allowing adaptation in accrual up to N_3^{Final}) zone, where Z_{Lower} divides the "Unfavorable" and "Promising" zones while Z_{Upper} divides the "Promising" and "fFavorable" zones.

standard normal distribution. For each pair of fixed Z_{Lower} and Z_{Upper}, we then compute the overall Type 1 error rate, i.e., the proportion of the number of statistically significance trials such that $\Phi(Z_{\text{Final}}) > 1 - \alpha$ based on all the simulations. We iterate this through a grid search consisting of all possible pairs of $\{p_{\text{Lower}}, p_{\text{Upper}}\}$ such that $p_{\text{Lower}} < p_{\text{Upper}} \leq 1 - \alpha$ where $p_{\text{Lower}} = \Phi(Z_{\text{Lower}})$ and $p_{\text{Upper}} = \Phi(Z_{\text{Upper}})$. This procedure is similar to the approach taken in Proschan and Hunsberger [1995] in the immediate setting.

One of the most common approach is to increase accrual to some number N_2 only in the promising region, i.e., when $\widehat{Z}_{\text{Interim}} \in [Z_{\text{Lower}}, Z_{\text{Upper}}]$ while letting $N_1 = N = N_3$ for other

zones. Thus, an adaptation is allowed only when the interim results $\widehat{Z}_{\text{Interim}}$ fall within this promising zone. Our final $Z_{\text{Final}}^{N_2}$ is computed based on this new accrual size when we obtain all 765 events. In this particular setting, when $\widehat{Z}_{\text{Interim}} \notin [Z_{\text{Lower}}, Z_{\text{Upper}}]$, we continue the trial with the original accrual size where $N_1 \equiv N_3 = 1000$, and analyze the results when all 765 events are obtained. This is then evaluated for all 100,000 simulation for each pair of fixed $\{p_{\text{Lower}}, p_{\text{Upper}}\}$.

Table 6.1: Summary of the potential accrual size adaptations depending on the whether the interim $\widehat{Z}$ statistic falls within the "Unfavorable", "Promising", or "Favorable" zone. The total number of events remain unchanged while the total number of patients recruited may change depending on the Rule. For example, in #4b, we increase accrual to 2000 when this interim estimate is in either the "Unfavorable" or "Promising" zone while continuing with 1000 subjects when in the "Favorable" zone.

Rule #	"Unfavorable" N_1	"Promising" N_2	"Favorable" N_3
1	2000^a, 3000^b, 5000^c	1500	1000
2	3000^a, 5000^b	2000	1000
3	5000	3000	1000
4	1500^a, 2000^b, 3000^c, 5000^d		1000
5	1000	1500^a, 2000^b, 3000^c, 5000^d	1000

We set $\alpha = 0.05$, and discretized values of p in steps of 0.05, from 0.05 to 0.8 (inclusive), and then in steps of 0.01, from 0.81 to 0.95. This gives us a total of 465 possible combinations of p_{Lower} and p_{Upper}. Each combination of p_{Lower} and p_{Upper} gives us an adaptive rule that enable us to characterize the overall Type 1 error. We then characterize regions in this space that leads to an inflation of the overall Type 1 error. Note that 564 such combinations are possible with this setup when $\alpha = 0.025$.

Table 6.1 presents some of the potential sample size strategy which we explore in the next section. Adaptive rule 1 may reflect the setting of continuing accrual when results may not be favorable but increasing accrual slightly when in the promising region. We consider

settings in rule#5 (a-d) to be most interesting, since they reflect the typical scenarios of increasing accrual when interim treatment results are "potentially promising". We describe our simulation results for the specific setting #5b when an interim analysis is conducted at 255 events, at level $\alpha = 0.05$, and assuming uniform accrual for the rest of this chapter. Other results are in Appendix E.

6.3.2 Simulation Results for Intention To Cheat Sensitivity Analysis

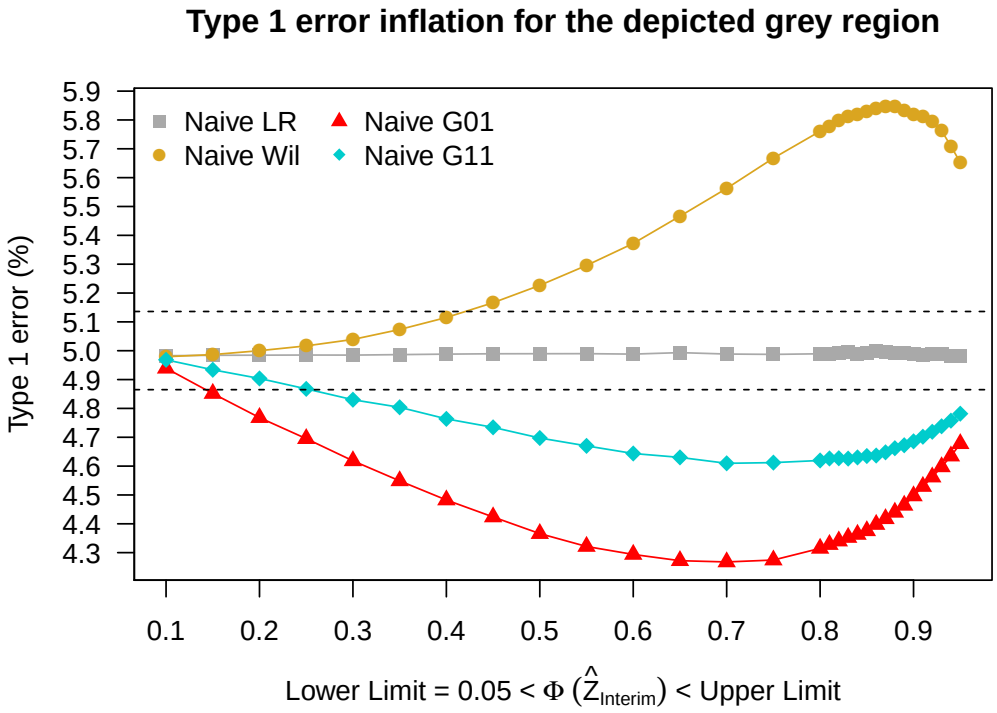

Figure 6.3: A particular slice of an adaptation with the lower limit fixed at $\Phi(Z_{\text{Lower}}) = 0.05$. The upper limit of $\Phi(Z_{\text{Upper}})$ is allowed to vary from 0.1 to 0.95. For any $\widehat{Z}_{\text{Interim}} > 0$, i.e., $\Phi(\widehat{Z}_{\text{Interim}}) > 0.5$, on average, doubling the accrual size following an interim analysis often leads to an inflation of Type 1 error for the $G^{1,0}$ statistic. This same approach leads to a conservative Type 1 error for the $G^{0,1}$ or $G^{1,1}$ statistic for fixed $\Phi(Z_{\text{Lower}}) = 0.05$ and any $\widehat{Z}_{\text{Interim}} > 0$ in this Figure.

Based on 100,000 simulations, at level $\alpha = 5\%$, the maximum Type 1 error obtained for the weighted logrank statistics ($G^{1,0}$, $G^{0,1}$, and $G^{1,1}$) are shown to be inflated with the application of this procedure. The overall Type 1 error for the logrank statistic does not appear to be inflated with the application of this procedure. To determine whether this inflation is only for specific choices of Z_{Lower} and Z_{Upper}, we first look at results when we hold $\Phi(Z_{\text{Lower}}) = 0.05$ (refer to Figure 6.3). The x-axis describes values of $\Phi(Z_{\text{Upper}})$ in this Figure which range from 0.1 to 0.95. The dotted lines correspond to the 95% confidence interval of a typical Type 1 error rate of 0.05 based on 100,000 simulations. Each point in the figure represents the average number of significant trials based on 100,000 simulations when we hold fixed the lower boundary, on the p-value scale, at 0.05 and select some upper limit, say $\Phi(Z_{\text{Upper}}) = 0.5$.

In Figure 6.3, the $G^{1,0}$ statistic has an overall Type 1 error at approximately 5.25% while both the $G^{0,1}$ and $G^{1,1}$ statistic show a conservative control of Type 1 error of approximately 4.7 and 4.4% respectively. Across other values of this Z_{Upper} that we investigated, majority of the regions show an inflation of overall Type 1 error for the $G^{1,0}$ statistic, other regions for the $G^{0,1}$ and $G^{1,1}$ demonstrate high conservatism.

In Figure 6.3, we see that the logrank statistic does not present such a behavior as we vary Z_{Upper}. Generally, we did not see an unacceptable inflation of the overall Type 1 error. The weighted logrank tests do not appear to preserve the overall Type 1 error with most of the adaptations. In fact, the overall Type 1 error is inflated for $G^{1,0}$ statistic. For $G^{0,1}$ and $G^{1,1}$, inflation of the overall Type 1 error occurs only when we make interim changes to the accrual size investigated under Rule 5 in the promising zone.

As we move $\Phi(Z_{\text{Lower}})$ away from 0.05, as shown in Figure 6.4, we see major inflation of overall Type 1 error for the weighted logrank statistics. This inflation of overall Type 1 error is dependent on the choices of Z_{Lower} and Z_{Upper} across the 33 slices explored. Note that the x-axis of Figure 6.4 now represents sets of $\Phi(Z_{\text{Lower}})$ that consist of values ranging from $0.05, 0.1, 0.15, \cdots, 0.95$. Within each vertical pair of consecutive dotted lines, the x-axis represent values of Z_{Upper} that are greater than some fixed value of Z_{Lower} similar to the way

the x-axis of Figure 6.3 is labeled.

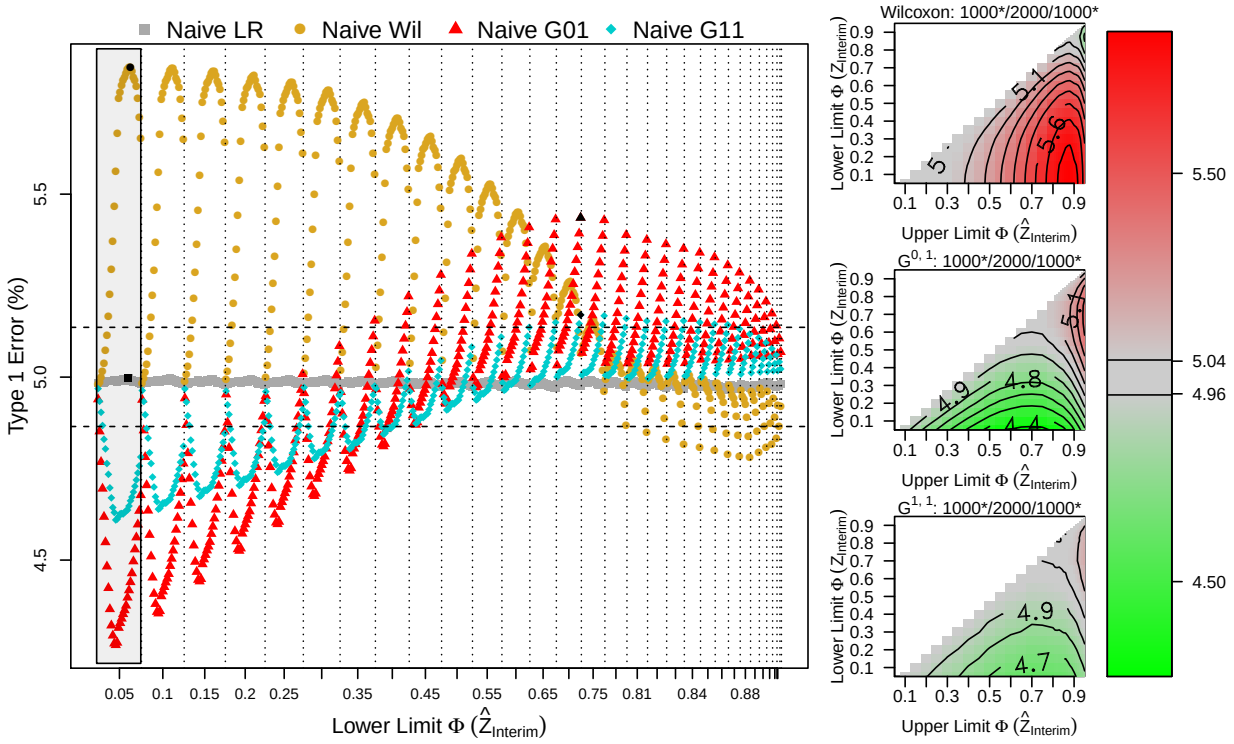

Figure 6.4: Degree of inflation of overall Type 1 error when increasing the accrual size to 2000 in the promising region under uniform accrual with interim analysis conducted at 1/3 of the total event size. Grey highlighted region is a slice as shown in Figure 6.3 with the lower limit fixed at 0.05 and the upper limit set at other values of $p \in (0.05, 0.95]$. In general, each point summarizes the amount of Type 1 error inflation for a combination of p_{Lower} and p_{Upper}, and a particular analysis method of choice. The behavior of the inflation of overall Type 1 error varies across different combinations of p_{Lower} and p_{Upper}.

Alternatively, these regions can be viewed as contour plots for each analysis method presented on the right of Figure 6.4. These contour plots describe the heatmap of the overall Type 1 error for values of Z_{Lower} vs Z_{Upper}. In these contour plots, increasingly red colored regions indicate an overall Type 1 error beyond the upper limit of the 95% CI of $\alpha = 0.05$ using 100,000 simulations. Increasingly bright green regions indicate an overall Type 1 error

that is more conservative than expected based on the lower limit of the 95% CI of $\alpha = 0.05$ using 100,000 simulations.

The general behavior for the overall Type 1 error for these test statistics vary. However, we generally see an inflation of overall Type 1 error when we consider an adaptation in accrual within this promising zone regardless of the timing of the adaptation considered. In summary:

- Unblinded adaptations for the logrank statistic do not seem to affect the overall Type 1 error in this case. This is to be expected, as the true information growth of the unweighted logrank statistic is proportional to the number of events.

- The overall Type 1 error can be inflated as high as 6.2% for the $G^{1,0}$ statistic. This is seen across values ranging from $\Phi^{-1}(0.05)$ to 1.96.

- For $G^{0,1}$ and $G^{1,1}$, the behavior of the Type 1 error inflation is in direct contrast to the $G^{1,0}$ statistic.

- This often leads to more conservative overall Type 1 error for $G^{\rho,1}$ for $\rho \geq 0$ in regions where $G^{1,0}$ has an inflation of overall Type 1 error.

- Regions that result in an inflation of the overall Type 1 error for $G^{0,1}$ are typically smaller (i.e., smaller range of interim Z estimates), have the estimated interim treatment effect close to statistical significance in the preferential direction for superiority of the treatment over the placebo. On the other hand, when the promising region is broader, the $G^{0,1}$ statistic tends to be more conservative in terms of overall Type 1 error. This behavior is similar for the $G^{1,1}$ statistic.

- In order to inflate the overall Type 1 error beyond 10% with the use of the $G^{1,1}$ statistic, the total accrual size has to be considerably larger than 2000.

Intuitively, if the number of events is some linear surrogate to measuring the information growth with the use of these weighted statistics, as in the immediate setting, then any potential increase in accrual during the course of the trial through the use of unblinded adaptations without modifying our total number of events should ideally not inflate the overall Type 1 error. Since the overall number of "events" has not been adapted, naïvely speaking, we should be protected in terms of the overall Type 1 error.

The results in this section indicate that the number of events is no longer appropriate to characterize information growth when applying these "less well-understood" survival methods. The additional accrual of subjects via this intent-to-cheat procedure must thus be affecting the information growth in some aspects. We now expand our understanding of the behavior of the information growth of these statistics from Gillen and Emerson [2005] and investigate the impact of changing accrual part way through the study in an unblinded manner on our information growth to better understand the root cause of this inflation of overall Type 1 error.

6.4 What Goes Wrong: Impact of Censoring on Information Growth

We investigate the impact of changing the censoring distribution on the information growth with the use of these weighted logrank statistics. To do so, we considered plausible underlying survival scenarios that may be practical in a typical clinical trial. For example, we can consider survival settings with a pattern of long term survival which is typical in prevention settings, where the event rate may take a sufficiently long time to accumulate. In more serious, life-threatening rare disease settings such as Ebola, or Stage 4 Melanoma, they may be represented by survival distributions with shorter median time. We investigate some of these settings under long term survival based on the previous example, as well as short term survival (a Weibull distribution with shape parameter 0.5, and "rate" parameter of 1) in order to characterize the information growth.

Using the design specifications as in the previous section, we compute the information fraction by taking the ratio of the variance estimate at the interim analysis relative to the

total variance at the end of the trial (defined as 765 events). Under balanced randomization, the maximum statistical information when using the logrank statistic would be $\mathcal{V} = D/4 = 191.25$. A naïve user might presume that this was also the maximal statistical information when using weighted logrank statistics. To illustrate the error associated with such naïve assumptions, we compare the statistical information for weighted logrank statistics (as estimated from simulations) to the naïve ratio of the total number of events at the interim analysis relative to the total number of events at the final analysis. This will enable us to characterize the degree to which the true information growth is well approximated as a linear function of the number of events. We note that we do expect the true information growth to depend on both the underlying survival distribution and the censoring distribution.

6.4.1 Information Growth without Accrual Size Adjustment

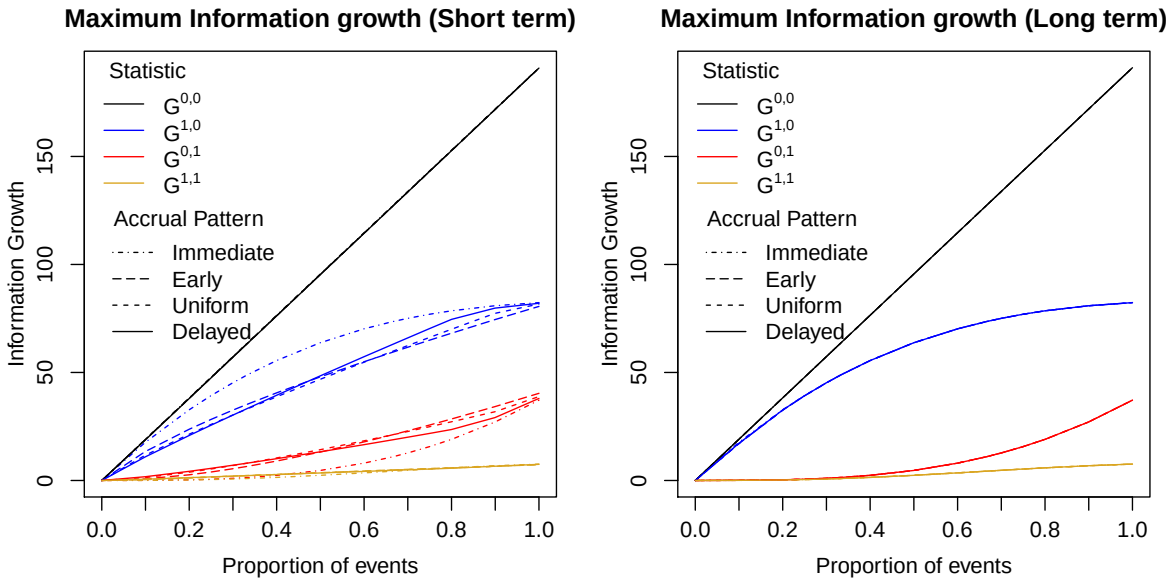

Figure 6.5: The average information growth for the various test statistics are different. In addition, the magnitude for which the average maximum statistical information at the planned final event size is vastly different for each test statistic of interest such that the assumption of a linear trend in number of events is no longer appropriate.

The average maximum statistical information for our choice of $G^{\rho,\gamma}$ family accumulates differently depending on the weighting scheme (Figure 6.5). The $G^{1,1}$ test statistic is seen to have the smallest maximum statistical information relative to the unweighted logrank test within the $G^{\rho,\gamma}$ family class of statistics. Relative to the proportion of total events, only the unweighted logrank statistic has linear (maximum) information growth unaffected by patterns of accrual.

The accrual distribution is seen to impact the information growth under short term survival more drastically relative to a long term underlying survival. The immediate entry characterizes the information growth that is unaffected by accrual. Compared to the immediate setting, the information growth presents more variability depending on the type of accrual pattern for the weighted logrank statistics. This induces a differential rate of growth in the short term survival setting as seen in Figure 6.5. Because the accumulated number of events is happening at a much slower rate in the long term setting, there is less variability seen in the information growth.

Figure 6.6 describes the cumulative proportion of information at interim analyses relative to the final statistical information as a function of the cumulative fraction of events under the short term survival setting. For the $G^{0,0}$ or the logrank statistic, the information growth is linear with respect to the proportion of events under all patterns of accrual and all survival distributions. Under the extreme setting of immediate entry, the information growth tends to be non-linear for the $G^{1,0}$, $G^{0,1}$, or $G^{1,1}$ statistic. With staggered entry, we observed more variations to the information growth for the weighted statistics.

The rate of information growth is related to the choice of weighting in these weighted statistics. The $G^{1,0}$ statistic places emphasis on early survival and the information growth is seen to increase more rapidly at earlier fraction of events as compared to the information growth later. This is evidenced by the larger spacing between consecutive points at earlier fraction of events relative to later fractions. The $G^{0,1}$ statistic, on the other hand, places relatively more weight on the late differences than early differences. The $G^{1,1}$ statistic places relatively more weight on mid differences rather than early or late differences.

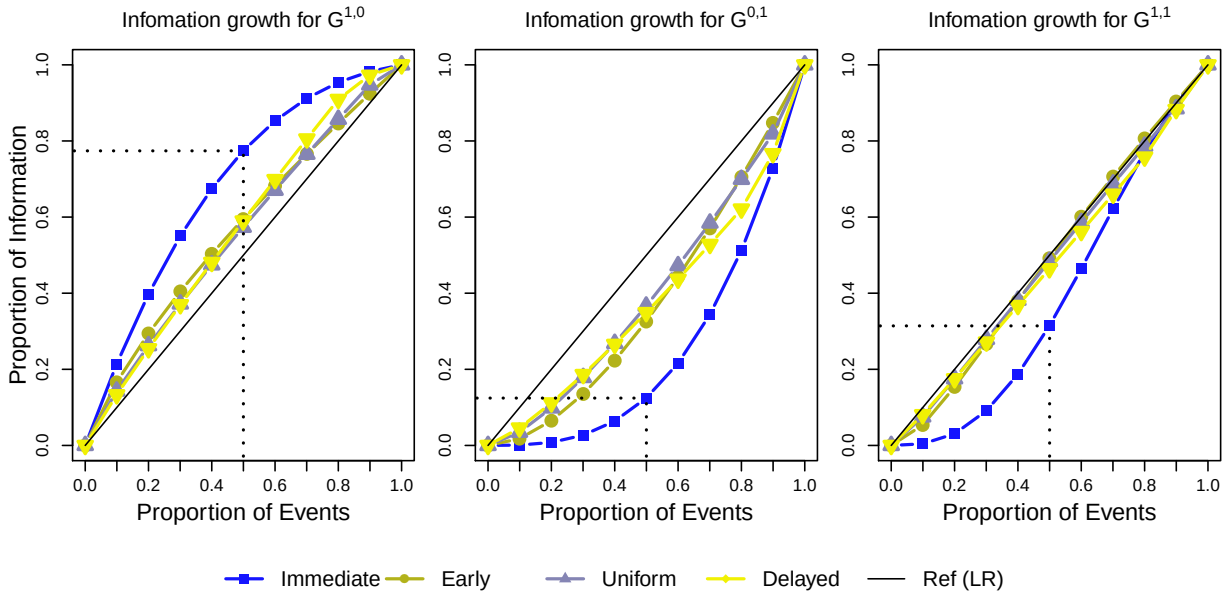

Figure 6.6: Cumulative proportion of statistical information relative to the fraction of the total number of events at each analysis for the weighted logrank statistics under short term survival. The logrank statistic is presented as a benchmark by the diagonal black line. Cumulative proportion of statistical information for the long term survival correspond to the results under immediate entry (blue lines).

Proportionate information tends to the limiting case of immediate accrual under long term survival. The (average) maximum information in the weighted logrank statistics are no longer linearly related to the event size. We now investigate the information growth when we make changes to accrual partway through the study.

6.4.2 Information Growth with Accrual Size Adjustment

We investigate the setting when we modify the accrual size of the study at an interim analysis while presuming the total number of events to be fixed at the end of the study. Following interim analysis at predetermined event size, we accrue the remaining 1000 patients that may not have entered the trial at double the original rate of accrual. We also included

the immediate accrual setting to exaggerate the degree to which extreme accrual can affect information growth. Thus, following an interim analysis under the immediate accrual setting, we immediately recruit the additional required patients into the study.

We describe the results for this section based on the long term survival setting simulated from the Weibull distribution with "shape" parameter of 0.5 and "rate" parameter of 120.1 for both treatment and placebo arm. The results based on short term survival simulated under the strong null present trends in information growth with slightly more variability under different accrual patterns.

6.4.2.1 Long Term Survival

Figure 6.7 shows the plot of the information growth of the various test statistics vs the proportion of events (relative to final event size of 765). For the logrank statistic, information growth is predictable in the sense that given the final event size and the randomization ratio, we can estimate the information at each interim analysis to determine the proportionate information. This observed linear trend also suggests that the accrual size does not modify the information growth of the test statistics when we hold the total number of events fixed.

In the case of the $G^{1,0}$ statistic, we see in Figure 6.7 that the estimated information growth following interim modifications to the accrual changes drastically. We note that in the extreme scenario when we have immediate accrual of patients following interim analysis, the information growth increases beyond the presumed maximum statistical information based on 1000 subjects. At later interim analysis, we observed similar trends when we attempt to increase our sample size following interim analysis immediately. Under various accrual patterns, we see similar patterns for the information growth as we make modifications to the sample size after interim analysis (Figure 6.7). Information growth tends to be more linear following a large sample size modification regardless of accrual pattern.

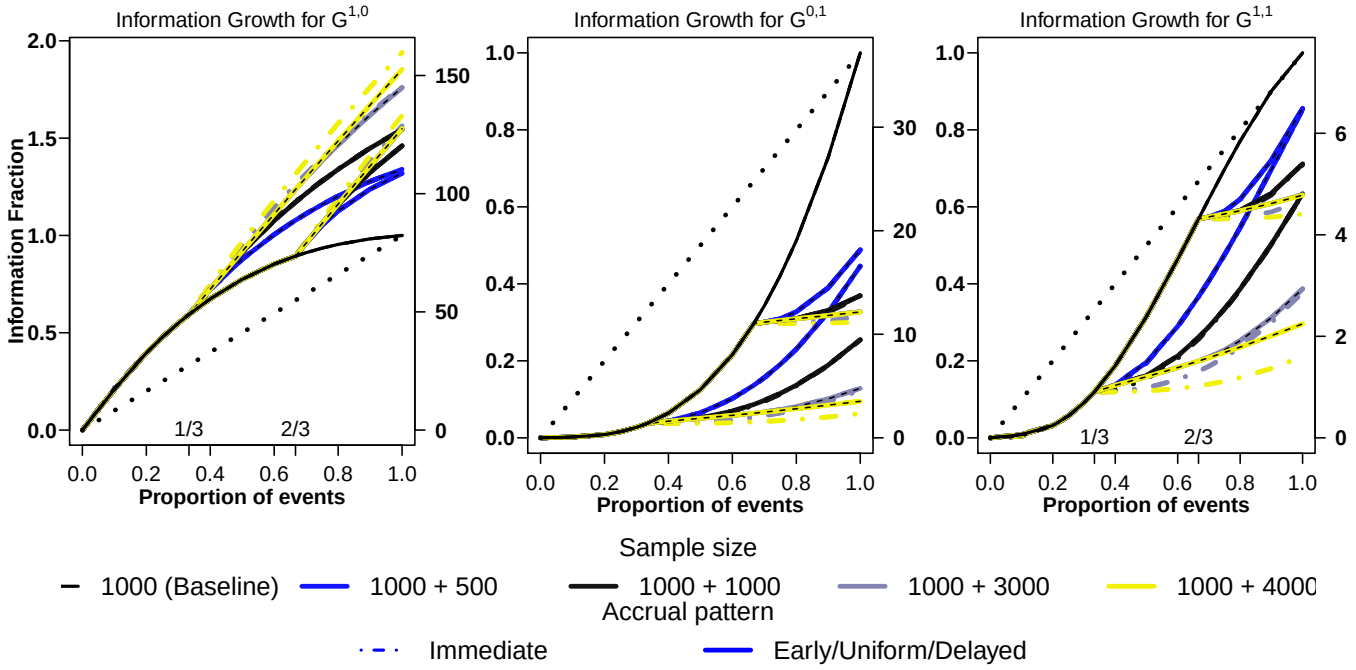

Figure 6.7: Proportionate information for various $G^{\rho,\gamma}$ family under the various scenarios when we increase the accrual size at 1/3 or 2/3 of the final events for long term survival scenario. Information growth remains linear for the logrank statistic while we see (1) non-linear growth for other families of the $G^{\rho,\gamma}$ statistic, and (2) changes to the maximum statistical information relative to the original design with 1000 subjects. The diagonal dotted line represents the setting where information growth is proportional to the number of events at each interim analysis.

With either the $G^{0,1}$ ot $G^{1,1}$ statistic, we see that any modification to the censoring distribution also affects the information growth. Under the most extreme scenario when we accrue all the sample size we want following an interim analysis, while holding the maximum number of event fixed, the maximum statistical information is very much reduced. This is because with this huge amount of subjects accrued, the entire information growth is weighted heavily by $1 - S(t^-)$. Thus, unless the number of events is increased or the accrual is allowed to be conducted in a prolonged manner, the information growth will be be weighted away from the information growth relative to the original design based on 1000 subjects. That is, the additional events accrued under this adaptive accrual scheme contribute very little information to the final analysis.

With various accrual patterns, the information growth also demonstrates minor increase in the statistical information. This slower rate of information growth, as a consequence of this additional influx of subjects contributing immense weight to the earlier portion of the pooled survival (and thus the pooled probability of not being censored), also increases the variability, and thus reduces the precision to which we can accurately measure our statistical information across interim analyses. In specific scenarios described in Appendix E.5, we see that this results in the apparent behavior of "backward" information when our statistical information is not growing sufficiently fast across interim analyses.

6.5 Application of Adaptive Procedures to Control for Inflation of Type 1 Error

It is sometimes described in the adaptive literature that if no adaptation was made to the statistical information during an unblinded analysis, one need not adjust for the adaptive analysis [Müller and Schäfer, 2001]. When a clinical trialist naively believes that statistical information is proportional to the number of events, he/she may regard that changes limited to accrual patterns (i.e. with no changes to the maximal number of events) do not represent an adaptation. This naïve principle does not apply to the weighted logrank statistics when unblinded analysis is used to decide on the amount of additional accrual necessary. We have

seen the consequence of how additional accrual of subjects under the strong null has lead to an inflation of overall Type 1 error with the use of these weighted logrank statistics, because such changes in accrual patterns modify the censoring distribution. Additionally, the root cause leading to this inflation is that our maximum statistical information, within the class of weighted logrank statistics, has been changed when we adaptively increase accrual in an unblinded manner. Because the information growth has changed, a statistical adjustment is now necessary when using the weighted logrank statistics.

We considered the general procedure described in Cui et al. [1999] to adjust our final critical value to ensure control of the overall Type 1 error after making an adjustment in statistical information. We note that other authors [Proschan and Hunsberger, 1995, Lehmacher and Wassmer, 1999, Müller and Schäfer, 2001, Chen et al., 2004, Gao et al., 2008] also describe the other versions of the above procedure and they have been shown to be equivalent under the two-stage setting by Jennison and Turnbull [2003]. For simplicity, we shall abbreviate this procedure as CHW since Cui et al. [1999] formally introduced this as a closed form adjustment in the two-stage setting.

The procedure for this approach is as follows: At the penultimate analysis, we modify our statistical information from $V(t_{J-1})$ at information time t_{J-1} to the new statistical information $V(\zeta_J)$ rather than continuing to the originally planned information $V(t_J)$ based on the observed (unblinded) estimated treatment effect. To ensure control of the overall Type 1 error, we adjust the final critical value from z_J to $z_J^\star$ based on the current estimated treatment effect $\widehat{\theta}$ or interim test statistic $\widehat{Z}(t_{J-1})$, and current information $V(t_{J-1})$ to

$$z_J^\star = \frac{1}{\sqrt{V(\zeta_J)}}\left[\frac{\sqrt{V(\zeta_J)-V(t_{J-1})}}{\sqrt{V(t_J)-V(t_{J-1})}}\left(z_J\sqrt{V(t_J)}-\sqrt{V(t_{J-1})}Z(t_{J-1})\right)+\sqrt{V(t_{J-1})}Z(t_{J-1})\right]$$

where we substitute $Z(t_{J-1})$ by the interim estimated $\widehat{Z}(t_{J-1})$ statistic. The degree to which the maximum statistical information is correctly specified is later seen to also affect the degree of Type 1 error control.

6.5.1 Flexible Procedures: Only Adjust when Adapting the Accrual Size

There are several challenges with the use of the CHW procedure. First, this requires understanding the information growth for the test statistics of choice. Second, when an adaptive procedure is used, i.e., when we adapted to a new design, we need to know what is the original maximum statistical information, $\mathcal{V}(t_J)$ of the original design. Thus, if the design protocol is unclear, then we do not know what is the original design. Otherwise, following an interim analysis, after we adapted to the new design, we no longer "observe" the original design. As such, there is a need to prespecify what is the maximum statistical information of the original design so as to implement the CHW approach. In particular, when the information growth is highly nonlinear as seen in the previous section, this can give rise to misspecification since we may not know the true censoring distribution or underlying survival in a clinical trial.

We first investigate the common claim in the adaptive literature [Mehta and Pocock, 2011] that one only needs to adjust using CHW for the adaptation when the "sample size" is modified. We consider pre-specifying the statistical information $\mathcal{V}(t_2)$ and only apply CHW when the accrual is modified based on unblinded interim analysis. We also define the original sample size of the design to be based on 1000 subjects. Thus, $\mathcal{V}(t_2)$ is defined based on the presumed accrual distribution, as well as underlying survival based on 1000 subjects.

When one does not adapt, $\mathcal{V}(t_2)$ is equivalent to $\mathcal{V}(\zeta_J)$, and thus the naïve z_2 can be used as a critical value. However, when an adaptation is made to the accrual, the information growth can be modified such that we may not have the precision to accurately estimate what would have been the information growth had we not adapted. This difficulty is further amplified by the non-linear behavior of the information growth for these weighted logrank statistics.

Hence, we first consider the potential strategies to only adjust when we adapt. Then, we investigate the consequences when $\mathcal{V}(t_2)$ is not correctly estimated. The above two settings are sufficient to characterize some issues with the adjustment procedure. We note that in the G^ρ family when $\gamma = 0$, the estimated information growth will always increase since the

precision of $S(t^-)$ will improve with additional accrual of subjects into the trial.

6.5.1.1 Only Adjust when We Adapt the Sample Size; $\mathcal{V}(t_2)$ is Correctly Specified

We use the mean of $\mathcal{V}(t_2)^{\text{True}}$ that is based on all the simulations and specify this as the final statistical information for the original design to be used after an adaptation to increase accrual size is made. At level $\alpha = 5\%$, we then compute the proportion of times the Z statistic is rejected when all the total number of events are accumulated. With this procedure, there will be times when the naïve critical value of $z_J = 1.645$ is used, and at other times $z_J^\star$ will be used based on $\mathcal{V}(\zeta_2)$ and prespecified $\overline{\mathcal{V}}(t_2)^{\text{True}}$. In these settings, there is adequate control of the overall Type 1 error after application of the adjustment procedure.

6.5.1.2 Only Adjust when We Adapt the Sample Size; $\mathcal{V}(t_2)$ is Incorrectly Specified

We investigate whether the adaptive approach using CHW can control the overall Type 1 error at level $\alpha = 5\%$ when $\mathcal{V}(t_2)$ is incorrectly specified. We present the simulation results where $\mathcal{V}(t_2)$ was incorrectly specified for $G^{1,0}$, $G^{0,1}$, and $G^{1,1}$. We refer to setting #5b to illustrate the results since this most resembles the approach of the "promising" zone in the adaptive literature. In particular, we specify $\mathcal{V}(t_2)^{1000}$ for $G^{1,0}$, $G^{0,1}$, and $G^{1,1}$ to be 110, 20, and 6.5 respectively. Had no design changes been made, the true average statistical information $\mathcal{V}(t_2)$ for each of the test statistics $G^{1,0}$, $G^{0,1}$, and $G^{1,1}$ would be 82.5, 37.25, and 7.75 respectively.

The overall Type 1 error of $\alpha = 5\%$ is not controlled everywhere when we only adjust the critical value when an adaptation is made at an interim analysis to double the original total sample size while presuming the naïve d_J at other times (Figure 6.8) . The $G^{1,0}$ statistics has an inflated Type 1 error when adaptations are made in the "promising" region. When the Z_{Lower} is positive, the overall Type 1 error can be inflated as high as 5.6% (>10% higher than the nominal Type 1 error on the relative scale). The $G^{0,1}$ statistic is seen to be more

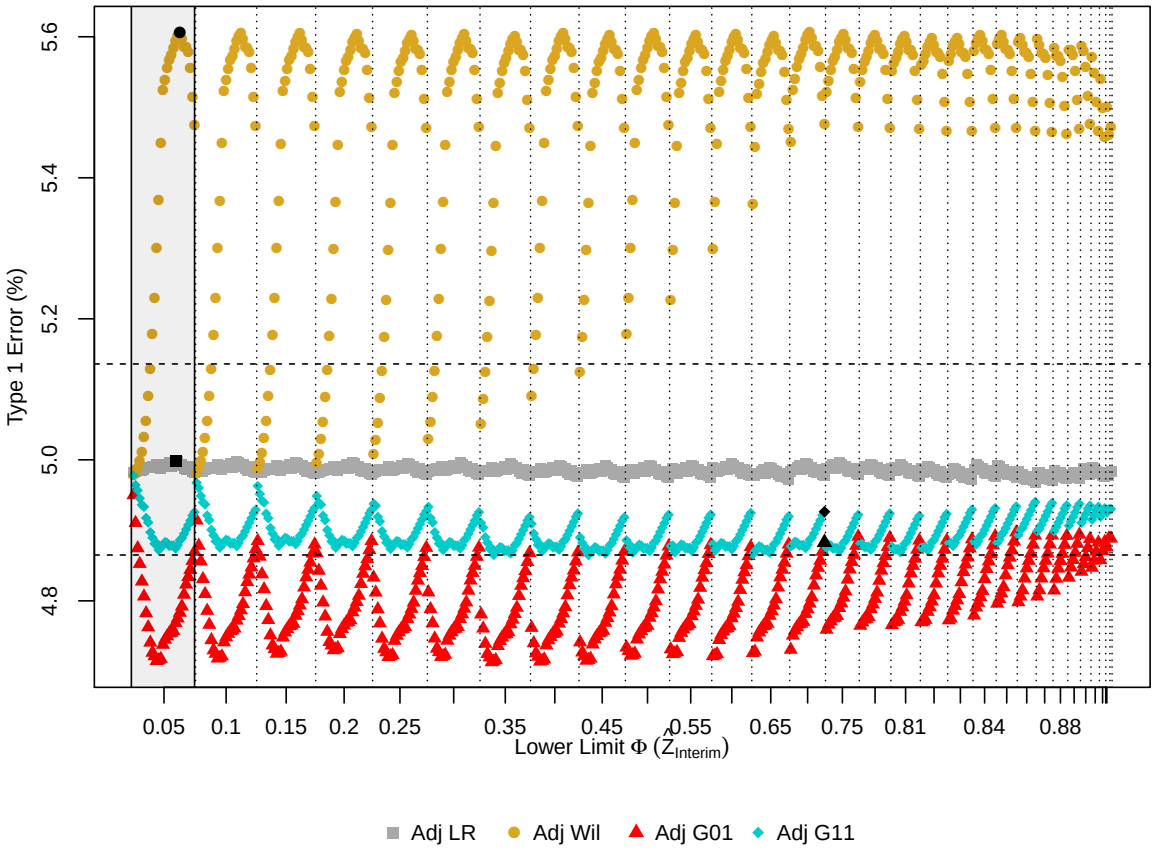

Figure 6.8: Overall Type 1 error rate for the procedure where we only adjust when one makes an adaptation, and incorrectly specify the maximum statistical information at design stage.

conservative than the nominal 5% when adaptations are made in the "promising" region. The $G^{1,1}$ statistic is borderline conservative.

These results indicate that the naïve concept of only adjusting when an adaptation is made do not necessary hold true in general. In particular, when there is lack of precision in accurately quantifying the information growth of our original design, and also choosing to adjust when an adaptation is made, the CHW procedure does not offer the protection of the overall Type 1 error as claimed. Worse still, when applying these weighted logrank statistics,

the non linear and highly unpredictable behavior of the information growth gives rise to great difficulty in accurately quantifying the statistical information when no adaptation is made.

6.5.2 Fully Adjusted Procedures

It thus seems that if we always adjust using CHW, we should ideally control for this uncertainty involved when there is substantial variability in our test statistics when these weighted statistics may be affected by the censoring distribution. We consider the ideal setting where we "always" adjust for this unspecified adaptation regardless of whether an unplanned, unblinded adaptation was made to increase accrual (possibly decreasing accrual too).

We first investigate the scenario of always adjusting for the correct original statistical information defined by the average $\widehat{\mathcal{V}}(t_2)$ based on simulations regardless of adaptation. We then relax this assumption and attempt to apply this procedure consistently even when this prespecified $\mathcal{V}_{\text{Prespecified}}(t_2)$ is incorrect. In other words, when we do not change the course of the design, i.e., we assume our original accrual size of 1000 and holding the total number of events at 765, we adjust the critical value using CHW based on $\mathcal{V}_{\text{Prespecified}}(t_2)$ rather than the observed $\widehat{\mathcal{V}}(t_2)$. As such, it is possible that $\mathcal{V}_{\text{Prespecified}}(t_2)$ may be different from the observed $\widehat{\mathcal{V}}(t_2)$.

6.5.2.1 Simulation Results

The results for adjusting for the right $\mathcal{V}(t_2)$ or original statistical information regardless of an adaptation controls the overall Type 1 error. When this prespecified maximum statistical information $\mathcal{V}_{\text{Prespecified}}(t_2)$ is different from what was obtained (i.e., $\widehat{\mathcal{V}}(t_2)$) had we not adapted, and we consistently adjust for what we prespecified (i.e., $\mathcal{V}_{\text{Prespecified}}(t_2)$), this overall Type 1 error is also maintained. These results have direct implications when applying the CHW procedure not only with these weighted logrank statistics, but also with many of the other "less well-understood" analyses methods in the time to event setting. Because the information growth in the survival setting is much harder to predict or quantify with these "less well-understood" approaches, the concept of adaptation becomes more difficult when

we cannot accurately predict information growth. Based on our explorations, unless we can correctly quantify the statistical information when obtaining any number of events (such as with the logrank statistics), the adjustment procedure of CHW has to applied regardless of the analyses method of choice.

This approach of always adjusting based on the prespecified statistical information at design stage when applying the procedure of CHW appears counter-intuitive. However, the procedure of CHW can be interpreted as some average across all potential adaptations that have occurred, as well as those that had not have occurred but may be considered. By consistently adjusting for this imprecise ("incorrect") but prespecified statistical information, we are considerably adjusting for some form ancillary statistic that accounts for all potential adaptations one would have imagined based on the unblinded interim results of the primary endpoint. Thus, in some sense, we *have* accounted for potential adaptation choices that have happened as well as those adaptations that did not lead us to increase/decrease accrual.

6.6 Summary

In this chapter, we demonstrated that the lack of understanding of information growth can affect the degree of control of the overall Type 1 error when applying these adaptive procedures to these "less well-understood" survival methods. When we want to emphasize clinical importance and efficiently weight the survival curves at different times, we may choose weighted statistics to gain power under these hypothesized alternatives. However, the evaluation of the overall Type 1 error is typically performed under the strong null hypothesis of proportional hazards. In the class of weighted logrank statistics, information growth is a function of the censoring distribution, the entry distribution, and the number of events. Any form of adaptive modification to only the censoring distribution thus has a direct impact on the information growth of the family of weighted logrank statistics. Hence, by naïvely presuming that the number of events is always a surrogate for information growth in general time to event settings, this can result in undesirable inflation of the overall Type 1 error when using the weighted statistics.

Additionally, while we can apply the adaptive methods to adjust for such unblinded adaptations, the naïve claim of only having to adjust when an adaptation is made does not hold true unless we can precisely quantify the entire information growth when an adaptation is made, as well as when an adaptation is not performed. With weighted logrank statistics, information growth lacks predictability when we modify our accrual. Thus, if we lack precision to estimating information growth, then we are no longer able to apply CHW efficiently to help control for this unblinded adaptations. Such is true when we change accrual or even change analysis methods. Application of adaptive designs in TTE studies using "less well understood" survival analysis is much harder to implement in practice.

In summary, our investigation led to some interesting conclusions regarding the use of adaptive methods in the time to event setting. So long as one cannot precisely quantify the information growth for the weighted analysis, the CHW procedure should be applied regardless of whether an adaptation is made in order to preserve the overall Type 1 error. Additionally, investigators must also prespecify what is their predicted original statistical information based on hypothesized accrual patterns under the null hypothesis to allow for CHW to be applied. This means that even when we can/cannot precisely quantify the maximum statistical information, we have to use this "imprecise" quantity to adjust using CHW for consistency.

While the results of this section are obtained based on assumption of the strong null, later, under time varying treatment effect, we may be interested in controlling our probability of rejection under the weak null hypothesis. In these settings, the censoring distribution and information growth are intricately tied together such that any form of interim analysis can further lead to differential weighting of the survival curves with the use of logrank statistic. Under such scenarios, we may no longer choose the (weighted) logrank statistics but other test statistics may be favored.

Chapter 7

Evaluation of Designs in the Setting of Anticipated Crossing Survival Curves

In a time to event clinical trial, researchers may posit the possibility that survival curves may cross at some time point. Logan et al. [2008] and Logan and Mo [2015] considered the analysis of censored time to event data with the objective of detecting the "better" treatment when crossing survival curves are quite plausible. We investigate the alternatives used by Logan et al.'s composite statistics, and use simulations to assess how their test statistics behave under the setting of non proportional hazards, both with stochastically ordered and with potentially crossing survival. We further investigate the behavior of the alternative test statistics as a function of the censoring distribution. We then examine how this may impact any sequential designs, where a DMC may make an interim decision to stop the trial early or adapt trial parameters based on unanticipated differences in survival. We find little advantage to the use of Logan et al.'s proposed composite statistics over judicious choice among the commonly used test statistics in the sequential survival setting.

7.1 Introduction

Both "well-understood" and/or "less well-understood" designs involve quantifying the information growth at each interim analysis. To enable application of these procedures under time varying treatment effects, it is crucial we understand some of the issues lingering with the use of group sequential methods, because GSD is a special case of adaptive design. Of concern in the time to event setting is that many of such methods are "less well-understood" or in-

adequately characterized to enable understanding of application of these statistical methods in planning a sequential study. The natural censoring that arises from the use of interim analyses truncates the survival, potentially modifying the estimates based on any of the statistical methods, and most importantly, the rate of information growth. The theme in this chapter is: How does censoring affect our knowledge of these interim estimates as well as the information growth under the presumption of time varying treatment effects (such as might be characterized in the time to event setting as non proportional hazards)?

We motivate this chapter based on the clinical setting taken from Logan et al. [2008] with the objective of identifying the better treatment with the better survival by some fixed time point as claimed in Logan et al. [2008]. Using the composite statistics introduced in Logan et al. [2008] and subsequently extended by Logan and Mo [2015] to sequential testing, we describe the scientific issues in this poorly characterized bivariate parameter space in section 7.2. We evaluate their test statistics under the weak null hypothesis in the fixed sample design framework. In section 7.3, we consider both stochastically ordered survival distributions, as well as distributions having crossing hazards. We then return to investigate the consequences of censoring on estimates of the treatment effect and information growth in section 7.4.

In section 7.5, we note how the lack of guidance on the use of composite statistics may present dilemmas whereby the DMC may not be able to reliably judge the trial using various statistics. When unanticipated large differences in survival are observed early on during DMC monitoring, the DMC may need to act on the emerging data to judge whether there is equipoise. While approaches considered in section 7.5.1 can be used to recalibrate the boundaries to circumvent such situations from occurring in the clinical setting, the lack of understanding of some of these "less well-understood" analyses methods in the time to event setting present as much issues with "less well-understood" adaptive designs. We conclude with a summary of the issues to ponder upon during the design stage of the study and the considerations that must be anticipated when dealing with any time to event driven trial with sequential analyses.

7.2 Use of Composite Statistics

Using autologous vs allogeneic bone marrow transplantation as a motivating example in their 2008 paper, both Logan et al. [2008] and Logan and Mo [2015] argued that the treatment with the greatest long term survival would generally be preferred. Logan et al. [2008] and Logan and Mo [2015] proposed basing inference on a two-sided test based on a quadratic form constructed from the joint distribution of an estimate of the survival curve at a pre-specified time and a weighted log rank statistic that has zero weight assigned prior to the pre-specified time under a fixed sample design. They recommended the use of two composite statistics that might *a priori* be thought to provide greater power to distinguish such late differences in survival in the fixed sample setting as compared to more commonly used approaches in limited simulation settings. Subsequently, Logan and Mo [2015] extended their procedures to the group sequential framework by proposing suitably normalized versions of the linear composite statistics that have the desired "independent increments" property. Furthermore, they demonstrated that their composite statistics beat other competing, and commonly used test statistics in the sequential setting to test for "long term" survival benefit under specific alternatives they constructed.

We find that insufficient characterization has been made in their description of the composite statistic(s) as well as evaluating the operating characteristics in both papers. There are key scientific and statistical issues with the use of their test statistics. Generally, the lack of guidance on how to interpret the results based on their statistics, such as a point estimate, or confidence interval, makes it difficult to quantify the effect of long term benefit that is the objective in many randomized clinical setting. We describe some of the statistical issues with interpretation of the parameter space with the use of the composite statistics. We characterize the asymptotic properties of their test statistics using the distribution of the Z statistic that has interpretation as standardized alternatives (Refer to Appendix F.1). In summary, their statistics are truly directed towards crossing hazard functions than crossing survival curves, and may mislead naïve clinical trialists when planning a time to event study.

7.2.1 Composite Statistics

Logan et al considered application of the composite statistics that included an "appropriately" calibrated sum of Nelson-Aalen and Log-rank test statistics to address a modified null hypothesis, namely, a composite null hypothesis. The null hypothesis is stated as $\mathbb{H}_0 : S_0(t) = S_1(t), \forall t \geq \tau_0$ where $S_1(t)$, $S_0(t)$ denote the survival curves at time t for the treatment (1) and placebo (0) group, and τ_0 denotes some prespecified time of potentially crossing survival curves. They note that this null hypothesis is equivalent to testing $\mathbb{H}_0 : \{S_1(\tau_0) = S_0(\tau_0)\} \cap \{\lambda_1(t) = \lambda_0(t), t > \tau_0\} = \mathbb{H}_{01} \cap \mathbb{H}_{02}$ where $\lambda_k(t)$ represents the hazard function at time t for group k taking values 0 and 1. By formulating the hypothesis as such, they thus considered testing the hypothesis of equality of survival curves at τ_0, and the hypothesis of no difference in hazard functions after time τ_0.

Under $\mathbb{H}_0$, they formulated alternative test statistics such that $\mathbb{H}_{01} : S_1(\tau_0) = S_0(\tau_0)$ can be tested using the standardized difference in Kaplan Meier estimates or the standardized difference in Nelson Aalen estimate of the cumulative hazard. Then, the hypothesis $\mathbb{H}_{02} : \lambda_1(t) = \lambda_0(t), t > \tau_0$ can be tested using the left-truncated log-rank test statistic.

The "linear combination test statistic" is defined by $Z_{\text{OLS}}(\tau_0, t) = \frac{Z_{\text{NA}}(\tau_0,t)+Z_{\text{LR}}(\tau_0,t)}{\sqrt{2}}$. The "quadratic test statistic" is defined by a sum of the squares of the individual components of the linear test statistics such that $\chi^2(\tau_0, t) = Z^2_{\text{NA}}(\tau_0, t) + Z^2_{\text{LR}}(\tau_0, t)$ at some fixed time t. $Z_{\text{NA}}(\tau_0, t)$ is the Nelson-Aalen statistic computed from time 0 to the time of anticipated crossing τ_0 with the survival curves updated to time t for $t > \tau_0$. $Z_{\text{LR}}(\tau_0, t)$ is the truncated logrank statistic computed from time of crossing τ_0 to time t for all $t > \tau_0$. These two statistics can be found as equation (5) and (7) from Logan et al. [2008] respectively.

Since events happening prior to τ_0 are independent of events happening after τ_0 under noninformative censoring, $Z_{NA}(\tau_0, t)$ and $Z_{LR}(\tau_0, t)$ are thus independent and have joint asymptotic bivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$ equivalent to the identity matrix. Using suitably normalized weighting which they found most appropriate for the setting they are interested in, they proposed two composite test

statistics with asymptotic distributions of the following

$$Z_{\text{OLS}} = \frac{Z_{NA}(\tau_0, t) + Z_{LR}(\tau_0, t)}{\sqrt{2}} \;\dot{\sim}\; \mathcal{N}(0, 1)$$
$$Z_{\text{Quad}} = [Z_{NA}(\tau_0, t)]^2 + [Z_{LR}(\tau_0, t)]^2 \;\dot{\sim}\; \chi^2_2.$$

Z_{OLS} has the analogous interpretation to the ordinary least squares statistic proposed by O'Brien [1984] to accommodate multiple testing of endpoints. Logan et al. [2008] recommend in particular the quadratic test as an omnibus test against various simulation setting explored in their paper while favoring also the Z_{OLS} to have better power to identify the better treatment.

7.2.1.1 Under local alternatives

Under some local alternatives, appealing to asymptotic normality, we note that the individual components of the composite statistics can be represented in the following:

$$Z_{NA} \;\dot{\sim}\; \mathcal{N}(\theta_{NA}/\sqrt{V_{NA}}, 1) \cong Z^2_{NA} \;\dot{\sim}\; \chi^2_1(\delta^2_{NA} = \theta^2_{NA}/V_{NA})$$
$$Z_{LR} \;\dot{\sim}\; \mathcal{N}(\theta_{LR}/\sqrt{V_{LR}}, 1) \cong Z^2_{LR} \;\dot{\sim}\; \chi^2_1(\delta^2_{LR} = \theta^2_{LR}/V_{LR})$$
$$Z_{NA}(\tau_0, t) \;\dot{\sim}\; \mathcal{N}(\delta_1, 1)$$
$$Z_{LR}(\tau_0, t) \;\dot{\sim}\; \mathcal{N}(\delta_2, 1)$$

where $V_{\text{NA}}, V_{\text{LR}}$ denote the variance of the Nelson-Aalen estimator and the variance of the logrank statistic respectively.

We shall refer to $\delta_1, \delta_2, \delta, \delta_{NA}, \delta_{LR}$ as standardized alternatives since they are a function of the parameter of interest and the precision of the parameter of interest. This standardized notation is later of interest for characterizing the time varying treatment effect for comparison with other time to event analyses methods.

The composite statistics thus have asymptotic distribution of the form

$$Z_{\text{OLS}} = \frac{Z_{NA}(\tau_0,t) + Z_{LR}(\tau_0,t)}{\sqrt{2}} \quad \dot{\sim} \; \mathcal{N}\left(\delta_{\text{OLS}} \equiv \frac{\delta_1+\delta_2}{\sqrt{2}}, 1\right) \quad \cong \chi^2_1(\delta^2_{\text{OLS}})$$

$$Z_{\text{Quad}} = [Z_{NA}(\tau_0,t)]^2 + [Z_{LR}(\tau_0,t)]^2 \; \dot{\sim} \sum_{i=\{NA,LR\}} (Z_i+\delta_i)^2 \quad \cong \chi^2_2(\delta_{\text{Quad}} = \delta_1^2+\delta_2^2)$$

7.2.2 Scientific Interpretation with the Use of Composite Statistics

To describe the scientific issues with the use of the composite statistics, we consider the comparison of two treatments groups, A and B, and let τ_0 to be some anticipated time of potentially crossing survival curves. The null hypothesis $\mathbb{H}_0^L = \mathbb{H}_0^{\text{NA}} \cap \mathbb{H}_0^{\text{LR}}$ is the intersection of:

- $\mathbb{H}_0^{\text{NA}} : \text{S}_\text{A}(\tau_0) = \text{S}_\text{B}(\tau_0)$ (which is tested using $\text{Z}_{\text{NA}}(\tau_0,t)$ based on the Nelson-Aalen estimate of the cumulative hazard functions), and

- $\mathbb{H}_0^{\text{LR}} : \lambda_A(t) = \lambda_B(t), \; \forall t \geq \tau_0$ (which is tested using $\text{Z}_{\text{LR}}(\tau_0,t)$ based on a log rank statistic that places no weight prior to time τ_0 and equal weights to failure times thereafter).

There are interpretation issues with the use of the composite statistics by Logan et al. [2008]. In their paper, they concluded that their test is suitable as a two-sided test to obtain a "p-value" as evidence against the strong null hypothesis. However, when the objective of the main paper was to "identify the better treatment", we would tend to prefer a one-sided test in order to determine the better treatment. Secondly, the comparison of survival curves based on Logan's hypotheses is conditional upon the fact that the crossing is correctly anticipated. Additionally, the authors have failed to provide adequate guidance on the interpretation of the test statistics when this presumed crossing of survival curves may not be real.

To better delineate the issues mentioned above, we examine the sample space for which each component of the composite statistics by defining the following notations H_K^T where T

denotes the test type (NA or LR), and K to represent the treatment that is better and takes values A, B, or 0 where 0 indicates that treatment A and B are not different.

For comparisons of survival probabilities at τ_0:

- (treatment B is better) $\mathrm{H}_B^{\mathrm{NA}} : S_A(\tau_0) < S_B(\tau_0)$, and
- (treatment A is better) $\mathrm{H}_A^{\mathrm{NA}} : S_A(\tau_0) > S_B(\tau_0)$.

For comparisons of hazards after τ_0:

- (treatment B is better) $\mathrm{H}_B^{\mathrm{LR}} : \lambda_A(t) \geq \lambda_B(t);\ \forall t \geq \tau_0$ with $\lambda_A(t) > \lambda_B(t)$ for some $t > \tau_0$, and
- (treatment A is better) $\mathrm{H}_A^{\mathrm{LR}} : \lambda_A(t) \leq \lambda_B(t);\ \forall t \geq \tau_0$ with $\lambda_A(t) < \lambda_B(t)$ for some $t > \tau_0$.

We now consider the clinical interpretation of the various hypotheses. Note that our nomenclature is such that the parameter space is only partially ordered by these hypotheses:

- $\mathrm{H}_0^{\mathrm{NA}}$ and $\mathrm{H}_0^{\mathrm{LR}}$ are the strong null hypothesis
- $\mathrm{H}_0^{\mathrm{NA}}$, $\mathrm{H}_A^{\mathrm{NA}}$, and $\mathrm{H}_B^{\mathrm{NA}}$ partition the parameter space for $(S_A(\tau_0), S_B(\tau_0))$.
- $\mathrm{H}_0^{\mathrm{LR}}$, $\mathrm{H}_A^{\mathrm{LR}}$, and $\mathrm{H}_B^{\mathrm{LR}}$ do not partition the parameter space for $(\lambda_A(t), \lambda_B(t))$ for $t \geq \tau_0$, as it is unclear which treatment might be preferable if hazards cross after time τ_0.
- We definitely prefer treatment A if ($\mathrm{H}_0^{\mathrm{NA}}$ and $\mathrm{H}_A^{\mathrm{LR}}$), or ($\mathrm{H}_A^{\mathrm{NA}}$ and $\mathrm{H}_0^{\mathrm{LR}}$), or ($\mathrm{H}_A^{\mathrm{NA}}$ and $\mathrm{H}_A^{\mathrm{LR}}$)
- We definitely prefer treatment B if ($\mathrm{H}_0^{\mathrm{NA}}$ and $\mathrm{H}_B^{\mathrm{LR}}$), or ($\mathrm{H}_B^{\mathrm{NA}}$ and $\mathrm{H}_0^{\mathrm{LR}}$), or ($\mathrm{H}_B^{\mathrm{NA}}$ and $\mathrm{H}_B^{\mathrm{LR}}$)

However, for other combinations of the hypotheses, it is more difficult to characterize the preference, and in fact $\mathrm{H}_A^{\mathrm{LR}}$, $\mathrm{H}_0^{\mathrm{LR}}$, and $\mathrm{H}_B^{\mathrm{LR}}$ do not partition the parameter space following τ_0, because we could have crossing hazards. Nonetheless, the above are sufficient to explore the issues we find with the combination statistics in practice. Table 7.1 shows the hypotheses that result in crossing hazards and the possible conclusions from the use of the bivariate statistics on the outcome space.

Table 7.1: A tabular summary of the potential conclusions one can draw from this partially ordered sample space where we presume monotonic hazard ratio after τ_0.

		$H_0^{LR}\ \forall t > \tau_0$		
		$\lambda_A(t) < \lambda_B(t)$	$\lambda_A(t) = \lambda_B(t)$	$\lambda_A(t) > \lambda_B(t)$
H_0^{NA}	$S_A(\tau_0) < S_B(\tau_0)$	**Inconclusive**[1]	Treatment B	Treatment B
	$S_A(\tau_0) = S_B(\tau_0)$	Treatment A	**Inconclusive**	Treatment B
	$S_A(\tau_0) > S_B(\tau_0)$	Treatment A	Treatment A	**Inconclusive**[2]

[1,2] represents the scenario whereby we have crossing hazards.

The authors have constructed a sample space that is ultimately ordered by choosing one of their univariate composite statistics. We find it more important to consider our clinical preference for treatments in the partially ordered space defined using the two hypotheses above, namely, $H_0^{NA} : S_A(\tau_0) = S_B(\tau_0)$, and $H_0^{LR} : \lambda_A(t) = \lambda_B(t), \forall t \geq \tau_0$ respectively. According to the authors, the hypothesis can be tested by considering appropriately calibrated combinations of a linear or quadratic statistic, $Z_{NA}(\tau_0, t)$ or $Z_{LR}(\tau_0, t)$, in order to compare survival curves at some pre-specified time τ_0, and hazard functions after time τ_0.

In the context of their problem, their composite statistics appear to rely on the assumption as specified before the trial. However, it is possible to have stochastically ordered survival curves over time such that treatment A is always better than treatment B, and have the composite test statistics instead favor treatment B with high probability.

We note that such partitioning of the parameter space as seen in Table 7.1 covers four quadrants in the real space where the x-axis describes the standardized alternatives of the Nelson-Aalen statistic δ_1, and the y-axis represents the standardized alternatives of the truncated logrank statistic δ_2. In this parameter space, Quadrant I and III are consistent with the standardized alternatives pointing in the same direction, thus consistently identifying either treatment A, or B as the better treatment. On the other hand, Quadrant II and IV represent regions where the directions of the standardized alternatives are not in the same direction, thus leading to uncertainty regarding a conclusive identification of the better

treatment. More details about these standardized alternatives are described in terms of the probability of rejecting the null hypothesis in the Appendix F.1.

7.2.3 Naïve Interpretation of the Composite Statistics

One naïve interpretation of the test statistic is to presume that the sign of Z_{OLS} conveys the direction of the treatment effect. Note that the cumulative hazard $\Lambda(t) = -\log S(t)$. Then $-\log[S_A(\tau_0)] - \{-\log[S_B(\tau_0)]\} = \Lambda_A(\tau_0) - \Lambda_B(\tau_0)$ compares the difference in cumulative hazards at time τ_0.

- When $Z_{\text{OLS}} < 0$, we might (naïvely) presume that the cumulative hazard prior to time of crossing at τ_0 for treatment group A and the integrated weighted hazard past τ_0 is less than that of treatment group B. Thus, A is the preferred treatment.
- When $Z_{\text{OLS}} > 0$, we might (naïvely) presume that the cumulative hazard prior to time of crossing at τ_0 for treatment group B and the integrated weighted hazard past τ_0 is less than that of treatment group A. Thus, B is the preferred treatment.

However, even with this naïve interpretation that appears "consistent" with how the test statistics behave, this interpretation can be problematic. We now discuss the issues with the use of composite statistics under several non proportional hazards settings not considered in either Logan et al. [2008] or Logan and Mo [2015] in the FSD and sequential settings respectively.

7.3 Issues with the Use of Composite Statistics in the Fixed Sample Setting

In this section, we investigate Logan's composite statistics in the fixed sample setting. Recall in section 3, an adaptive design can be chosen to either expand the study based on a FSD or modify the maximum statistical information based on a group sequential design. In some sense, Logan et al's test statistics can be interpreted as an adaptive switching of the test statistics in the fixed sample setting. It is thus of use to investigate any potential issues with the use of Logan's composite statistics in such settings. If statistical issues related to

interpretation of results, or quantifying the level of evidence cannot be resolved in the fixed sample setting, then the use of the composite statistics will be expected to pose similar issues when extended to the group sequential setting.

7.3.1 Simulation Study Setup

Our main objective is to identify the treatment with the better survival based on the use of the test statistics. Thus, we want to have some form of summary measure at the end of the study that would quantify the evidence in favor of, or against a treatment strategy. We conducted a simulation study to describe the shortfalls of the composite statistics, and the issues pertaining to interpretation. We present summary statistics as well as the less commonly used test statistics in the fixed sample setting. We let $\tau_0 = 2$ to be the anticipated time of potentially crossing survival curves. For simplicity, we assumed that patients were immediately accrued in the fixed sample setting.

We simulate our survival curves based on some mixtures described as follows: Denote $M \sim \mathcal{B}\text{ernoulli}(\pi)$ to be the random variable that characterizes the mixture of exponential survival distributions after being assigned randomized treatment (either treatment A or B). After being randomized to treatment $k = \{A, B\}$, the survival time for a patient has some probability π of coming from the distribution corresponding to $M = 1$ with exponential rate λ_k^1, and probability $1 - \pi$ coming from the distribution corresponding to $M = 0$ with exponential rate λ_k^0. Thus, our survival time distribution based on the following mixtures of exponential distribution can be described as follows

$$f_k(t) = \begin{cases} \lambda_k^1 \exp(-\lambda_k^1 t) & M = 1 \text{ with probability } \pi \\ \lambda_k^0 \exp(-\lambda_k^0 t) & M = 0 \text{ with probability } 1 - \pi \end{cases}$$

For each mixture $M = \{0, 1\}$, denote the hazard ratio (Λ^k for $k = \{A, B\}$) between each treatment group to be $\lambda_A^1 = \Lambda^1 \lambda_B^1$ and $\lambda_A^0 = \Lambda^0 \lambda_B^0$. The expected survival at time t, $S_A(t)$, for *any* patient in the control group is $\pi \exp(-\lambda_A^1 t) + (1 - \pi) \exp(-\lambda_A^0 t)$. Sim-

ilarly, the expected survival at time t, $S_B(t)$, for *any* patient in the treatment group is $\pi \exp(-\lambda_B^1 t) + (1-\pi)\exp(-\lambda_B^0 t)$. This general formulation can allow us to simulate both the proportional hazards and non-proportional hazards setting. For the proportional hazards setting, it suffices to set either $\pi = 0$ or 1 so that the simulated survival distribution will always come from one of the pair of mixtures. Extension to mixtures of Weibull distributions are described in Appendix F.3 for more flexible survival curves.

We simulated 10,000 survival curves as shown in Figure 7.1 where I corresponds to the stochastically ordered survival curves without true crossings, and II for crossing survival curves. The hazard functions for both simulated scenarios cross approximately at year 0.5. Furthermore, in our simulated setting, although our true survival curves for I are stochastically ordered over the period of 5 years, the estimated survival curves will appear to cross with a probability arbitrarily close to 50%. At the calendar time of analyses, we censor the survival time for all subjects under the immediate accrual or uniform accrual setting. We compute summary statistics based on the 10,000 simulation: the total number of events, number of events in each treatment group, the estimated hazard ratio based on the logrank statistic with treatment group B as the reference group ($\text{HR}_{\text{Ref:B}}$), the estimated restricted mean survival for each treatment group, and the estimated survival probability at each calendar time. In addition, we compute the total number of events, number of events in each treatment group after this prespecified time of crossing, $\tau_0 = 2$. We then compute the number of survival curves that are observed to have crossed by τ_0, i.e., $\widehat{S}_A(t) > \widehat{S}_B(t)$ for $t > \tau_0$.

We can further compute the following test statistics at the end of the trial: the logrank statistics at time 5 (LR), the Nelson-Aalen at time 5 (NA), the restricted mean statistics at time 5 (RMS), the Nelson-Aalen at τ_0 based on all the data collected up to time 5 ($\text{NA}(\tau_0, t)$), the left truncated logrank statistic restricted to data after τ_0 ($\text{LR}(\tau_0, t)$), the linear composite statistics (OLS), and the quadratic statistic (Quad).

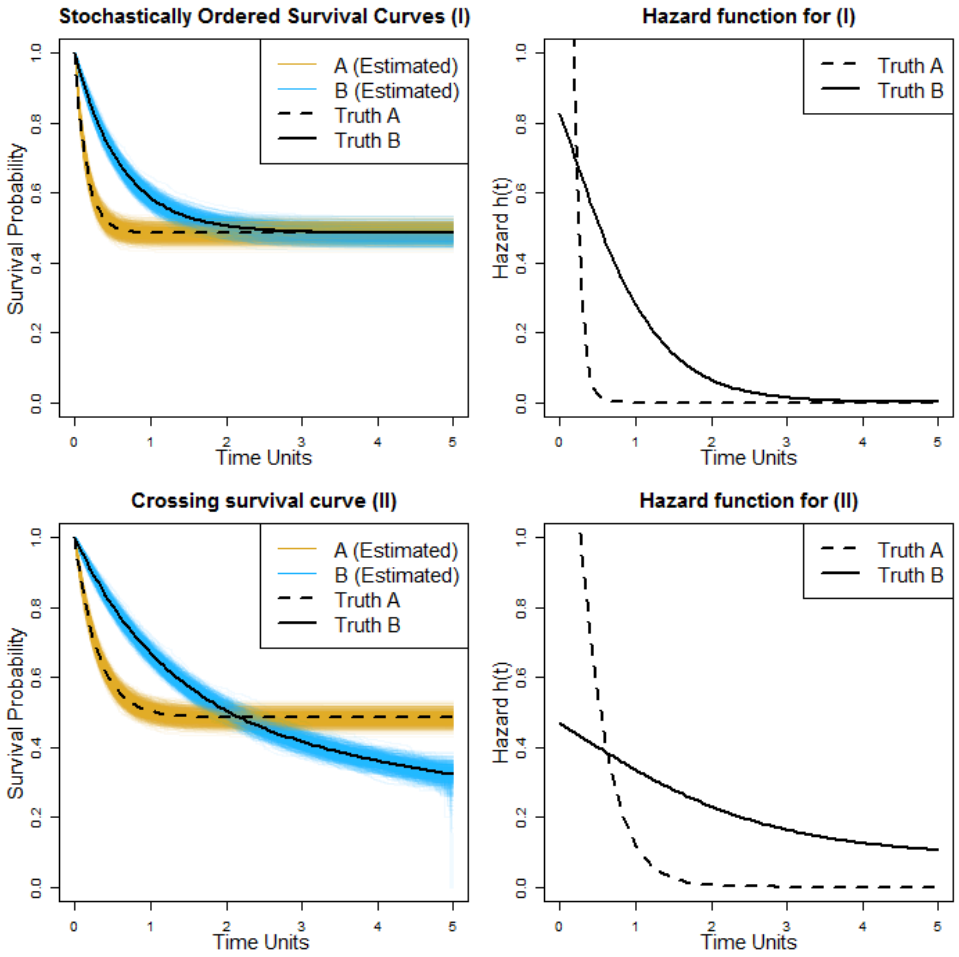

Figure 7.1: (I) Simulated scenario under the setting where our survival curves are stochastically ordered without true crossings over the first five years. However, spurious crossings are observed based on a random sample of 1000 simulations. The hazard ratio are $(\Lambda^1, \Lambda^0) = (4, 0.01)$. The hazard rate for each treatment group are $(\lambda_B^1, \lambda_B^0) = \left(\frac{-\log(0.04)}{2}, \frac{-\log(1-5e^{-13})}{5}\right)$ with respective mixing probability $\{\pi_1, \pi_0\} = \{0.515, 0.485\}$. (II) Simulated scenario under the setting where our survival curves are truly crossing within the first five years. The corresponding hazard functions for the simulated survival curves are presented in the right column.

7.3.2 Simulation Results for Stochastically Ordered, Crossing Hazards Survival Curves

We describe results for immediate accrual setting for setting (I) to investigate how the test statistics behave when there is only administrative censoring at time of analysis. Additional results for stochastically ordered with crossing hazards with uniform accrual, and simulation settings for crossing survival curves (both censored and uncensored) have also been investigated and are in Appendix F.2.

7.3.2.1 Descriptive Statistics

Under immediate accrual (based on 10,000 simulations each with 1000 subjects per treatment group), on average, a total of 1030 events are observed by year 5 with an equal number of events seen in either treatment group A or B (Table 7.2). By analysis time 2, there are more events observed in treatment group A as a consequence of the high hazard we imposed at earlier survival times.

Based on our simulations, we observed that 17.2% of the time, a crossing in survival curves is observed at time 2 such that treatment A has an estimated survival probability better than treatment B, i.e., $\widehat{S}_A(2) > \widehat{S}_B(2)$ and $\widehat{S}_A(1) < \widehat{S}_B(1)$. This probability of crossing increases to 48.4% by analysis time 5 . In addition, the estimated survival curves are approximately equal by $t = 2$.

Based on our simulated setup, all the events that occurred after the time of crossing now accumulate in treatment B group as a consequence of its higher hazard. Despite that, our estimated hazard ratio consistently concludes that B is the better treatment at each calendar time of analyses. On average, the difference in survival probability(group B vs group A) at time 2 decreases from 0.021 to 0.0002 by calendar time 5, favoring group B earlier on with this advantage to diminish at later calendar time. However, the use of the 5-year restricted mean statistics pointed to sufficiently large survival benefit in terms of 0.22 years saved by being in the treatment A group. This difference (~ 0.24) is consistently observed across all

the calendar time.

Table 7.2: Summary statistics based on 10,000 simulations under immediate accrual where (@) survival curves are stochastically ordered without true crossings over the first five years with treatment group B being the preferred treatment in terms of survival probability at all time. Descriptives are presented in the format mean (standard deviation).

	Stochastically ordered survival curves[@]				
	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Total number of events	926 (22)	1010 (22)	1026 (22)	1030 (22)	1030 (22)
Events (B vs A)	412 vs 514	495 vs 515	511 vs 515	514 vs 515	515 vs 515
Number of events $\geq \tau_0$	-	-	17 (4)	20 (4)	20 (5)
Events (B vs A)	-	-	17 vs 0	20 vs 0	20 vs 0
$\text{HR}_{\text{Ref:B}}(\pm \text{logSD})$	1.58 (0.066)	1.30 (0.064)	1.26 (0.063)	1.25 (0.063)	1.25 (0.063)
$\text{RMS}_A(t)^{\ddagger}$	0.44 (0.010)	0.93 (0.026)	1.41 (0.042)	1.90 (0.058)	2.38 (0.074)
$\text{RMS}_B(t)^{\ddagger}$	0.59 (0.008)	1.15 (0.022)	1.65 (0.036)	2.14 (0.051)	2.62 (0.066)
$\widehat{S}_A(t)$	0.4856 (0.02)	0.4848 (0.02)	0.4848 (0.02)	0.4848 (0.02)	0.4848 (0.02)
$\widehat{S}_B(t)$	0.5880 (0.02)	0.5055 (0.02)	0.4890 (0.02)	0.4856 (0.02)	0.4850 (0.02)
% of $\widehat{S}_A(t) > \widehat{S}_B(t)^{\dagger}$	-	17.2	41.4	47.2	48.4

†: Percentage of times a crossing is observed.
‡: The restricted mean statistic is truncated to 3 months just prior to the calendar time t.

7.3.2.2 Test Statistics

It is also important to point out the lack of power in the use of the Nelson-Aalen test to detect a 0.02 difference in survival probability at this prespecified crossing time. Based on our significant results obtained at time 5, we stratify these results at time 5 based on those survival curves that have crossed by time 2 vs those that have not crossed at time 2. In this case, 17.2% of them have crossed by time 2. Among those that crossed at time 2, only 1.38% of them conclude that treatment A is statistically significantly better than treatment B (24 in favor of A out of 1723 crossing in Table 7.3). Furthermore, among those survival curves in which a crossing is not observed at $\tau_0 = 2$, the Nelson-Aalen test ($\text{NA}(\tau_0, 2)$) concludes that treatment B is statistically significantly better than treatment A 18.9% of the time (1564 out of 8277). This indicates that the Nelson-Aalen test does not have high power to detect a difference in survival at time of prespecified crossing.

Table 7.3: Table of the statistically significant results at level $\alpha = 2.5\%$ based on 10,000 simulations under the setting where survival curves are stochastically ordered without true crossings over the first five years under immediate accrual of subjects. The columns, A and B, indicate the total number of times the test statistic concludes the trial in favor of the treatment arm. Note that analyses are conducted on the calendar time.

	Analyzed at time 5			Crossing at time 2				Crossing by time 5			
		Overall sig in favor of		Crossing $\widehat{S}_A(2) > \widehat{S}_B(2)$ n= 1723		No Crossing $\widehat{S}_A(2) < \widehat{S}_B(2)$ n= 8277		Crossing $\widehat{S}_A(5) > \widehat{S}_B(5)$ n= 4836		No Crossing $\widehat{S}_A(5) < \widehat{S}_B(5)$ n= 5158	
Statistic	Sig	A	B	A	B	A	B	A	B	A	B
Z_{LR}	9403	0	9403	0	1127	0	8276	0	4239	0	5164
Z_{NA}	520	279	241	279	0	0	241	279	0	0	241
Z_{RMS}	6816	0	6816	0	0	0	6816	0	1653	0	5163
$Z_{\text{NA}}(\tau_0, t)$	1588	24	1564	24	0	0	1564	24	0	0	1564
$Z_{\text{LR}}(\tau_0, t)$	10000	10000	0	1723	0	8277	0	4836	0	5164	0
Z_{OLS}	7394	7394	0	1723	0	5671	0	4832	0	2562	0
Z^{**}_{Quad}	9994	9994		1723		8271		4836		5158	

Crossing at time 2 is defined as $\widehat{S}_A(2) > \widehat{S}_B(2)$ and $\widehat{S}_A(1) < \widehat{S}_B(1)$.
Crossing at time 5 is defined as $\widehat{S}_A(5) > \widehat{S}_B(5)$ and $\widehat{S}_A(1) < \widehat{S}_B(1)$.
**: Direction of the quadratic test is selected based on OLS's direction.

Using a 1-sided level $\alpha = 0.025$, our overall unweighted log-rank test detects a statistically significant difference in survival between the two treatment groups 94.4% of the time and concludes in favor of treatment B. The average hazard ratio comparing A with respect to B is 1.25 ($\pm \log$ SD $= 0.063$). However, the use of the (weighted) log rank test that is restricted to the "clinically meaningful" time interval, τ_0 and $t = 5$, detects a significant difference 100% of the time. A naïve user using the results based on this truncated logrank test may interpret the number of events arising from treatment B to perhaps conclude that treatment A past τ_0 is better than treatment B!

7.3.3 Interpretation in Terms of Preferred Treatment

Suppose we naïvely assume that for all the test statistics described in Table 7.3, the direction of the test statistics indicates the preferred treatment. In other words, the log-rank test statistic (LR) and the restricted mean statistics (RMS) both conclude with probability of

94.0% and 68.2% that treatment group B is the preferred treatment with lower average hazard and higher average years of life saved respectively. The Nelson Aalen test statistic concludes 2.4% of the time that there is a difference in survival probability with B being the preferred treatment. The survival probability of being on treatment B is on average higher than the survival probability when randomized to treatment group A.

Both the linear and quadratic combination tests reject the composite null hypothesis with high statistical significance. However, the directionality of the quadratic test is lacking but rejects the composite null hypothesis almost 100% of the time based on level $\alpha = 0.05$, with a crossing in estimated survival curves observed approximately 50% of the time. Note that while the linear combination test does not have such high probability of rejecting the null composite hypothesis, the direction of the test statistics concludes that treatment group A has lower hazard as compared to treatment group B.

When we further examine the number of statistically significant survival curves that crossed by year 5, all 48.36% of the observed crossings by year 5 conclude A as the preferred treatment. This is despite the fact that only 5.8% (279/4836) of them were significantly in favor of A based on Nelson-Aalen test statistic conducted at year 2. By construction, as a consequence of this spurious crossing, the OLS statistic significantly rejects the (composite) null hypothesis almost 100% of the time, making a Type 1 error of at least $\approx 48\%$ of the time.

7.3.4 Issues with Lack of Guidance for Clinicians

If the Nelson-Aalen test statistic at $\tau_0 = 2$ is not statistically significant, a naïve researcher using the composite statistics might erroneously assume that $S_A(2) = S_B(2)$. However, because at time 5, $\widehat{S}_A(5) > \widehat{S}_B(5)$ and with $Z_{LR}(2)$ being highly statistically significant, the researcher may assume that $\lambda_A(t) < \lambda_B(t)$ when $\widehat{S}_A(2) > \widehat{S}_B(2)$ since the crossing has already occurred by time 5. Hence, by concluding that the long term survival at time of analysis, $t = 5$, is better for treatment A rather than for B.

However, on the other hand, if $Z_{NA}(2)$ is statistically significant in favor of treatment B

but that the weighted log rank statistic $Z_{LR}(2,5)$ concludes in favor of B, the naïve researcher might be smart enough to realize that the $Z_{LR}(2,5)$ is non-diagnostic. For anyone using the above statistics, a naïve researcher will make a type I error of $\approx 49\%$ of the time based on what we observed in our simulations. In observing an extremely low hazard for treatment group A past τ_0, it is incorrect to infer and extrapolate that this continuing high hazard ratio (comparing B relative to A) would necessarily lead to crossing survival over the time frame of interest.

7.4 Impact of Censoring on the Treatment Effect (Finite Follow-up)

With sequential sampling, interim analyses are conducted when part of the survival distribution is obtained and/or in presence of incomplete patient accrual. We investigate the impact of censoring on the estimates of the treatment effect in the finite follow-up setting for the various time to event analysis method described in Chapter 4. We assume patients are administratively censored at the time of interim analyses.

Characterizing this time varying treatment effect under the setting of non PH is more challenging when the different test statistics are estimating different quantities across time. For example, when using the logrank statistic, this estimate of the treatment effect as represented using the hazard ratio is no longer interpretable since we are averaging over risk sets that change over time. When using other test statistics that may be capturing other aspects of the contrasts of functionals, then the estimates of the treatment effect are no longer comparable.

To mitigate this problem, we consider a naïve approach in the context of sequential sampling and provide more details in section 7.4.1. In summary, with sequential sampling, for some local alternative δ, at some analyses time t, these Z statistics on the "standardized scale" are asymptotically $\mathcal{N}(\delta, 1)$. Since the planning of sequential rules can be planned on various scales that are 1-1 functions of each other as described in section 2.2, it makes natural sense to characterize this time varying treatment effect based on the "standardized

alternative" as represented on the Z statistics scale under both proportional hazards and non proportional hazards settings.

7.4.1 Standardized Alternatives

Consider the sample size formulation in section 2.1. Our sample size for each group can be determined based on some given level α, and statistical power β, to discriminate between $\mathbb{H}_0 : \theta \leq \theta_0$ vs the alternative of interest $\mathbb{H}_A : \theta \geq \theta_A$ using the general formula

$$N = \frac{(z_{1-\alpha} + z_\beta)^2 V}{(\theta_{\text{Alt}} - \theta_0)^2}$$

where z_p denotes p^{th} quantile of a standard normal distribution for $p \in (0, 1)$, and V is the variance contributed by a single sampling unit. Note that

$$\left(\frac{\theta_{\text{Alt}} - \theta_0}{\sqrt{V/N}} \right)^2 = (z_{1-\alpha} + z_\beta)^2$$

The LHS can be interpreted as our usual standardized Z statistic which is asymptotically distributed with $\delta = z_{1-\alpha} + z_\beta$ and variance 1.

Table 7.4: Table of standardized alternatives δ for various β while holding fixed $\alpha = 0.025$ ($z_{1-\alpha} = 1.96$), $N = 1$, and $V = 1$.

z_β	δ
-1.96	0
0	1.96
1.28	3.24
1.96	3.92

In particular, we may make the following interpretation. When holding fixed $N = 1 = V$ and $\alpha = 0.025$, then $Z = 3.24 = \delta_{\alpha\beta}$. We may interpret that a constant treatment effect

across time will tend to attain this power of β as we accrue more statistical information across time. We will use the Z statistic to describe our standardized alternatives for the various test statistics computed so as to characterize the time varying alternative. Some values of Z_β and the standardized alternative δ are as shown in Table 7.4.

7.4.2 Description of Simulation Setup

In this section, we regard these test statistics, namely, Nelson-Aalen test ($Z_{\mathrm{NA}}(t)$), logrank test ($Z_{\mathrm{LR}}(t)$), and the restricted mean statistics (Z_{RMS}), as commonly used statistics. Additionally, we describe the composite statistics based on Logan to consist of the following: Nelson-Aalen test at time of crossing ($Z_{\mathrm{NA}}(\tau_0, t)$), truncated logrank statistic after anticipated time of crossing ($Z_{\mathrm{LR}}(\tau_0, t)$), linear composite statistic (Z_{OLS}), and the quadratic statistic (Z_{Quad}) We prespecify $\tau_0 = 2$ to denote the anticipated time of crossing survival curves for Logan's statistics.

We parameterize positive values of the Z statistic to be consistent with the treatment being superior over the placebo at the interim analysis of time 1. For the restricted mean statistic, we computed the standardized difference in the area under the survival curve for the experimental treatment arm with respect to the placebo arm and define our support from 0 up to $t - 3$. We take the square root of the χ^2 statistics to represent our "standardized alternatives" for the quadratic composite statistics. We set the critical value as $\Phi^{-1}(1-\alpha)$, with $\alpha = 0.025$ for comparison with all test statistics except for the quadratic statistics where we use the square root of $\chi^2_{2,1-2\alpha}$. All standardized alternatives Z are averaged over 10,000 simulations.

We presume accrual of subjects uniformly ($q = 1$) over various accrual periods ($A = 2, 3, 4$) to characterize the degree of censoring that will affect these standardized alternatives based on equation 6.1. The immediate entry setting is used to characterize the setting with only administrative censoring at time of analyses. We analyze the data at calendar time t at 1, 2, 2.75, 3.5, 4.25, and 5 to describe the trends of the standardized alternatives across time.

7.4.3 Proportional Hazards Alternatives

Without loss of generality, we simulate true survival distributions from a Weibull distribution with common shape parameter of 0.5 for both groups, and "mean" parameter corresponding to 5.467002 and 11.48289 for the placebo (black) and treatment group (blue) respectively (left column of Figure 7.2). 10,000 simulations were performed under these proportional hazards alternatives and we computed the mean of all the Z statistics at each interim analysis and show the estimates on the right column of Figure 7.2. They are then plotted over time where the x-axis corresponds to the timing of the interim analysis. The thick solid black lines on the right corresponds to the immediate entry scenario, with increasingly lighter shades of blue (from dark blue to light blue) to represent increasing accrual over a longer period of time (from 2 to 4). The thin red lines corresponds to the critical value based on $\sqrt{\chi^2_{2,1-\alpha}} = 2.447747$ (dashed), $\Phi^{-1}(1-\alpha/2)$ (dash-dotted-dash) and the x-axis (dotted).

In the absence of censoring, under proportional hazards, each of the commonly used test statistics are consistently estimating the same ordering of survival curves over time. The average of the estimated Z statistics for all three commonly used test statistics are consistently favoring treatment over placebo and estimating the same functional of survival curves across analyses time. Logan's statistics also consistently favor the treatment over placebo. However, with the left truncated logrank statistic, this average of the $\widehat{Z}$ quantity is weaker as a consequence of the truncation.

In presence of censoring as reflected by staggered accrual, each of the commonly used statistics is estimating the same quantity with more variability. At early interim analyses, these estimates are often attenuated away from the solid line. The Nelson-Aalen estimator behaves differently under staggered entry and the estimates do not converge to the solid lines even at the end of the analyses, indicating that censoring can impact the average estimate of the results with Z_{NA}. However, the general behavior is consistent in terms of the ordering of the survival curves, reassuring us that under the PH setting, comparison of survival curves via any of the methods (including the composite statistics) will yield similar conclusions,

with differences in overall power depending on the efficiency of the test statistics under PH.

The OLS statistic, a weighted average of the Nelson-Aalen at time of crossing and the left truncated log rank statistic, has the average of the estimated Z being partway in between the Nelson-Aalen at time of crossing and truncated log rank test. The precision of these estimates of OLS is dependent on the sample size of the study. The OLS appears to be consistent in terms of estimating alternatives under PH which comprises of the $NA(\tau_0, t)$ and the left truncated logrank statistic. In comparison with the commonly used statistics, the mean $\widehat{Z}$ is also estimating the same functional consistently under PH.

The information growth for the various test statistics is presented in Table 7.5. In general, with censoring, the rate of information growth is expected to decrease when we extend the calendar time of accrual. Under the proportional hazards setting, this amount of information growth for the logrank test statistic is relatively close to the average of the number of events divided by the total average at calendar time 5. Under the censored setting, this rate of information growth decreases as we extend the accrual period. Consequently, the maximum statistical information also decreases.

Consider the information growth of the Nelson-Aalen test restricted to time τ_0 and the truncated log-rank test statistics after time τ_0 that are each calibrated to its own respective maximum statistical information on the "score" scale. The Nelson-Aalen test, $\mathrm{NA}(\tau_0, t)$, typically obtain its maximum statistical information much earlier relative to time 5. With censoring, this can more or less control this information growth but requires the accrual to be sufficiently long to obtain complete characterization by time 5. The $\mathrm{LR}(\tau_0, t)$ behaves differently in terms of magnitude and the growth of statistical information. With censoring, this drastically affects the total number of events coming in after τ_0 that can be used to estimate the statistical information, because events prior to τ_0 are not counted towards the test statistic. When this accumulation of events after τ_0 is slow, the information growth is thus slower and possesses more variability. As such, when all the data is analyzed by time 5, the truncated logrank statistic may not have attained the maximum statistical information, i.e., participants may not have complete followup between τ_0 and 5. This can lead to differ-

ent amounts of information contributed by each component of the test statistic, leading to difficulties in calibrating the individual components when considering weighted versions of the information growth of these test statistics as described in Logan and Mo [2015].

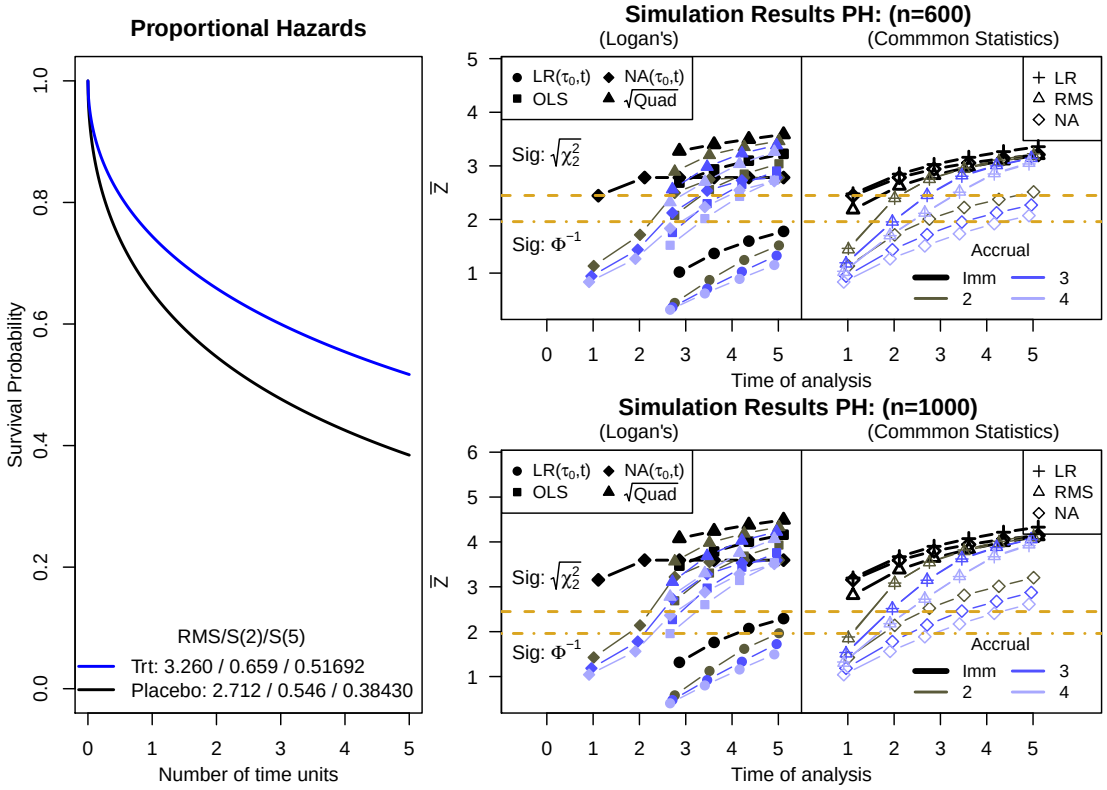

Figure 7.2: Survival curves and plot of standardized alternatives for proportional hazards survival curves under various accrual patterns and interim analyses for $n = 600$ and 1000. The average standardized alternatives $\widehat{Z}$ are consistently positive (respectively in quadrant I) for the commonly used or composite statistics.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table 7.5: Average information growth under proportional hazards for the various test statistics and the average number of events at each calendar time. The information growth conducted on the calendar time is affected by censoring even for the log rank statistics under constant treatment effect across time. The maximum statistical information is affected by censoring. Because the Nelson-Aalen test at $\tau_0 = 2$ generally obtain all statistical information when accrual is complete at the time of crossing. Since this depends on the accrual patterns, we see that this information growth is different with respect to the logrank statistic.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	180.97	238.31	268.29	292.27	312.35	329.55	
	2.00	63.19	169.12	222.58	257.04	283.50	305.09	
	3.00	42.12	112.74	176.24	230.88	264.16	289.61	
	4.00	31.59	84.65	132.24	184.83	237.28	270.21	
Events (τ_0, t)	Imm			29.98	53.96	74.04	91.24	
	2.00			5.83	21.76	45.19	66.78	
	3.00			3.90	14.47	30.59	51.30	
	4.00			2.93	10.87	22.92	38.47	
Z_{LR}	Imm	55.18	72.57	81.62	88.84	94.86	100.00	81.85
	2.00	20.73	55.59	73.13	84.39	93.00	100.00	75.84
	3.00	14.51	38.97	60.96	79.84	91.28	100.00	72.02
	4.00	11.63	31.31	48.95	68.45	87.86	100.00	67.21
$Z_{\text{LR}}(\tau_0, t)$	Imm			33.00	59.31	81.26	100.00	22.45
	2.00			8.67	32.56	67.69	100.00	16.44
	3.00			7.50	28.12	59.58	100.00	12.62
	4.00			7.47	28.11	59.49	100.00	9.45
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	224.12
	2.00		43.17	80.83	97.60	100.00	100.00	224.12
	3.00		31.00	59.09	83.41	96.41	100.00	224.12
	4.00		25.79	46.81	68.20	87.94	100.00	213.32

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic.
Imm refers to immediate accrual.

7.4.4 Non Proportional Hazards with Stochastically Ordered Survival Curves

Previously, in section 7.3, we evaluated the composite statistics in the fixed sample setting under stochastically ordered, crossing hazards scenario when patients are accrued immediately. We now describe the impact of censoring on the estimated standardized alternatives as well as the information growth in presence of this time varying treatment effect. Other scenarios are investigated in Appendix F.3.2 where we (1) varied this probability of survival at analyses time 5 from 90% (Figure F.6) to 10% (Figure F.7), (2) survival curves that diverge (Figure F.11), and (3) survival curves where this long term benefit diminishes at different survival times (Figure F.8, F.9, and F.10). Of interest here is the setting when this long term benefit diminishes.

In the absence of censoring (as defined by immediate accrual of subjects), the commonly used test statistics are now estimating different standardized alternatives over time (Figure 7.3). The overall LR test Z_{LR} and the restricted mean statistic Z_{RMS} both account for the entire history of the survival curves and are "consistent" in concluding the experimental treatment arm to be better relative to the placebo arm. Even though the PH assumption is violated under the weak null and under alternatives using the LR test, the stochastically ordered survival curves allow the LR test to consistently estimate the standardized alternative and to conclude that the experimental treatment to be better. However, in presence of censoring, the magnitude of the $\widehat{Z}_{\text{LR}}$ is attenuated. This is similar for $\widehat{Z}_{\text{RMS}}$ and can have implications when we later implement sequential monitoring.

On the other hand, the Nelson Aalen test, $\widehat{Z}_{\text{NA}}$, is targeted towards identifying whether there is a difference in survival across different analyses time. Thus, if the long term treatment effect was of interest, then the overall Nelson-Aalen test at time 5 (as shown by the unshaded diamonds in the plot titled "Common Statistics") will be the right test to use since this standardized alternative is close to 0 and should ideally be rejecting the null hypothesis close to level α. The choice of which test statistics to use then depends on the scientific interest and whether the survival at time 5 is clinically relevant or the entire survival experience is

more relevant.

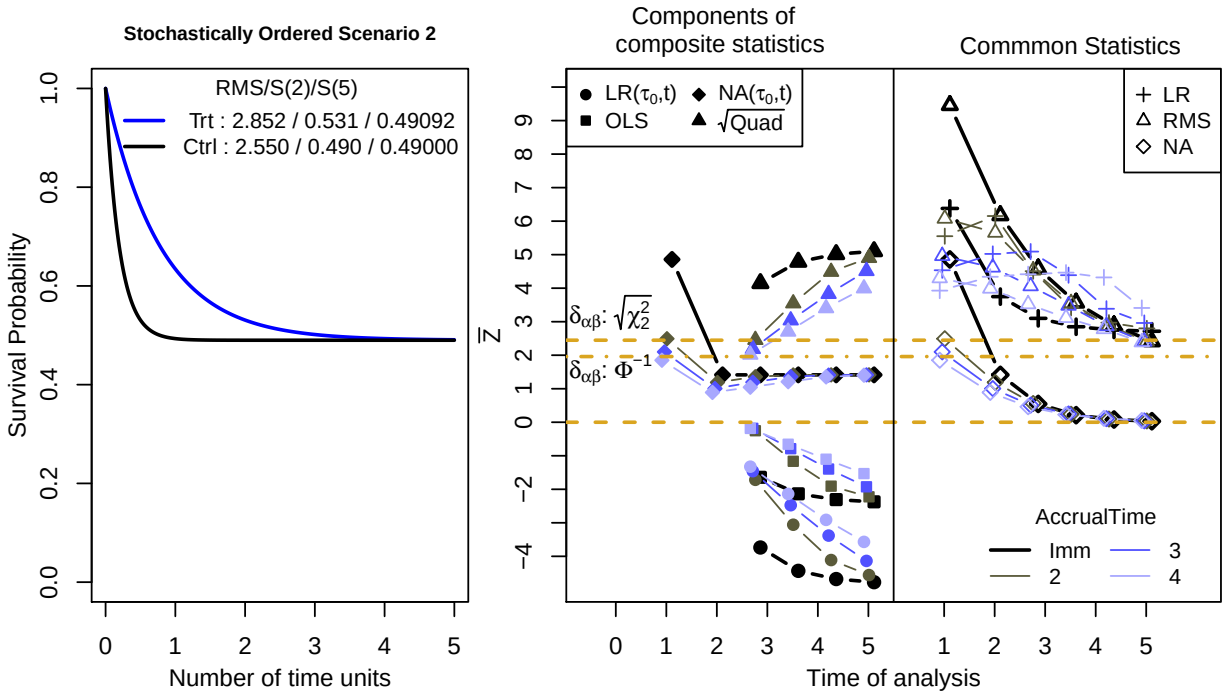

Figure 7.3: Standardized alternative for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 2) under various accrual patterns and different interim analyses. Survival curves for scenario 2 under stochastic ordering with roughly 50% probability of survival. The combination of alternatives for $\text{NA}(\tau_0, t)$ and $\text{LR}(\tau_0, t)$ resides in quadrant IV and describes conflicting conclusions prior to crossing and after crossing. The net effect for the linear composite statistic concludes that placebo is better than the treatment despite treatment being "better" at all times from 0 to 5 relative to placebo.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
$\text{LR}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
$\text{NA}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

We no longer observe the consistency when using the components of the composite statistics as seen under the proportional hazards setting. In particular, the standardized alternative for the truncated version of the log-rank test statistics is estimating a quantity that places preferential treatment on the placebo rather than the treatment. This relatively large imbalance in number of events creates a large, negative standardized alternative using $\widehat{Z}_{\mathrm{LR}}(\tau_0, t)$, hence ordering our outcome incorrectly when we place 0 weights prior to crossing.

The Nelson-Aalen test restricted to time of crossing ($\widehat{Z}_{\mathrm{NA}}(\tau_0, t)$) has low power to conclude a statistically significant difference in the majority of the time as reflected by being partway between 0 and $\Phi^{-1}(0.975)$. Even though this standardized alternative is in the right direction, the overwhelmingly negative average estimate of the $\widehat{Z}_{\mathrm{LR}}(\tau_0, t)$ now weights the Z statistic to favor the placebo as the preferential treatment with sufficiently high probability. The quadratic version of the test fares "slightly better" by taking the sum of the squares but ignores the direction of these misleading estimates. However, the result is having high probability of perhaps claiming statistical significance that a crossing has occurred even though such is not the case, with the added problem of lacking interpretability.

With censoring, the maximum information growth for the various test statistics are affected. In particular, when accrual is conducted over 4 time units, the $Z_{\mathrm{NA}}(\tau_0, 2)$ has accumulated at least 30% of the maximum statistical information. Because we place 0 weights on the truncated logrank statistic $Z_{\mathrm{LR}}(\tau_0, 2)$, we can only measure this statistical information at time 2.75. By time 2.75, the $Z_{\mathrm{NA}}(\tau_0, 2.75)$ has accumulated $> 50\%$ of its maximum statistical information while the $Z_{\mathrm{LR}}(\tau_0, 2.75)$ only obtain 12% of its maximum statistical information.

Suppose we choose a one-sided symmetric OBF boundary with equally spaced analyses to be conducted at calendar time 2.75, 3.5, 4.25, and 5. At 2.75 where we are thought to be relatively conservative, the $Z_{\mathrm{NA}}(\tau_0, 2.75)$ has accumulated sufficient statistical information to at least reject the futility boundary when there is harm in the use of the treatment. On the other hand, the $Z_{\mathrm{LR}}(\tau_0, 2.75)$ has only accumulated 12% of the statistical information and is thus relatively conservative in making a decision in favor of harm of efficacy. This partitioning of the parameter space by the composite statistics creates a "discontinuity" in

the information growth, giving an additional difficulty in deciding whether to combine the information growth equally or in some efficient weighted version to truly reflect the same degree of early "conservatism" as the monitoring rule.

Table 7.6: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 2) for the various test statistics under patterns of accrual and different interim analyses. The information growth conducted on the calendar time is affected by censoring even for the log rank statistics under constant treatment effect across time. The maximum statistical information is affected by censoring. Because the Nelson-Aalen test at $\tau_0 = 2$ generally obtain all statistical information when accrual is complete at the time of crossing. Since this depends on the accrual patterns, we see that this information growth is different with respect to the logrank statistic.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	523.28	587.31	602.39	608.22	610.47	611.36	
	2.00	189.08	470.02	567.88	595.03	605.39	609.37	
	3.00	125.97	313.36	462.33	568.20	595.51	605.58	
	4.00	94.39	235.17	346.78	460.33	563.60	594.69	
Events (τ_0, t)	Imm			15.08	20.91	23.16	24.05	
	2.00			3.27	10.19	18.08	22.06	
	3.00			2.19	6.77	12.34	18.27	
	4.00			1.65	5.10	9.23	13.68	
Z_{LR}	Imm	85.47	96.01	98.51	99.48	99.85	100.00	150.52
	2.00	31.01	77.11	93.11	97.61	99.34	100.00	150.03
	3.00	20.76	51.69	76.31	93.78	98.31	100.00	149.08
	4.00	15.82	39.49	58.27	77.39	94.77	100.00	146.36
$Z_{\text{LR}}(\tau_0, t)$	Imm			62.71	86.95	96.32	100.00	6.00
	2.00			14.74	46.09	81.92	100.00	5.51
	3.00			11.84	36.91	67.46	100.00	4.56
	4.00			11.88	37.09	67.41	100.00	3.41
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	312.74
	2.00		57.60	90.86	99.17	100.00	100.00	312.74
	3.00		39.45	69.41	91.40	98.63	100.00	312.74
	4.00		30.83	53.19	73.59	92.32	100.00	306.49

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic.
Imm refers to immediate accrual.

In general, under our stochastically ordered, non proportional hazards survival setting,

commonly used test statistics will order our standardized alternatives consistently with the survival curves. Our standardized alternatives tend to remain consistently in quadrant I across different interim analysis. In other words, these standardized alternatives do not switch over from the positive to the negative quadrant. This is consistent even when subjected to different accrual patterns.

Logan's composite statistics, however, are characterized as having $\delta_1 \geq 0$ and $\delta_2 < 0$, thus belonging to other quadrants instead of I. The inconsistency between the direction of δ_1 and δ_2 is partly due to weighting of the hazards prior to time of crossing as 0 for the truncated log rank test. Without knowing the history of the number at risk, the truncated log rank test orders the hazards in this partial space as if the survival curves are evaluated with followup starting from time τ_0.

7.4.5 Non Proportional Hazards with Crossing Survival Curves

In this section, we characterize the time varying treatment effect in the setting of non proportional hazards settings with crossing survival curves based on Figure 7.4. Other crossing survival curves were considered in Appendix F.3.3 and their behavior is similar so long as there is some large non-negligible difference in survival by the end of the study (Figure F.12 and F.13). By construction, the difference in the area under the survival curves by the end of the study in all these scenarios are negligible and 0.

The commonly used statistics all exhibit properties of switching conclusions of the preferred treatment (as exhibited by the change in sign of the standardized alternative) over the course of time. For example, with censoring via patient accrual, the conclusions based on the overall log rank test are further dependent on when the crossing of survival takes place. When crossing happens later (after τ_0), further accentuated by a slow and long accrual, the alternative of the overall logrank test attenuates less rapidly to 0, indicating that this change of sign is reflected much later in the trial (may happen outside the calendar time of the study and unless we ensure that everyone has complete followup for 5 time units). In contrast, with rapid accrual, the alternative converges more rapidly to alternatives as observed un-

der immediate entry. Additionally, only the immediate setting captures this change in sign much earlier than time 5. Under long accrual times, the pattern of changing quadrants is not obvious.

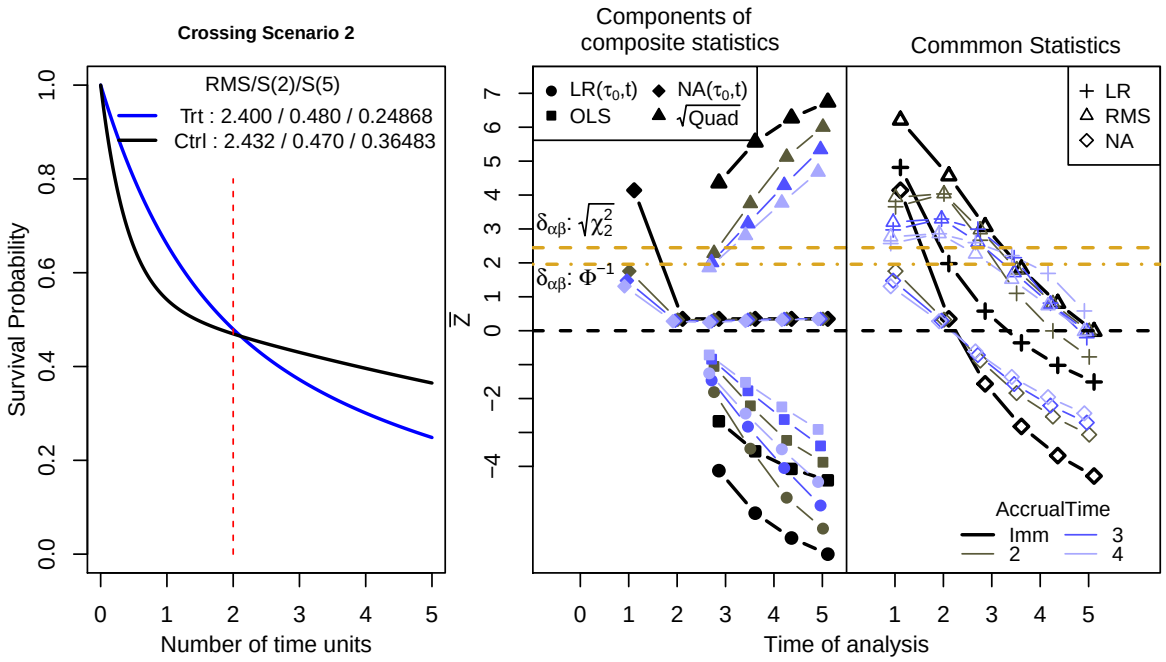

Figure 7.4: Standardized alternatives for various test statistics when we have crossing survival curves (Crossing scenario 2) under various accrual patterns and different interim analyses. Other variations on when the survival curves cross are in Appendix F.3.2: prior to τ_0, or after τ_0. Here, the survival curves cross at approximately τ_0. The combination of composite alternatives changes from Quadrant III (-,-), Quadrant III/IV(0, -), Quadrant IV (+,-) with the net result providing conclusion of preferring the placebo over treatment. The commonly used statistics is seen to have a changing alternative that switches from positive to negative.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

The Nelson-Aalen test statistic captures this change point quite accurately since it directly measures the difference in survival probability. In presence of censoring, the average estimated Z is weakened towards 0. The restricted mean statistics, on the other hand, are seen to decrease from an extremely large and positive alternative to almost 0, providing consistent standardized alternatives with our simulation setup in which this difference in area under the curve is negligible by time 5.

Table 7.7: Average information growth for crossing survival curves (Crossing Scenario 2) for the various test statistics under patterns of accrual and different interim analyses. The statistical information when analyzed on the calendar time is affected by censoring even for the log rank statistics under constant treatment effect across time.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (τ_0, t)	Imm			69.10	122.00	165.04	201.39	
	2.00			13.66	49.89	102.27	149.26	
	3.00			9.11	33.27	69.28	115.19	
	4.00			6.82	24.94	51.98	86.37	
Events (t)	Imm	477.78	630.45	699.56	752.45	795.50	831.85	
	2.00	148.85	430.62	587.56	672.88	732.72	779.71	
	3.00	99.18	287.23	453.80	604.48	687.90	745.65	
	4.00	74.39	215.29	340.42	476.59	615.26	700.01	
Z_{LR}	Imm	57.50	75.90	84.25	90.61	95.73	100.00	206.38
	2.00	19.08	55.23	75.36	86.33	94.02	100.00	193.65
	3.00	13.27	38.49	60.85	81.07	92.27	100.00	185.20
	4.00	10.59	30.71	48.60	68.08	87.91	100.00	173.84
$Z_{\text{LR}}(\tau_0, t)$	Imm			34.67	61.04	82.29	100.00	49.74
	2.00			9.15	33.52	68.69	100.00	37.01
	3.00			7.87	28.90	60.21	100.00	28.56
	4.00			7.82	28.86	60.22	100.00	21.39
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	271.62
	2.00		35.60	79.05	97.34	100.00	100.00	271.62
	3.00		25.23	56.30	81.76	96.03	100.00	271.62
	4.00		20.94	44.66	66.30	86.53	100.00	257.24

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic.
Imm refers to immediate accrual.

The estimates provided by the composite statistics behave similarly under the crossing

survival, crossing hazards setting. The linear composite statistic is now averaging out the magnitude and direction of the estimated Z from the individual components. The quadratic statistic on the other hand has high probability of rejecting the null hypothesis with a larger magnitude.

7.4.6 Summary

In general, our simulation results suggest that unless we have proportional hazards alternatives, the different test statistics considered can lead to different conclusions on the preferred treatment. Under non proportional hazards, when we have stochastically ordered, crossing hazards survival curves, the behaviors of these commonly used test statistics are generally consistent in presence of censoring and can identify the preferred treatment. When we have crossing survival with crossing hazards at some point in time, all test statistics are essentially estimating different quantities across interim analyses, selecting different treatments across analyses time. This effect is exaggerated when there is further censoring as introduced by accrual of subjects.

The use of composite statistics, and their individual components, are observed to perform poorly in terms of being able to identify the preferred treatment when we have stochastic ordering of survival curves. Particularly, the truncated logrank statistic, by placing 0 weights prior to time of crossing, can no longer order the outcome space correctly with the ranked failure times. This at times creates the illusion of identifying the placebo arm as preferred when logically it would be less desirable. When the Nelson-Aalen statistics conducted at the time of prespecified crossing do not have strong standardized alternatives to compensate for this apparent behavior seen in the truncated logrank statistic, the conclusion of the composite statistics will be weighted heavily by this direction of the standardized alternative based on the truncated log rank test.

In the clinical setting, if we were to apply Logan et al's test statistics to pick the preferential treatment, then under our setting of non proportional hazards with stochastic ordering, there is high probability of identifying the wrong treatment based on the estimated alter-

native. This can be as high as the probability of observing a spurious crossing as seen in the fixed sample setting. In our exploration using the standardized alternatives, the bizarre behavior of the truncated log rank statistic (almost) always orders the treatment effect incorrectly under stochastic ordering where the presumed crossing is unreal. When a huge treatment benefit is observed prior to time of crossing, their test statistic often ignores these differences when ordering the outcome space after crossing, i.e., the preferential treatment after averaging the effect over time.

In this section, many of these non proportional hazards survival curves examined can be considered a version of some weak null hypothesis. For example, in crossing survival curves, the timing at which the analysis is conducted presents different conclusions depending on the choice of the test statistics used. Because all these test statistics provide different interpretations of the preferred treatment, they thus may not consistently be testing the same average null hypothesis. This differential preference of what may or may not be clinically or statistically relevant at the timing of the analysis is of concern with testing the weak null hypothesis. Here, we may be making a Type 1 error if we incorrectly reject the null hypothesis when this rejection does not correspond to making the right clinical decision in favor of the preferred treatment.

7.5 Sequential Planning: Concerns and Considerations

As seen in section 7.4, in the presence of time varying treatment effect, censoring as induced by interim analysis/incomplete accrual of subjects can weight the survival curves differently across time, changing the relative importance of the treatment effect. We discuss some of the potential decisions the DMCs can make based on the use of summary statistics presented to summarize the differences in survival as described in section 4.6 to address the primary question of identifying the better treatment quantified by possibly 5 year survival. We now tie in the concepts of what we investigated in section 7.3 and 7.4 to now address more general issues concerned with planning a GSD in the time to event setting where we may be concerned with the weak null hypothesis as seen in the previous section.

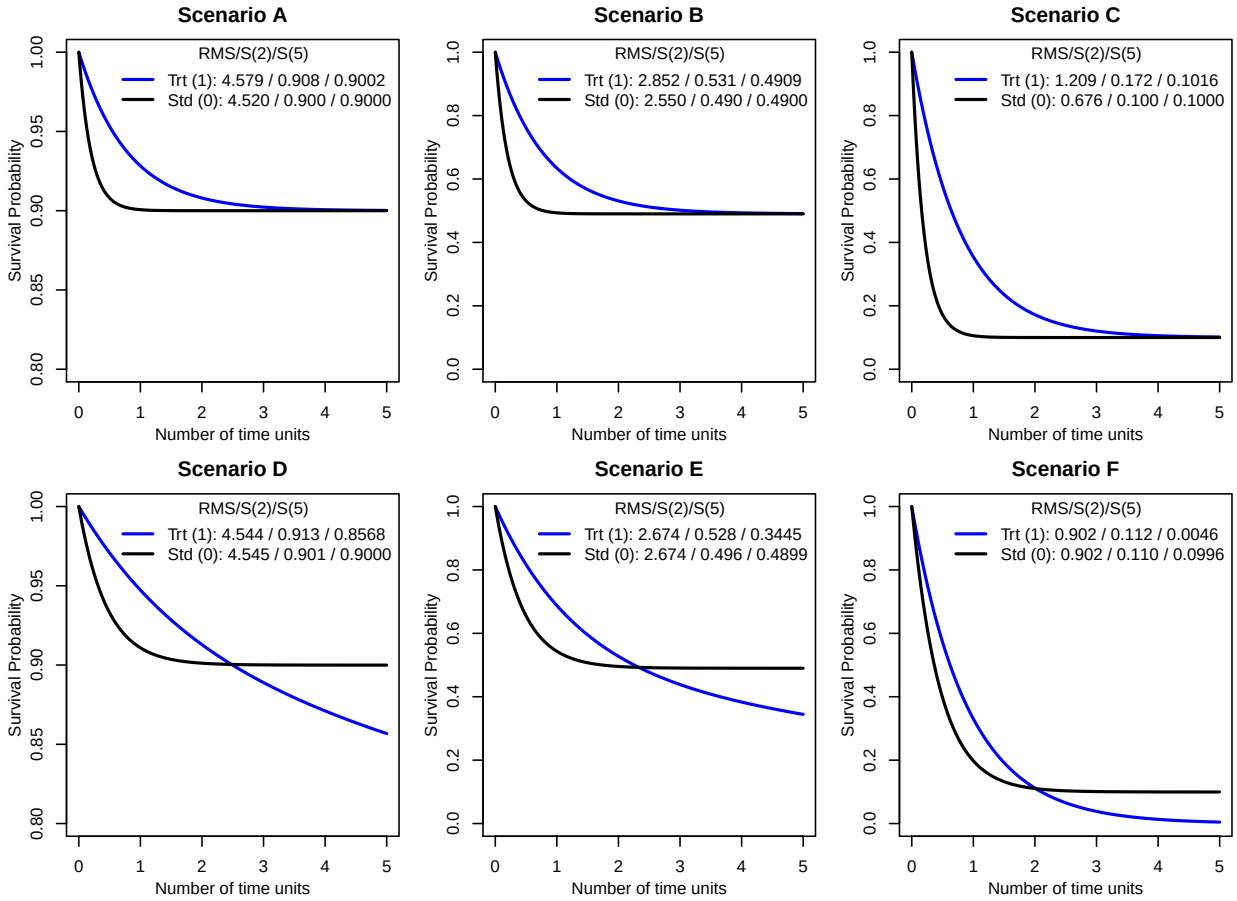

Figure 7.5: Survival curves (Truth) exhibiting crossing hazards: Stochastically ordered (Scenarios A-C) with high probability of crossing vs partially ordered survival curves with probability one of crossing (Scenarios D-F). In the bottom row, the survival curves are partially ordered, with the treatment group exhibiting better survival probability at pre-specified time of crossing ($\tau_0 = 2$) relative to the standard of care (SOC) group. However, the restricted mean survival (RMS) time is similar for the treatment and standard of care group by year 5.

Scenarios A-F are simulated from on a mixture of exponential distributions based on section 7.3.1. The degree of mixing was chosen for π=(0.1, 0.51, 0.9), the respective rates, hazard ratios for the mixtures are provided in Table 7.8. The survival curves for the various weak null alternatives are shown in Figure 7.5.

Each of the scenarios in A-C represents stochastically ordered, non PH alternatives over the first 5 years with high probability of the survival curves exhibiting crossing by year 5. By way of construction, the hazards for these three scenarios all cross much earlier than 2 years with the survival curves observed to cross spuriously by 2 years. Note that each of these scenarios represents different clinical settings where judgement of long term benefit is dependent on the disease setting and the probability for which this long term benefit happens. For example, in childhood cancer, when we are interested in cure rates, then scenario B and C may not be acceptable as compared to scenario A due to the high probability of survival past time 5. Additionally, it may be the case that maybe A is not reasonable if this long term survival benefit does not persist beyond year 5 and thus may be clinically less relevant to the child in the trial.

Table 7.8: Table of parameter values chosen for each of the scenarios shown in Figure 7.5 where Scenario A - C represent stochastic ordering with crossing hazards over the first 5 years where Group 0 is the preferred treatment while D-F denotes crossing survival curves.

Scene	e_{Trt}, e_{Std}	Λ_{Trt}, Λ_{Std}	π	N	$\text{RMS}_{\text{Trt}}, \text{RMS}_{\text{Std}}$	$\text{S}_{\text{Trt}}(2)$, $\text{S}_{\text{Std}}(2)$	$\text{S}_{\text{Trt}}(5)$, $\text{S}_{\text{Std}}(5)$
A	0.08, 5e-13	4, 0.01	0.1	5400	4.58, 4.52	0.9080, 0.900004	0.900181, 0.9
B	0.08, 5e-13	4, 0.01	0.51	1058	2.83, 2.54	0.5257, 0.490012	0.490661, 0.49
C	0.08, 5e-13	4, 0.01	0.90	600	1.21, 0.66	0.1720, 0.100037	0.101629, 0.1
D	0.33, 0.055	4, 0.001	0.1	5400	4.54, 4.54	0.913, 0.901	0.857, 0.9
E	0.210, 0.318	2.892, 5e-4	0.51	1058	2.67, 2.67	0.528, 0.496	0.340, 0.49
F	0.121, 0.99985	2.1, 5e-4	0.90	600	0.90, 0.90	0.112, 0.111	0.005, 0.1

Each of the scenario in D-F, on the other hand, was constructed such that the survival curves cross definitely after 2 years in addition to crossing hazards. By construction, our outcome space or the probability of survival for each treatment group changes depending

on the time of analyses. Treatment group (1) appear to be preferred prior to year 2 on the survival scale with a higher survival probability over standard of care (0). Standard of care (0) has a higher survival probability when followed longer up to year 5 over treatment group (1). Additionally, by construction, the true restricted mean comparing difference in the areas between the curves when analyzed at year 5 is 0. In other words, the absolute number of years of life saved comparing treatment group 1 with respect to treatment group 0 by year 5 is thus equivalent.

When we are more concerned about the long term benefit of the treatment effect, and less about the early differences, then any of the above scenarios are potentially testing a weak null hypothesis. As such, the conclusions may differ depending on when the timings of these analyses are restricted to. In A-C, they all represent the weak null setting when we consider the use of Nelson-Aalen to test at specific calendar time corresponding to year 4 and beyond. Other test statistics may potentially conclude benefit in favor of group 0 when evaluation is made over the support of the survival curves. In D-F, application of the restricted mean statistic makes different conclusions at say 2 year vs 4 year. At year 5, the restricted mean statistic will conclude no benefit between treatment and placebo. Similarly, other test statistics may make different conclusions across the calendar time.

For discussion purposes, we describe results for scenarios B and E for the rest of this section whereby this time of crossing is suspected to take place by 2 years. Scenario B only has $< 0.1\%$ difference in survival probability at time 5. Scenario E only has small differences in survival probability at time 2 but the preferred treatment changes at time 5. Both scenarios exhibit relatively large early differences that may or may not matter clinically depending on the clinical setting.

We discuss some of the statistical considerations when choosing to either pick a fixed sample design, and/or a sequential monitoring rule to guide statistical monitoring when the objective is concerned with identifying the better treatment with long term survival as defined possibly by 5 years. We consider some potential dilemmas that may force a DMC to make a recommendation to stop the study early, such as in the CHER trial (Appendix A),

or in a bone marrow transplant setting, using various summary statistics and/or measures at interim analyses. We consider how to possibly incorporate this early difference when choosing sequential rules.

The ultimate goal is to pick the treatment with the long term benefit while ignoring any early differences in survival, then it may be possible that a fixed sample design can do as well. However, when we are anticipating a scenario E to possibly happen, then it may no longer be ethical to wait until year 5 to terminate the trial when this crossing indicates unacceptable harm to participants on the trial. Then, our monitoring rules must be selected such that at early interim analyses, the probability of stopping and rejecting the null is not so large enough that we cannot identify clinically meaningful differences when the full time period of interest has been observed. While other scenarios are not discussed here, the concept can be applied when planning a trial to guide monitoring. We next describe some of the calibration approaches to handling potential non independent increments structure.

7.5.1 Calibration Approaches for GSDs to Preserve the Overall Type 1 Error

As a consequence of constructing the sequential rules to address the scientific considerations, and assess the potential considerations in these weak null setting, our control of the strong null may not be preserved. We first describe some of the approaches that can be used to recalibrate our boundaries to ensure control of the strong null Type 1 error rate. Additionally, some of these approaches may be used with non-independent increment statistics, such as, in the setting of the restricted mean survival or Nelson-Aalen test when the support of the statistics change.

For any design scenario of interest, we may consider the following boundaries construction

1. Using specified information growth based on the Z statistic scale.
    - Naïve (equal) information growth at each calendar time of analyses
    - True information growth under the null hypothesis at each calendar time of analyses
2. Constrained boundaries approach based on the Z statistic scale.

- Continuously revise our monitoring boundaries based on current estimated statistical information
- Revise our boundaries at the end of the trial

3. Error spending approach

    - Specify our monitoring boundaries based on the amount of error one wants to spend on each analysis.

In (1), test statistics without independent increment property are presumed to assume monitoring boundaries that are equally spaced at calendar time or assume statistical information based on the logrank test statistic. These test statistics include the Nelson-Aalen test at year 5, restricted mean statistic at year 5, quadratic test statistic. To apply the approach of constrained boundaries (2), one requires approximately independent increment property. Test statistics such as Nelson-Aalen, restricted mean statistic, both performed at time of analyses, have support that changes across analyses time and no longer possess the property of independent increments. The quadratic test statistic also does not possess such a property. Thus, in (2), we restricted ourselves to the overall log rank test statistic (both with 3 or 4 interim analyses), Nelson Aalen at year 2, and the appropriately calibrated OLS.

7.5.1.1 Pre-specified information growth based on the Z statistic scale

The first approach assumes boundaries that are fixed on the Z statistic scale based on either naïve information growth or the true (average) information estimated under the null hypothesis. In other words, at each interim analysis, we use the pre-specified boundaries as defined on the Z-statistic scale when we monitor our interim analysis on the calendar time scale. Thus, we can,

- Construct boundaries based on the specified information growth/fraction either presuming naïve information growth or the true information growth
- At each interim analysis, we compare the Z-statistic at a interim analysis and compare that to our monitoring boundaries on the Z scale. If the estimated Z statistic is either above the upper boundary or below the lower boundary, we stop the trial and conclude the trial based on the direction of the test statistic.

- If not, we continue the trial to the next interim analysis. We do not modify our monitoring boundaries at any stage.

The merits of this first approach is the simplicity and can be applied when test statistics are not computed with a fixed support that can lead to correlated increments. We can directly calibrate any test statistic to ensure control of our overall Type 1 error under the strong null hypothesis and apply comparison of boundaries directly on the Z statistic scale. However, the consequence of such a simple approach is that this naïve approach may not be appropriate when the statistical information across interim analysis are non monotonic. When the statistical information fraction is different from planned, one may overspend or underspend the appropriate amount of Type 1 error at any interim analysis. This leads us to consider a slightly more flexible approach to calibrate the monitoring boundaries over the course of the trial.

7.5.1.2 Constrained Boundaries Approach Based on the Z Statistic Scale

The constrained boundaries approach described in Burington and Emerson [2003] provides a flexible way to revise the boundaries when dealing with changes in observed statistical information from the planned levels of statistical information. The appeal of the constrained boundaries approach stems from the fact that one may presume the pre-specified maximum statistical information at design stage and allows updating of this maximum statistical information when there is accrued data becomes more reliable. One can either (a) calibrate the boundary at every stage, or (b) only recalibrate the boundaries at the end of the trial or when the statistical information is used up.

The constrained boundaries approach (a) is as follows:

- Construct the monitoring boundaries ($\{a_j, d_j\}_{j=1}^{J}$ where a_j represent the futility boundary and d_j the efficacy boundary) using the specified information growth (Π_j) with the maximum statistical information defined as $\mathcal{I}_J$. This can be performed either using the naïve or true information growth for the test statistic.

- At the first interim analysis, if the observed statistical information, $\widehat{\mathcal{I}}_1$, does not match the planned statistical information, $\mathcal{I}_1$, we revise $\{a_1, d_1\}$ to $\{a_1^*, d_1^*\}$ based on the current estimated statistical information, $\widehat{\mathcal{I}}_1$, while holding the future monitoring boundaries $\{a_j, d_j\}_{j=2}^J$ fixed. We then compare the current z statistic with the revised boundaries, $\{a_1^*, d_1^*\}$, to determine whether the interim z statistic has crossed either the revised efficacy or futility boundaries. If not, we continue the trial.

- At the j^{th} interim analysis, if the observed statistical information, $\widehat{\mathcal{I}}_j$ does not match the planned statistical information, $\widehat{\mathcal{I}}_j$, we use the sequence of monitoring boundaries before j that are possibly revised, $\{a_l^*, d_l^*\}_{l=1}^{j-1}$, based on the prior sequence of statistical information $\{\widehat{\mathcal{I}}_l\}_{l=1}^{j-1}$, we hold the future boundaries fixed and constrain on the current statistical information, $\widehat{\mathcal{I}}_j$, to obtain $\{a_j^*, d_j^*\}$. We then compare the current Z_j statistic with the revised boundary, Z_j^*, to determine whether the interim Z_j statistic has crossed either the revised efficacy or futility boundary.

- At the final analysis, if $\widehat{\mathcal{I}}_J < \mathcal{I}_J$, we spend the remaining unused error using this estimate of the final statistical information, $\widehat{\mathcal{I}}_J$.

Alternatively, a simpler version of the constrained boundary approach (b) is as below:

- Construct the boundaries ($\{a_j, d_j\}_{j=1}^J$ where a_j represent the futility boundary and d_j the efficacy boundary) using the specified information growth (Π_j) with the maximum statistical information be $\mathcal{I}_J$. This can be performed either using the naïve or true information growth for the test statistic.

- At the j^{th} interim analysis, we presume the monitoring boundaries, $\{a_l, d_l\}_{l=1}^j$ as prespecified and compare the current Z statistic with these boundaries. We do not revise the boundaries according to the observed statistical information. We then decide whether the Z statistic has cross the efficacy/futility boundary. If not, we continue the trial.

- At the final analysis, we spend the remaining unused error by specifying our true observed sequence of statistical information, $\{\widehat{\mathcal{I}}_l\}_{l=1}^{J}$.

In situations when the test statistics have the independent increment property, either one of the above approaches can allow one to handle changes from the planned schedule of analyses. Operationally, it is possible for statistical information to "flow" backwards due to imprecision of the prior estimate of the statistical information. This can result when the statistical information across analyses present more variability at later interim analysis than the previous, leading to little changes in information fraction. This can happen when the interim analyses are conducted too close to each other on the information fraction scale. To circumvent such issues from occurring, it is possible for one to skip the current interim analysis and move this analysis to the next planned calendar time when this information fraction is close to 1. However, Proschan et al. [1992] has shown that even with common monitoring rules, revision of monitoring boundaries when the information growth is negligible can inflate the overall Type 1 error by up to 20% with conservative monitoring strategies such as OBF. Burington and Emerson [2003] suggested the approach of revising the previous estimate of the statistical information based on this updated knowledge of the test statistics.

7.5.1.3 Error spending approach

We describe the error spending approach in the more general fashion that can be used to handle mild deviations from the independent increments by re-calibrating our monitoring boundaries.

We may start off by planning the study using a pre-specified information growth (defined either based on the naïve or true) to obtain our monitoring boundaries on the Z scale. We then

1. Apply the sequential boundaries to the planned monitoring boundaries.

2. Compute the empirical overall Type 1 error rate under the strong null hypothesis.

3. If the overall Type 1 error rate is within $(\mathbb{F}^{-1}_{\text{Bin}(n,\alpha)}(0.025), \mathbb{F}^{-1}_{\text{Bin}(n,\alpha)}(0.975))$, we can keep our monitoring boundaries as pre-specified. If not, we can adjust the critical value at the final analysis such that it is fixed level α.

Typically, to employ the above approach it is useful to simulate sufficient data to obtain this revised boundaries and then test this revised boundaries on another simulated set of data to evaluate the behavior of this revised rule.

Alternatively, we can pre-specify the amount of rejection at each successive interim analysis. This concept is similar to the error spending approach. In order to obtain the sequential boundaries for this method, we can then

1. Specify the sequence of errors to be spent at each interim analysis at each information time such that $\sum_{j=1}^{J} \alpha_j = \alpha$.

2. At the first interim analysis, for $j = 1$, determine the empirical $\mathbb{F}^{-1}(\alpha_1/2, \text{lower})$ and $\mathbb{F}^{-1}(1 - \alpha_1/2, \text{upper})$. Set them to be either the lower or upper boundary.

3. At subsequent analysis, $j = 2, \cdots, J$, among the remaining simulations that are not statistically significant, determine the $\mathbb{F}^{-1}(\alpha_j/2, \text{lower})$ and $\mathbb{F}^{-1}(1 - \alpha_j/2, \text{upper})$ such that only α_j are statistically significant.

4. At the final analysis, determine the $\mathbb{F}^{-1}(\alpha_J/2, \text{lower})$ and $\mathbb{F}^{-1}(1 - \alpha_J/2, \text{upper})$ such that the remaining simulations are only significant for α_J of the n simulations.

The above approaches can be implemented even for test statistics with the usual independent increments. When simulation scenarios under the strong null gives rise to weak correlations across calendar time, we can typically apply the above rule to recalibrate our overall Type 1 error.

7.5.2 Fixed Sample Designs vs Group Sequential Designs

We present some summary statistics described yearly from year 1 to 5 that are typical of the setting when the DMC may convene to deliberate on the data obtained. The summary statistics in Table 7.10 are: the total number of events (nEv_k), total number of events past τ_0 ($\text{nEv}_k(\tau_0, t)$), restricted mean statistic truncated to 3 months prior to interim analysis

($\text{RMS}_k(t)$), average hazard ratio (unweighted), probability of survival for each treatment group ($S_k(t)$), and the proportion of times $S_{\text{Std}}(t) > S_{\text{Trt}}(t)$ for treatment group $k = \text{Trt}, \text{Std}$ where $\tau_0 = 2$ is again the pre-specified time of crossing.

We included a fixed sample design based on a two-sided level $\alpha/2 = 0.025$ test for comparison purposes. To compare how the decisions based on the monitoring boundaries can be used across various test statistics, we assume that our planned schedule of analyses are conducted at time 3, 4, and 5. We note that if other choices of test statistics were used, the interim analyses are typically conducted when accrual is incomplete. However, the composite statistics are not well behaved to make any decisions prior to the prespecified time of crossing. Thus, this makes decision making in terms of efficacy of the treatment relative to the standard of care difficult.

We present summary statistics for the following test statistics in Table 7.13: the log-rank test statistics ($Z_{\text{LR}}(t)$), the Nelson-Aalen test statistics ($Z_{\text{NA}}(t)$), the Nelson-Aalen test statistics restricted to time of crossing ($Z_{\text{NA}}(\tau_0, t)$, the restricted mean test statistics ($Z_{\text{RMS}}(t)$ used interchangeably with $Z_{\text{WKM}}(t)$), the linear combination test statistic ($Z_{\text{OLS}}(\tau_0, t)$, and the quadratic test statistic ($Z_{\text{Quad}}(\tau_0, t)$) where τ_0 denotes some prespecified time of potentially crossing survival curves. Note that the composite statistics ($Z_{\text{OLS}}(\tau_0, t)$ and $Z_{\text{Quad}}(\tau_0, t)$) cannot be summarized at $t = 1, 2$.

We have considered various GSDs ranging from

- *OBF(Equal)*: two-sided symmetric OBF with information fraction that is equally spaced on the calendar time,

- *OBF(Info)*: two-sided symmetric OBF with information growth calibrated to the test statistic of choice at each calendar time,

- *HP*: two sided symmetric Haybittle-Peto design with interim analysis requiring $p < 0.001$ for efficacy/futility,

- *Equal*: two-sided symmetric equal error design,

- *Poc*: two-sided symmetric Pocock with information fraction that is equally spaced analysis on the calendar time,

- *Poc(Info)*: two-sided symmetric Pocock with equally spaced analysis calibrated to the test statistics of choice at each calendar time.

Additionally, we note that for test statistics with independent increments across analyses time, we can incorporate the true information growth to get a more precise calibration of the monitoring boundaries.

In our exploration, we calibrated $Z_{\text{NA}}(t)$, $Z_{\text{RMS}}(t)$ to the information growth of $Z_{\text{LR}}(t)$. The $Z_{\text{OLS Fixed}}$, $(Z_{\text{LR}}(\tau_0, t) + Z_{\text{NA}}(\tau_0, t))/\sqrt{2}$, is assumed to use the OBF boundary with equal information fraction. Note that the $Z_{\text{OLS Fixed}}$ can be calibrated to allow for the property of independent increments as discussed in Logan and Mo [2015]. The first (denoted as $Z^{\text{N}}_{\text{OLS}}$) that we considered is the $Z^{\text{N}}_{\text{OLS}} = \mathcal{U}_{\text{N}}/\sqrt{\text{Var}(\mathcal{U}_{\text{N}})}$ which is the non-standardized version of the calibration that is a weighted average of the information growth for each of the two components. The second (denoted as $Z^{\text{S}}_{\text{OLS}}$) is calibrated such that each component of the composite statistic is weighted equally by summing up the information fraction for each component relative to its maximum statistical information at the end of the trial. In this case, by the final calendar time, the maximum statistical information based on $Z^{\text{S}}_{\text{OLS}} = \mathcal{U}_{\text{S}}/\sqrt{\text{Var}(\mathcal{U}_{\text{S}})}$ is always 2. For discussion, we focus on the monitoring boundaries *OBF(Equal)* and *OBF(Info)*.

We calibrated the boundaries to ensure a fixed two-sided level 0.05 test under the strong null hypothesis with the stopping probability for each interim analysis presented in Table 7.9. The respective application under scenario B is shown later in Table 7.11. Scenario E is calibrated by assuming a different set of strong null simulation. Note that it may often be useful to calibrate, in this case scenario E, to the same derived monitoring rule under the strong null as in Table 7.9. The disparity between *OBF* and *OBF(Info)* becomes apparent when we reject too frequently at some early interim analysis that may not balance the scientific goals of the study.

Table 7.9: Overall Type 1 error rate, probability of stopping at interim analyses, the respective conclusion for the two monitoring rules considered for the various test statistics evaluated under the strong null hypotheses. Monitoring rules are based on equally spaced information growth assumption with recalibration of the final monitoring boundary under the strong null to ensure fixed 5% type 1 error.

		$t=3$	$t=4$	$t=5$	Overall
		Std/Trt	Std/Trt	Std/Trt	Both/ Std/Trt
OBF (Equal)	Z_{LR}	0.01/0.02	0.59/0.62	1.77/1.99	5.00/2.37/2.63
	Z_{NA}	0.02/0.00	0.70/0.55	1.92/1.81	5.00/2.64/2.36
	Z_{RMS}	0.02/0.03	0.58/0.59	1.74/2.04	5.00/2.34/2.66
	$Z_{\text{NA}}(2,t)$	0.05/0.02	0.57/0.65	1.91/1.80	5.00/2.53/2.47
	$Z_{\text{OLS Fixed}}$	0.00/0.05	0.60/0.64	1.88/1.83	5.00/2.48/2.52
	$Z^{\text{S}}_{\text{OLS}}$	0.00/0.02	0.64/0.67	1.84/1.83	5.00/2.48/2.52
	Z_{Quad}	0.04/0.02	0.78/0.75	1.84/1.57	5.00/2.66/2.34
OBF (Info)	Z_{LR}	0.88/0.93	1.10/1.12	0.44/0.53	5.00/2.42/2.58
	Z_{NA}	0.94/0.73	1.16/1.27	0.46/0.44	5.00/2.56/2.44
	Z_{RMS}	0.90/0.89	1.07/1.16	0.43/0.55	5.00/2.40/2.60
	$Z_{\text{NA}}(2,t)$	0.66/0.58	1.27/1.32	0.60/0.57	5.00/2.53/2.47
	$Z_{\text{OLS Fixed}}$	0.07/0.17	0.81/0.90	1.61/1.44	5.00/2.49/2.51
	$Z^{\text{S}}_{\text{OLS}}$	0.04/0.12	0.82/0.89	1.64/1.49	5.00/2.50/2.50
	Z_{Quad}	0.16/0.08	1.23/0.99	1.41/1.13	5.00/2.80/2.20

7.5.2.1 Monitoring under the scenario with spurious crossing

We first describe the issues when this prespecified crossing is spurious as defined by Scenario B. Summary results are in Table 7.10. On average for scenario B, the average total number of events at the first interim analysis is 222 with an average of 35% of the events coming from treatment group, with the remaining to come from standard of care. We note that there is an excess of 77 deaths on the standard of care, with a hazard ratio of 2.29. The difference in probability of survival comparing the treatment and standard of care is at least 0.1487, the difference in area under the survival curves comparing the treatment and standard of care is 0.1421. There is a 2.3% chance that some of the survival curves may have crossed in B.

Table 7.10: Summary statistics based on 10,000 simulations under scenario B, E and the strong null setting comparing treatment (Trt) vs standard of care (Std).

		$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Null	No of events by t	156 (12)	457 (19)	801 (23)	1001 (23)	1057 (23)
	Events (Trt vs Std)	78 vs 78	229 vs 229	400 vs 400	500 vs 500	529 vs 529
	No of events in (τ_0,t)	0 (0)	0 (0)	12 (4)	37 (6)	64 (8)
	Events (Trt vs Std)	0 vs 0	0 vs 0	6 vs 6	18 vs 18	32 vs 32
	$HR_{\text{Ref: Trt}}$	1.00 (0.162)	1.00 (0.093)	1.00 (0.071)	1.00 (0.063)	1.00 (0.062)
	$RMS_{\text{Std}}(t)^{\ddagger}$	0.6145	1.217	1.739	2.237	2.73
	$RMS_{\text{Trt}}(t)^{\ddagger}$	0.6146	1.217	1.739	2.237	2.73
	$\widehat{S}_{\text{Std}}(t)$	0.6351	0.5311	0.5017	0.4931	0.4909
	$\widehat{S}_{\text{Trt}}(t)$	0.6352	0.531	0.5019	0.4931	0.4908
	% of $\widehat{S}_{\text{Std}}(t) > \widehat{S}_{\text{Trt}}(t)^{\dagger}$	50.0	49.9	49.7	50.2	50.5
Scenario B: Alt	No of events by t	222 (14)	553 (20)	904 (23)	1040 (23)	1068 (23)
	Events (Trt vs Std)	78 vs 145	229 vs 324	400 vs 504	500 vs 539	529 vs 540
	No of events in (τ_0,t)	0 (0)	0 (0)	6 (3)	18 (4)	32 (6)
	Events (Trt vs Std)	0 vs 0	0 vs 0	6 vs 0	18 vs 0	32 vs 0
	$HR_{\text{Ref: Trt}}$	2.29 (0.140)	1.77 (0.085)	1.57 (0.067)	1.35 (0.062)	1.27 (0.062)
	$RMS_{\text{Std}}(t)^{\ddagger}$	0.4659	0.9581	1.448	1.938	2.428
	$RMS_{\text{Trt}}(t)^{\ddagger}$	0.6146	1.217	1.739	2.237	2.73
	$\widehat{S}_{\text{Std}}(t)$	0.4931	0.4898	0.4899	0.4899	0.49
	$\widehat{S}_{\text{Trt}}(t)$	0.6352	0.531	0.5019	0.4931	0.4908
	% of $\widehat{S}_{\text{Std}}(t) > \widehat{S}_{\text{Trt}}(t)^{\dagger}$	2.3	13.4	31.7	44.5	48.6
Scenario E: Alt	No of events by t	170 (12)	484 (19)	846 (23)	1065 (23)	1156 (23)
	Events (Trt vs Std)	62 vs 109	203 vs 281	386 vs 460	534 vs 531	617 vs 539
	No of events in (τ_0,t)	0 (0)	0 (0)	18 (4)	62 (8)	122 (11)
	Events (Trt vs Std)	0 vs 0	0 vs 0	17 vs 1	59 vs 3	117 vs 5
	$HR_{\text{Ref: Trt}}$	2.01 (0.160)	1.61 (0.092)	1.38 (0.069)	1.15 (0.062)	1.00 (0.060)
	$RMS_{\text{Std}}(t)^{\ddagger}$	0.5516	1.079	1.573	2.063	2.553
	$RMS_{\text{Trt}}(t)^{\ddagger}$	0.6466	1.289	1.792	2.216	2.588
	$\widehat{S}_{\text{Std}}(t)$	0.5435	0.4956	0.4904	0.4899	0.4899
	$\widehat{S}_{\text{Trt}}(t)$	0.6884	0.5285	0.4394	0.3836	0.3448
	% of $\widehat{S}_{\text{Std}}(t) > \widehat{S}_{\text{Trt}}(t)^{\dagger}$	2.8	22.6	92.2	100.0	100.0

Descriptives are presented in the format mean (standard deviation).
†: Percentage of times a crossing is observed.
‡: The restricted mean statistic is truncated to 3 months just prior to the analyses time.

We see a similar situation at interim analyses at year 2. In B, we see that this excess death has accumulated to 95 deaths on the standard of care, a difference in area under the survival with extra 0.2589 years saved on the treatment group, and a difference in survival

probability of 4.1%.

Interim analyses conducted at time $t \geq 3$ have similar total number of events for the standard of care past the interim analysis conducted at $\tau_0 = 2$. As noted, the average number of events contributing to the log-rank statistic after the pre-specified τ_0 for scenario B all falls on the treatment arm. By the final analysis, on average, more than 45% of the survival curves have spuriously crossed, indicating that the standard of care has a better survival as compared to the treatment. As such, the probability that the DMC observed a crossing in survival curves by $t = 3, 4,$ and 5 would be at least 30%, 40%, and 45% on average respectively.

Without sequential monitoring at early interim analyses, the DMC have to make judgment based on the available data. This large treatment benefit is observed for both scenarios, presenting a challenging situation to the committee when the monitoring guidance precludes having an appropriate boundary for the early analyses. On the basis of the data presented, the DMC have to judge collectively on whether this excess of death arising on the standard of care arm is sufficient to warrant stopping the trial early when there appears to be this overwhelming benefit. Other summary measures are consistently pointing to early benefit with treatment over standard of care. The DMC may not call for stopping early possibly there may be random high bias, although with over 100 events, this may not be the case.

If accruing data demonstrates an even larger difference in survival, the DMC may potentially have to act on the results despite the objective of identifying a better long term treatment. This is true in exaggerated scenarios where the treatment group may, for example, have survival probability of 0.9 by time 2 relative to the standard of care, where this probability is 0.53. This sufficiently large difference in survival and high number of events falling on the treatment arm may no longer enable the trial to continue to achieve its objective for ethical reasons.

Table 7.11: Probability of rejecting the null hypothesis and the respective conclusion Type 1 error obtained for various methods of monitoring the various test statistics under different monitoring rules for Scenario B and E. Monitoring rules are either based on equally spaced information growth assumption with recalibration of the final monitoring boundary under the strong null to ensure fixed 5% type 1 error or based on the information growth of the test statistic.

			$t=3$ Std/Trt	$t=4$ Std/Trt	$t=5$ Std/Trt	Overall Both/Std/Trt
Scenario B	OBF	Z_{LR}	0.00/99.97	0.00/0.00	0.00/0.00	99.97/0.00/99.97
		Z_{NA}	0.00/0.33	0.31/1.03	1.63/1.70	5.00/1.94/3.06
		Z_{RMS}	0.00/95.42	0.00/1.07	0.00/0.00	96.49/0.00/96.49
		$Z_{\text{NA}}(2,t)$	0.00/3.64	0.00/22.96	0.00/18.87	45.47/0.00/45.47
		$Z_{\text{OLS Fixed}}$	0.00/0.28	6.23/0.00	69.98/0.00	76.49/76.21/0.28
		$Z^{\text{S}}_{\text{OLS}}$	0.01/0.00	15.63/0.00	60.23/0.00	75.87/75.87/0.00
		Z_{Quad}	6.85	92.59	0.56	100.00
	OBF (Info based)	Z_{LR}	0.00/100.00	0.00/0.00	0.00/0.00	100.00/0.00/100.00
		Z_{NA}	0.07/3.72	0.90/1.35	0.33/0.20	6.57/1.30/5.27
		Z_{RMS}	0.00/99.76	0.00/0.00	0.00/0.00	99.76/0.00/99.76
		$Z_{\text{NA}}(2,t)$	0.00/20.88	0.01/19.16	0.00/4.43	44.48/0.01/44.47
		$Z_{\text{OLS Fixed}}$	0.02/0.56	8.41/0.00	66.28/0.00	75.27/74.71/0.56
		$Z^{\text{S}}_{\text{OLS}}$	0.04/0.00	19.79/0.00	54.96/0.00	74.79/74.79/0.00
		Z_{Quad}	15.92	83.76	0.32	100.00
Scenario E	OBF	Z_{LR}	0.00/89.23	0.00/0.02	2.29/0.00	91.54/2.29/89.25
		Z_{NA}	0.32/ 0.00	84.48/0.00	15.08/0.00	99.88/99.88/0.00
		Z_{RMS}	0.00/71.79	0.00/0.23	0.88/0.00	72.90/0.88/72.02
		$Z_{\text{NA}}(2,t)$	0.00/1.29	0.01/14.01	0.05/15.33	30.69/0.06/30.63
		$Z_{\text{OLS Fixed}}$	0.14/0.01	75.06/0.00	24.79/0.00	100.00/99.99/0.01
		$Z^{\text{S}}_{\text{OLS}}$	2.37/0.00	94.52/0.00	3.11/0.00	100.00/100.00/0.00
		Z_{Quad}	47.48	52.52	0.00	100.00
	OBF (Info based)	Z_{LR}	0.00/98.29	0.00/0.00	0.94/0.00	99.23/0.94/98.29
		Z_{NA}	6.50/0.01	84.46/0.00	8.90/0.00	99.87/99.86/0.01
		Z_{RMS}	0.00/92.73	0.00/0.02	0.69/0.00	93.44/0.69/92.75
		$Z_{\text{NA}}(2,t)$	0.02/8.19	0.04/15.85	0.00/5.76	29.86/0.06/29.80
		$Z_{\text{OLS Fixed}}$	0.21/0.01	76.81/0.00	22.97/0.00	100.00/99.99/0.01
		$Z^{\text{S}}_{\text{OLS}}$	3.88/0.00	93.49/0.00	2.63/0.00	100.00/100.00/0.00
		Z_{Quad}	58.21	41.79	0.00	100.00

Z_{RMS} and Z_{NA} are calibrated based on the information growth of the logrank statistic.
$Z_{\text{OLS Fixed}}$ is calibrated to the information growth with the sum of equally weighted components from each test statistic.

For comparison, scenario E has summary measures similar to what was seen in scenario B, whereby in this scenario, the crossing is as hypothesized. Basically, at this point in time, when a judgement has to be passed on whether the trial should continue in favor of the treatment effect, neither B nor E may continue past time 2 if this difference in survival is clinically meaningful. However, stopping the trial at this time, again indicates that the scientific objective may not be addressed when we are interested in the long term benefit. This thus summarizes the conflicting goals and potential decisions the DMC may make.

7.5.2.2 Sequential monitoring plan in place

At interim analyses conducted at calendar time 3, we see a high probability of early stopping in favor of the treatment arm for both scenarios. In fact, the logrank and restricted mean statistic both identify the treatment as the better arm. The Nelson-Aalen, however, is not powered to detect a difference that is 0.02 and 0.05 for scenario B and E respectively. The restricted mean is larger on the treatment group for both scenarios. The DMC must then face a choice on deciding whether to stop the trial based on the choice of monitoring rule and the evidence provided so far.

The use of Logan's statistic is however inconsistent with what the data are presenting. Because of the dependence on this crossing being as anticipated and the Nelson-Aalen be powered adequately to tell this difference apart, we see that the two versions of the composite statistics are not making a recommendation for early stopping. However, the individual component of $\widehat{Z}_{\mathrm{NA}}(\tau_0, 3)$ is indicative of favoring the better treatment for the standard of care when considering the information based OBF monitoring. In contrast to E, this stopping probability falls short. The quadratic statistic is not useful here in identifying the better treatment strategy despite sufficiently high probability of crossing by time 3 for scenario E.

The overall probability of rejecting at time 5 is seen to be sufficiently high for scenario B where both logrank and restricted mean are identifying the better treatment as the treatment group over standard of care group. Even though the Nelson-Aalen does not have high power to detect the difference in survival probability, it is addressing the "right" question if

5 year survival is of clinical interest while early differences may be deemed to be less clinically important. Because this estimate is "essentially" disregarding all survival information earlier on, it would not identify potential treatment strategies as well in situations when, for instance, there is plausibility of a waning treatment effect in vaccine settings where a booster shot may be useful to take advantage of this sufficiently large effect early on. In such settings, sufficiently large early differences are of interest.

Table 7.12: Probability of rejecting the null hypothesis in favor of either treatment (Trt) or standard of care (Std) for scenario B, E based on a fixed sample design at time $t = 5$.

	Scenario B			Scenario E		
Stat	Overall	Std	Trt	Overall	Std	Trt
Z_{LR}	97.44	0.00	97.44	5.44	2.74	2.70
Z_{NA}	4.76	1.95	2.81	99.70	99.70	0.00
Z_{RMS}	88.95	0.00	88.95	7.00	0.90	6.10
$Z_{\text{NA}}(\tau_0, 5)$	25.88	0.02	25.86	17.08	0.04	17.04
$Z_{\text{OLS Fixed}}$	76.63	76.63	0.00	100	100	0.00
Z_{Quad}	100.00	-	-	100	-	-

Additionally, we contrast how the overall conclusions of the fixed sample design differs from the sequential monitoring. Under the fixed sample setting when there is spurious crossing:

- The Nelson Aalen (NA) rejects the null hypothesis at roughly 5%.
- The logrank and restricted mean statistic conclude with high probability in favor of the treatment.
- The linear composite statistic concludes with high probability in favor of the standard.
- The quadratic statistic concludes with 100% probability of rejecting the composite null hypothesis that a crossing occurs.

However, with sequential analyses,

- The Nelson Aalen (NA) rejects the null hypothesis at roughly 6% with slightly higher probability favoring the treatment.

- Both logrank and restricted mean statistics conclude with high probability in favor of the treatment by time 3.

- The use of the OLS that is calibrated to the information growth based on the sequential version of OLS^{S} has only high probability of stopping at the final analysis to reject the composite null of no crossing and identifying the standard of care as the preferred treatment.

- Use of the quadratic statistic stops early at time 4 with the conclusion that there is 100% probability of rejecting the composite null hypothesis.

In contrast to a fixed sample design, when there is truly a crossing such as in scenario E, the sequential analyses generally result in the same conclusions for the logrank statistic, and even higher probability of stopping with the linear composite statistics. The restricted mean, although has slightly lower probability, is overpowered at earlier analyses even with the assumption of naïve information growth. The Nelson-Aalen statistic, Z_{NA}, with sequential analyses is at least powered to ultimately detect a difference in E, that is consistent with results from the fixed sample design. The use of composite statistics do not always provide consistent conclusions with results from the other test statistics in scenario B, despite being similar in the setting of crossing survival. This indicates that while the composite statistics are "adequate" as promised to detect true crossing survival curves, many often, when this crossing is spurious in both fixed sample and sequential setting, the results are unreliable to enable guidance for statistical monitoring. The over reliance on using the p-value to guide stopping and the "less well-understood" aspect of how the composite statistics perform in a

variety of situations to provide conflicting results with other summary measures that make guidance difficult in practice.

7.5.2.3 Additional concerns with choice of boundaries

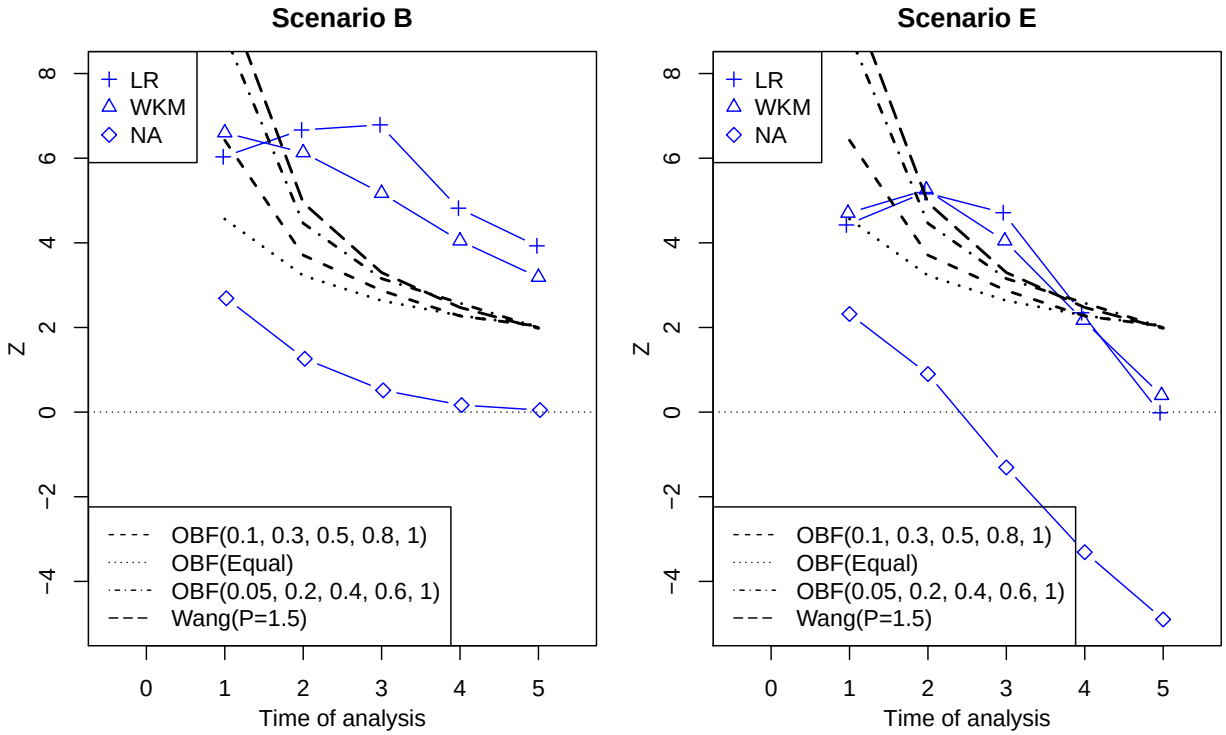

Figure 7.6: Standardized alternatives estimated for various test statistics at each interim analyses with several choices of monitoring boundaries to account for early differences that are not meaningful. Positive values of Z are consistent for the treatment being better than the standard while negative values of Z are consistent for standard of care being better. The black solid lines describe some potential sequential boundaries with varying degree of early conservatism. The crossing scenario (E) becomes problematic when these time varying estimates crosses the boundaries and then recrosses back later on.

The sequential monitoring plan can be chosen to take place at later analyses as what we just described as well as in Logan and Mo [2015]. It is often the case that data monitoring are conducted frequently to inspect the trial for any safety issues that might arise. Thus,

interim analyses may be conducted on a yearly basis even though a sequential rule is placed at later calendar time. However, during a DMC meeting, members of the committee are often presented efficacy data as well as safety data in order to be able to judge the trial and assess the benefit to risk of the treatment. When the sequential monitoring plan does not present a guidance to allow the DMC to judge the trial, they have to rely on other aspects of the data, such as summary statistics in order to assess this benefit to risk. However, poor choices of monitoring rules such as what we described earlier can lead to too frequent stopping at a calendar time that does not matter clinically when assessing long term benefit.

Suppose for now we may be able to make our first formal interim analysis with an appropriate monitoring rule based on previous section. Both the log rank and Nelson-Aalen test statistic have high probability of stopping by this time in favor of the treatment. However, by now, our survival curves may have crossed in scenario B and most often in E. Then, a difficulty arises as to whether the conclusions of the logrank and Nelson-Aalen are useful since there may be concern that this treatment effect past 3 years may no longer address the same clinical question. In such a setting, we see that the naïve choice of a monitoring rule that we presume is conservative may not be ruling out differences that matter clinically to us. In this setting, we see that the (average) standardized alternatives for these test statistics are sufficiently higher than that of our proposed boundaries as in Table 7.13. Consequently, if this difference does not balance the scientific considerations of the trial, then a more extreme rule should have been chosen to preclude rejecting the null hypothesis too often at early interim analyses.

Such a scenario is also observed at the second interim analyses. We see the problem of choosing a simple monitoring boundary that does not necessarily protect us under the weak null setting such as scenario B and E. Additionally, when attempting to calibrate the boundary with the true information growth, this makes the implicit assumption that the earlier information growth does not matter, particularly with the composite statistics. Incidentally, if long term benefit was of interest, then the Nelson-Aalen test would be sufficient to detect this long term benefit at some time point of interest since in E, the test has high power to

detect this difference. In scenario B when this difference in survival probability diminishes, then the test do not reject sufficiently often which is more appropriate than any of the other test statistics considered.

When we posit the possibility of crossing survival curves, then the decision is harder since we need to assess the relative benefit prior to crossing, the time of crossing as well as the probability that this crossing takes place as noted in earlier section. With these in mind, the restricted mean statistic may present a better candidate with the appropriate monitoring rule. In the fixed sample setting, neither treatment groups for scenario E presents an advantage as the late difference in area under the survival curves is negated by this early difference in area under the survival curves.

Figure 7.6 illustrates the time varying standardized alternatives for the various statistics. In particular, we present other plausible boundaries using a two-sided symmetric OBF designs that spanned across the graph with different levels of early conservatism. However, when the history of the survival is to be used, both the logrank and restricted mean statistics provide differing level of statistical evidence. We noted previously that with the sequential rules imposed, we were not able to control this probability of stopping at the first analyses. This is further illustrated in both Table 7.13 and Figure 7.6 where we see that the standardized alternatives are far away from 0 so that more conservative boundaries may need to be chosen. Below, we see some other potential rules that we have computed to reflect how early this level of conservatism needs to be.

1. *OBF(0.1, 0.3, 0.5, 0.8, 1)* : Two-sided symmetric OBF boundary with a total of 5 analyses conducted at information fraction of 0.1, 0.3, 0.5, 0.8, and 1.

2. *OBF(0.2, 0.4, 0.6, 0.8, 1)* : Two-sided symmetric OBF boundary with a total of 5 analyses conducted at equally spaced information fraction.

3. *OBF(0.05, 0.2, 0.4, 0.6, 1)* : Two-sided symmetric OBF boundary with a total of 5 analyses conducted at information fraction of 0.05, 0.2, 0.4, 0.6, and 1.

4. *Wang and Tsiatis P = 1.5* : Two-sided symmetric Wang and Tsiatis boundary ($P = 1.5$) with a total of 5 equally spaced analyses.

Table 7.13: Average of the Z statistics at each interim analysis under the weak null scenario of B and E that must be considered when wanting to exclude early differences that are not important. The monitoring boundaries on the Z-scale are also presented for a total of 3 analyses. Although these boundaries selected can be calibrated to control the overall Type 1 error rate, note that if early differences were not of concern, none of these boundaries are suitable since there is high probability the standardized alternatives will cross the boundaries (Table 7.11). As such, our level of early conservatism as represented via the calendar time is not maintained with the use of these common boundaries.

	Stat	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Scenario B	Z_{LR}	6.03	6.67	6.79	4.82	3.93
	Z_{NA}	2.69	1.26	0.51	0.17	0.05
	Z_{RMS}	6.60	6.13	5.17	4.05	3.19
	$Z_{\text{NA}}(\tau_0, t)$	2.69	1.26	1.67	1.84	1.87
	$Z_{\text{LR}}(\tau_0, t)$			-2.36	-4.14	-5.50
	$Z_{\text{OLS Fixed}}$			-0.48	-1.62	-2.57
	Z_{Quad}			9.65	21.75	35.00
Scenario E	Z_{LR}	4.42	5.19	4.71	2.35	-0.02
	Z_{NA}	2.32	0.90	1.23	1.42	1.46
	Z_{RMS}	4.70	5.25	4.05	2.17	0.40
	$Z_{\text{NA}}(\tau_0, t)$	2.32	0.90	1.23	1.42	1.46
	$Z_{\text{LR}}(\tau_0, t)$			-3.66	-7.20	-10.50
	$Z_{\text{OLS Fixed}}$			-1.72	-4.09	-6.39
	Z_{Quad}			16.41	55.36	113.80
Boundaries	*OBF(Equal)*			3.4711	2.4544	2.0040
	Haybittle-Peto			3.0902	3.0902	1.9704
	Equal Error			2.3941	2.2937	2.2002
	OBF(Info[Ref:LR])			2.3633	2.1139	2.0567

The last two boundaries can be chosen to preclude early termination of the study when there may be suspicion of crossing survival curves. Of note, each of the boundaries may address different levels of early conservatism but may not be best in the setting of E when

this time varying treatment effect can lead to crossing the boundaries and then crossing over again. In such settings, common sequential strategies as above may not be best suited for crossing survival without maintaining other scientific aspects of the trial. In other words, when we posit crossing survival curves, we are excluding the possibility that this crossing can be spurious. When choosing any sequential rules, we may be less concerned with scenario B when there may be no crossing over the calendar time we conduct the study.

Even though the boundaries can be selected to ensure early conservatism, the late crossing may not be adequately captured with these preferred rules. It may be the case whereby a more extreme version of the Haybittle-Peto boundary needs to be selected such that the boundary shape function are flat on the Z scale to ensure having captured sufficient long term data to establish the benefit and proper identification of the better long term treatment. However, such a monitoring rule is deemed to be less efficient than competing boundaries such as OBF or Pocock. Additionally, such choices also imply we are imposing high constant degree of conservatism to balance statistical goals, which may come with the price of not addressing competing constraints such as scientific or ethical goals.

7.6 Discussion

Sequential analyses of clinical trials are conducted for ethical and efficiency concerns with the goal of protecting the safety and well being of the patients on the trial. The DMCs are charged with many important ethical and often difficult scientific decisions related to protecting the integrity of the study as well as the patient's safety. Thus, the summary/test statistics that can elucidate understanding of any impending safety issues are of utmost important. Because many of these scientific decisions are complex, the use of simple statistics that enable understanding of the efficacy and effectiveness of the data accrued thus far far outweighs the use of complex statistics that do not address the scientific question in a reliable manner.

Potentially, when crossing survival are possible, we may have to consider the use of fixed sample strategies while putting additional considerations in how to judge a trial with the use

of the summary statistics. Kaplan Meier survival curves may enhance better understanding together with various summary statistics in order to make clear informed decisions to evaluate the strategy. Over-reliance purely on statistical significance obtained from sequential boundaries may not be the single statistic to provide reliable quantification of treatment benefit.

Chapter 8

Conclusions

In this dissertation, we have investigated the statistical issues associated with the application of adaptive designs in the time to event RCT setting. We do not regard each chapter to stand separately on its own when considering the issues that may arise during the planning of a clinical trial to determine whether a treatment/prevention strategy is worth pursuing. Instead, each of these chapters is focused on different issues that can arise during the design, conduct, and monitoring of any time to event clinical trial. We have attempted to separate out some of these issues that are currently lacking in the adaptive literature and conducted comprehensive evaluation of the potential benefit and risk when planning an adaptive interim analysis in the setting with delayed ascertainment of outcomes.

The setting with delayed ascertainment of outcomes presents a variety of challenges when there is an added dimensionality of time. Conventional adaptations made at the penultimate analysis in the immediate setting are no longer feasible logistically since these adaptations in presence of low event rate may be chosen late on the calendar time which are no longer considered late on the statistical information scale. Additionally, when there are logistical difficulties, the trial may need to be terminated earlier, thus lending itself the plausibility of adaptively stopping with a smaller sample size. In the immediate outcome settings, findings have not found adaptive designs to be markedly more efficient over competing group sequential designs. However, the statistical literature also has not been adequate to separate out efficiency issues, differentiating between good or bad flexible adaptive designs vs selecting the best time to make such an adaptation. Thus, we investigated the setting of finding the best flexible design while keeping the schedule fixed to determine whether such benefit can

be attained in the fixed sample setting with the aim of decreasing the final sample size. We found that when studies have to be terminated earlier, the best flexible adaptation tends to lead to significant loss of power over having prespecified these rules at design stage. And that when studies are enlarged, there is significant loss in power when such schedule of analyses are conducted late using the flexible adaptations. While gains in power are negligibly better at very earlier analyses, it also points out that under such circumstances, one should have started off designing the study with care, rather than choosing late adaptations.

Estimating the background rates in a time to event setting can be a difficult task. When an apparently "low background rate" arises during the conduct of the study, the options of either increasing accrual, terminating a study earlier than expected, and/or extending calendar time may well depend on whether this is a consequence of truly low event rate, and/or extreme treatment effect. In such settings, the utility of an adaptive strategy thus depends on how well the rule can distinguish this difference. We investigated the plausibility of comparing a fully blinded adaptive design based on a group sequential method vs an unblinded procedure that might be used to better distinguished between low event rates and/or extreme efficacy of treatment effect. When this low background rate is due to the presence of an extreme treatment effect, we found that fully blinded group sequential approaches are most often more than adequate. Furthermore, protecting the integrity of the trial without the additional complications of potential operational bias from third party to conduct the adaptive steps.

In limited settings when the treatment effect is moderately effective and that event rates are not sufficiently extreme, we found potential for benefit within the class of "prespecified adaptive designs" vs fully blinded group sequential procedures. However, when such prespecifications are lacking, which is very often the case in many of the adaptive literature, we found negligible benefit in a fully adaptive strategy. Additionally, when scientific considerations restrict extension of the calendar time, the best fully adaptive strategy loses efficiency over fully blinded strategies. Our results have implications on the notion of the cost of not planning to plan a RCT properly. The "price to pay" with a fully adaptive strategy is a

substantial loss of power when regulatory agencies are not convinced that sponsors have not demonstrated any form of "intent-to-cheat" due to potential operational bias with these "less well-understood" unplanned, unblinded procedures.

In many settings, investigators speculate on the potential possibility of a waning treatment effect and thus considered the plausibility of gaining power under such hypothesized alternatives with the use of approaches to differentially weigh the survival curves. This leads to the question of whether "less well understood" approaches can be used together with these "less well understood" adaptive designs. Many "less well-understood" time to event analyses methods may be considered to gain power under time varying treatment alternatives but are however evaluated otherwise under the strong null setting for their operating characteristics. However, many of such procedures to control the overall Type 1 error require characterizing the information growth. We investigate the use of weighted logrank statistics on the control of overall Type 1 error with unblinded adaptations to evaluate the robustness of adaptive procedures to misspecification of information growth.

We find that in situations when an unblinded interim analysis is made to modify aspects of the censoring distribution as quantified by the accrual of subjects, this can indiscriminately affect the control of the overall Type 1 error by as much as 20%. When further evaluating the robustness of current adaptive procedures to these "less well understood" survival procedures, the typical assumption of not needing to adjust when one has not adapted no longer holds true. We found that when adaptations are made to increase accrual based on unblinded results, or even potentially switching from logrank to Wilcoxon statistic to gain efficiency, there is added loss of precision to quantifying this nonlinear information growth that is dependent on both the censoring and survival distribution. This difficulty to accurately quantify the maximum information growth leads to difficulty in adequately controlling our Type 1 error inflation. In such situations, when the protocol allows for potential unblinded adaptation with these "less well-understood" analyses methods for censored data, the consequence of such is the need always adjust for any potential for operational bias.

In presence of time varying treatment effects, there is less clarity on how best to answer

any of these questions. We discussed plausible scenarios on how to describe this outcome space and investigated the degree to which censoring affects time varying treatment effects either as a consequence of interim data analyses and/or incomplete accrual of subjects. We found that in many situations, "less well-understood" analyses methods used in practice are similarly affected by the degree of censoring, such that they are affect the scientific interpretation of the study results.

This presents a challenge when investigators posit the possibility of crossing survival curves during the design of clinical trials and considered the plausibility of placing scientific weightings to switch between test statistics. We investigated the consequences of such use based on an example taken from Logan et al. [2008] to evaluate the potentially robustness of the test statistics when such crossings are spurious. We find difficulty in the use of their proposed statistics to distinguish between a true crossing as opposed to the scenario when such crossings are spurious. We then compared their results some of the commonly used test statistics and evaluated how they fare when the primary objective is to identify the better long term treatment with the plausibility that earlier on, there may be strong potential benefit of survival for one group but then this benefit essentially disappears.

We investigate further how the totality of these findings may be used in a DMC situation when they faced difficult ethical dilemmas. With sequential boundaries imposed using a group sequential rules, we investigate the degree to which naïvely chosen conservative rules that are thought to protect against early conservatism can affect decision making based on "less well-understood" statistics. We find that in such situations, the sole use of the monitoring boundaries to guide decision making is insufficient to judge survival benefit and that over-reliance on obtaining statistically significance without a proper quantification of the evidence of the treatment can results in poor decision making. When there is potential for time varying treatment effects, the planning of a sequential monitoring rule has to be even more cautiously specified so that we do not tend to stop too early for unanticipated differences that do not matter clinically.

The general theme of this dissertation is to evaluate the potential for benefits and risk

on the use of adaptive methods. There has been an explosion of literature pushing for more innovative designs to speed up the drug discovery process. While these creative approaches are laudable, many of these approaches are lacking comprehensive evaluations to what are potential operational and logistical issues in practice. Our results in the time to event setting have found that there are many potential issues that can arise that are not easily solved. It is not sufficient to assume a proposed monitoring rule that is known to be conservative to hold in practice. Instead, there is a need to consider a wide variety of designs to understand the operating characteristics, robustness to misspecification of event rates. Additionally, when positing time varying treatment effects, it is not sufficient to only consider the alternatives based on prior trials but to evaluate them with regard to potential possible scenarios that could well occur so as not to be surprised by the outcome.

There are many operational issues with the use of adaptive designs in the time to event setting that has not been considered in this dissertation. One of the main issues is that in many clinical trials, interim analyses often consist of many other endpoints that are measured immediately. Of concern is whether the implementation of these adaptive approaches has been made with these additional information. This issue was pointed out by Bauer and Posch [2004]. In such situations, there is high risk that approaches by Cui et al. [1999] are unable to correct for the use of these secondary endpoints.

The design of clinical studies entail experimenting with a treatment with unproven benefit to risk profile on human volunteers with the objective of being able to reliably establish evidence of efficacy, effectiveness, and safety. Such design and planning is often an iterative collaborative process where concerns from various communities, such as, the scientific researchers, regulatory agencies, ethical review boards, logistics etc, have to be accommodated. The thought process going into a design of a time to event setting is more complex when we are trying to determine the appropriate strategy over time. Thus, a wide range of designs should be evaluated to be properly evaluated against potential misspecification (Chapter 5) to unanticipated low event rates, inferior group sequential design that do not balance the scientific concerns when there may potential crossing in survivals. In such settings, sequen-

tial rules have to be chosen with care, stress-tested with misspecification of assumptions, plausible time varying alternatives, specify what are the not surprising outcomes, and what are the early important clinical differences that matter so as to provide a comprehensive, achievable protocol that balances multiple competing goals of clinical research.

BIBLIOGRAPHY

Denise R Aberle, Amanda M Adams, Christine D Berg, William C Black, Jonathan D Clapp, Richard M Fagerstrom, Ilana F Gareen, Constantine Gatsonis, Pamela M Marcus, and JD Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine*, 365(5):395–409, 2011.

P Armitage, CK McPherson, and BC Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, pages 235–244, 1969.

Jared M. Baeten, Deborah Donnell, Patrick Ndase, Nelly R. Mugo, James D. Campbell, Jonathan Wangisi, Jordan W. Tappero, Elizabeth A. Bukusi, Craig R. Cohen, Elly Katabira, Allan Ronald, Elioda Tumwesigye, Edwin Were, Kenneth H. Fife, James Kiarie, Carey Farquhar, Grace John-Stewart, Aloysious Kakia, Josephine Odoyo, Akasiima Mucunguzi, Edith Nakku-Joloba, Rogers Twesigye, Kenneth Ngure, Cosmas Apaka, Harrison Tamooh, Fridah Gabona, Andrew Mujugira, Dana Panteleeff, Katherine K. Thomas, Lara Kidoguchi, Meighan Krows, Jennifer Revall, Susan Morrison, Harald Haugen, Mira Emmanuel-Ogier, Lisa Ondrejcek, Robert W. Coombs, Lisa Frenkel, Craig Hendrix, Namandjé N. Bumpus, David Bangsberg, Jessica E. Haberer, Wendy S. Stevens, Jairam R. Lingappa, and Connie Celum. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *New England Journal of Medicine*, 367(5):399–410, 2012. doi: 10.1056/NEJMoa1108524. URL `http://www.nejm.org/doi/full/10.1056/NEJMoa1108524`. PMID: 22784037.

Stuart Barber and Christopher Jennison. Optimal asymmetric one-sided group sequential tests. *Biometrika*, 89(1):49–60, 2002.

AD Barker, CC Sigman, GJ Kelloff, NM Hylton, DA Berry, and LJ Esserman. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical*

Pharmacology & Therapeutics, 86(1):97–100, 2009.

P Bauer and K Köhne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4):pp. 1029–1041, 1994. ISSN 0006341X. URL http://www.jstor.org/stable/2533441.

Peter Bauer and Martin Posch. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller, statistics in medicine 2001; 20: 3741–3751. *Statistics in Medicine*, 23(8):1333–1334, 2004.

Yannis Bilias, Minggao Gu, and Zhiliang Ying. Towards a general asymptotic theory for cox model with staggered entry. *The Annals of Statistics*, 25(2):662–682, 1997.

Werner Brannath, Cyrus R Mehta, and Martin Posch. Exact confidence bounds following adaptive group sequential tests. *Biometrics*, 65(2):539–546, 2009.

Sean S Brummel and Daniel L Gillen. Flexibly monitoring group sequential survival trials when testing is based upon a weighted log-rank statistic. *Sequential Analysis*, 33(1):39–59, 2014.

Bart E Burington and Scott S Emerson. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics*, 59(4):770–777, 2003.

Michael A Carducci, Fred Saad, Per-Anders Abrahamsson, David P Dearnaley, Claude C Schulman, Scott A North, Darryl J Sleep, Jeffrey D Isaacson, and Joel B Nelson. A phase 3 randomized controlled trial of the efficacy and safety of atrasentan in men with metastatic hormone-refractory prostate cancer. *Cancer*, 110(9):1959–1966, 2007.

Myron N Chang. Confidence intervals for a normal mean following a group sequential test. *Biometrics*, pages 247–254, 1989.

YH Chen, David L DeMets, and KK Gordon Lan. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*, 23(7):1023–1038, 2004.

Jon Cohen. HIV treatment as prevention. *Science*, 334 (6063):1628, 2011. doi: 10.1126/science.334.6063.1628. URL http://www.sciencemag.org/content/334/6063/1628.short.

Myron S Cohen, Ying Q Chen, Marybeth McCauley, Theresa Gamble, Mina C. Hosseinipour,

Nagalingeswaran Kumarasamy, James G. Hakim, Johnstone Kumwenda, Beatriz Grinsztejn, Jose H.S. Pilotto, Sheela V. Godbole, Sanjay Mehendale, Suwat Chariyalertsak, Breno R. Santos, Kenneth H. Mayer, Irving F. Hoffman, Susan H. Eshleman, Estelle Piwowar-Manning, Lei Wang, Joseph Makhema, Lisa A. Mills, Guy de Bruyn, Ian Sanne, Joseph Eron, Joel Gallant, Diane Havlir, Susan Swindells, Heather Ribaudo, Vanessa Elharrar, David Burns, Taha E. Taha, Karin Nielsen-Saines, David Celentano, Max Essex, and Thomas R Fleming. Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 365(6):493–505, 2011. doi: 10.1056/NEJMoa1105243. URL `http://www.nejm.org/doi/full/10.1056/NEJMoa1105243`. PMID: 21767103.

L Cui, HM Hung, and Sue-Jane Wang. Modification of sample size in group sequential clinical trials. *Biometrics*, 55(3):853–857, 1999.

Robert B Davies. Rank tests for "lehmann's alternative". *Journal of the American Statistical Association*, 66(336):879–883, 1971.

David L DeMets and Gordon Lan. The α spending function approach to interim data analyses. In *Recent Advances in Clinical Trial Design and Analysis*, pages 1–27. Springer, 1995.

Jonathan S Denne. Sample size recalculation using conditional power. *Statistics in Medicine*, 20(17-18):2645–2660, 2001.

John D Eales and Christopher Jennison. An improved method for deriving optimal one-sided group sequential tests. *Biometrika*, 79(1):13–24, 1992.

Susan S Ellenberg, Thomas R Fleming, and David L DeMets. *Data monitoring committees in clinical trials: a practical perspective*. John Wiley &; Sons, 2003.

EMA. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. *Committee for Medicinal Products for Human Use and others: London, UK*, 2007.

Sarah C Emerson and Scott S Emerson. Detecting differential gene expression in subgroups of a disease population. *The International Journal of Biostatistics*, 9(1), 2013.

Sarah C Emerson, Kyle D Rudser, and Scott S Emerson. Exploring the bene-

fits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in Medicine*, 30(11):1199–1217, 2011a. ISSN 1097-0258. doi: 10.1002/sim.4156. URL http://dx.doi.org/10.1002/sim.4156.

Scott S Emerson. S+ Seqtrial technical overview. *Data Analysis Products Division, MathSoft, Inc*, 2000.

Scott S Emerson. Issues in the use of adaptive clinical trial designs. *Statistics in Medicine*, 25(19):3270–3296, 2006. ISSN 1097-0258. doi: 10.1002/sim.2626. URL http://dx.doi.org/10.1002/sim.2626.

Scott S Emerson. Some observations on the wilcoxon rank sum test. *UW Biostatistics Working Paper Series*, 380, 2011.

Scott S Emerson and Thomas R Fleming. Symmetric group sequential test designs. *Biometrics*, pages 905–923, 1989.

Scott S Emerson and Thomas R Fleming. Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):875–892, 1990. doi: 10.1093/biomet/77.4.875.

Scott S Emerson and Thomas R Fleming. Adaptive methods: telling "the rest of the story". *Journal of biopharmaceutical statistics*, 20(6):1150–1165, 2010.

Scott S Emerson, John M Kittelson, and Daniel L Gillen. On the use of stochastic curtailment in group sequential clinical trials. *UW Biostatistics Working Paper Series. Working Paper 243*, 2005. URL http://biostats.bepress.com/uwbiostat/paper243.

Scott S Emerson, John M Kittelson, and Daniel L Gillen. Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*, 26(28):5047–5080, 2007.

Scott S Emerson, Gregory P Levin, and Sarah C Emerson. Comments on "Adaptive increase in sample size when interim results are promising: A practical guide with examples". *Statistics in Medicine*, 30(28):3285–3301, 2011b. ISSN 1097-0258. doi: 10.1002/sim.4271. URL http://dx.doi.org/10.1002/sim.4271.

FDA. Draft guidance for industry: Adaptive design clinical trials for drugs and biologics. *Food and Drug Administration. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), Rockville MD, USA*, 2010.

FDA. Draft guidance for industry: Adaptive designs for medical device clinical studies. *Food and Drug Administration. Center for Devices and Radiological Health (CDRH) and Center for Biologics Evaluation and Research (CBER), Rockville MD, USA*, 2015.

Lloyd D Fisher. Self-designing clinical trials. *Statistics in Medicine*, 17(14):1551–1562, 1998.

Thomas R Fleming. Standard versus adaptive monitoring procedures: a commentary. *Statistics in Medicine*, 25(19):3305–3312, 2006. ISSN 1097-0258. doi: 10.1002/sim.2641. URL `http://dx.doi.org/10.1002/sim.2641`.

TR Fleming and DP Harrington. Counting processes and survival analysis. *New York*, 1991.

Food, Drug Administration, et al. Guidance for clinical trial sponsors: Establishment and operation of clinical trial data monitoring committees. *OMB Control*, (0910-0581), 2006.

Ping Gao, James H Ware, and Cyrus Mehta. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, 18(6):1184–1196, 2008.

Ping Gao, Lingyun Liu, and Cyrus Mehta. Exact inference for adaptive group sequential designs. *Statistics in Medicine*, 32(23):3991–4005, 2013.

Michael P Garcia. Adaptive randomization ratios in multi-arm clinical trials. Master's thesis, University of Washington, 2015.

Daniel L Gillen and Scott S Emerson. Information growth in a family of weighted logrank statistics under repeated analyses. *Sequential Analysis*, 24(1):1–22, 2005.

Daniel L Gillen and Scott S Emerson. Nontransitivity in a class of weighted logrank statistics under nonproportional hazards. *Statistics & probability letters*, 77(2):123–130, 2007.

Anthony H Goldstone, Susan M Richards, Hillard M Lazarus, Martin S Tallman, Georgina Buck, Adele K Fielding, Alan K Burnett, Raj Chopra, Peter H Wiernik, Letizia Foroni, et al. In adults with standard-risk acute lymphoblastic leukemia, the greatest benefit is achieved from a matched sibling allogeneic transplantation in first complete remission, and an autologous transplantation is less effective than conventional consolidation/maintenance chemotherapy in all patients: final results of the international all trial (mrc ukall xii/ecog e2993). *Blood*, 111(4):1827–1833, 2008.

A Lawrence Gould. Interim analyses for monitoring clinical trials that do not materially

affect the type i error rate. *Statistics in medicine*, 11(1):55–66, 1992.

A Lawrence Gould and Victor J Pecore. Group sequential methods for clinical trials allowing early acceptance of ho and incorporating costs. *Biometrika*, 69(1):75–80, 1982.

A Lawrence Gould and W Joseph Shih. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics-Theory and Methods*, 21(10):2833–2853, 1992.

A Lawrence Gould and W Joseph Shih. Modifying the design of ongoing trials without unblinding. *Statistics in Medicine*, 17(1):89–100, 1998.

The Look AHEAD Research Group. Look AHEAD (action for health in diabetes): design and methods for a clinical trial of weight loss for the prevention of cardiovascular disease in type 2 diabetes. *Controlled Clinical Trials*, 24(5):610 – 628, 2003. ISSN 0197-2456. doi: http://dx.doi.org/10.1016/S0197-2456(03)00064-3. URL `http://www.sciencedirect.com/science/article/pii/S0197245603000643`.

Ming Gao Gu and Tze Leung Lai. Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *The Annals of Statistics*, 19(3):1403–1433, 1991.

Sebastian Irle and Helmut Schäfer. Interim design modifications in time-to-event studies. *Journal of the American Statistical Association*, 107(497):341–348, 2012.

Holly Janes, Peter Gilbert, Susan Buchbinder, James Kublin, Magdalena E Sobieszczyk, and Scott M Hammer. In pursuit of an HIV vaccine: Designing efficacy trials in the context of partially effective nonvaccine prevention modalities. *AIDS research and human retroviruses*, 29(11):1513–1523, 2013.

Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4):347–356, 2011.

Christopher Jennison. Efficient group sequential tests with unpredictable group sizes. *Biometrika*, 74(1):155–165, 1987.

Christopher Jennison and Bruce W Turnbull. Group-sequential analysis incorporating co-

variate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997.

Christopher Jennison and Bruce W Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

Christopher Jennison and Bruce W Turnbull. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22(6):971–993, 2003.

Christopher Jennison and Bruce W Turnbull. Adaptive and nonadaptive group sequential tests. *Biometrika*, 93(1):1–21, 2006a.

Christopher Jennison and Bruce W Turnbull. Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, 25(6):917–932, 2006b.

Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

John M Kittelson and Scott S Emerson. A unifying family of group sequential test designs. *Biometrics*, 55(3):874–882, 1999.

KK Gordon Lan and David L DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

Walter Lehmacher and Gernot Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290, 1999.

Eric L Lehmann. The power of rank tests. *The Annals of Mathematical Statistics*, pages 23–43, 1953.

Gregory Levin. *An Evaluation of Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sampling Plan*. PhD thesis, 2013.

Gregory P Levin, Sarah C Emerson, and Scott S Emerson. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Statistics in medicine*, 32(8):1259–1275, 2013.

Gregory P Levin, Sarah C Emerson, and Scott S Emerson. An evaluation of inferential procedures for adaptive clinical trial designs with pre-specified rules for modifying the

sample size. *Biometrics*, 70(3):556–567, 2014.

DY Lin, L Shen, Z Ying, and NE Breslow. Group sequential designs for monitoring survival probabilities. *Biometrics*, pages 1033–1041, 1996.

Brent R Logan and Shuyuan Mo. Group sequential tests for long-term survival comparisons. *Lifetime Data Analysis*, 21(2):218–240, 2015. ISSN 1380-7870. doi: 10.1007/s10985-014-9298-4. URL `http://dx.doi.org/10.1007/s10985-014-9298-4`.

Brent R Logan, John P Klein, and Mei-Jie Zhang. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics*, 64 (3):733–740, 2008.

Dominic Magirr, Thomas Jaki, Franz Koenig, and Martin Posch. Adaptive survival trials. *arXiv preprint arXiv:1405.1569*, 2014.

Dominic Magirr, Thomas Friedrich Jaki, Franz Koenig, and Martin Posch. Sample size reassessment and hypothesis testing in adaptive survival trials. *PLoS ONE*, 2016.

Cyrus Mehta, Helmut Schäfer, Hanna Daniel, and Sebastian Irle. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine*, 33(26):4515–4531, 2014.

Cyrus R Mehta and Stuart J Pocock. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, 30(28):3267–3284, 2011.

Cyrus R Mehta and Anastasios A Tsiatis. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal*, 35(4):1095–1112, 2001.

Hans-Helge Müller and Helmut Schäfer. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):pp. 886–891, 2001. ISSN 0006341X. URL `http://www.jstor.org/stable/3068429`.

Hans-Helge Müller and Helmut Schäfer. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*, 23(16):2497–2508, 2004.

Peter C O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*,

pages 1079–1087, 1984.

Peter C O'Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

S Pampallona, AA Tsiatis, and K Kim. Spending functions for the type I and type II error probabilities of group sequential tests. Technical report, Technical Report, Dept. of Biostatistics, Harvard School of Public Health, Boston, 1995.

Margaret Sullivan Pepe and Thomas R Fleming. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, pages 497–507, 1989.

Margaret Sullivan Pepe and Thomas R Fleming. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 341–352, 1991.

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

Stuart J Pocock, Nancy L Geller, and Anastasios A Tsiatis. The analysis of multiple endpoints in clinical trials. *Biometrics*, pages 487–498, 1987.

Martin Posch and Peter Bauer. Adaptive two stage designs and the conditional error function. *Biometrical Journal*, 41(6):689–696, 1999.

Michael A Proschan and Sally A Hunsberger. Designed extension of studies based on conditional power. *Biometrics*, pages 1315–1324, 1995.

Michael A. Proschan, Dean A. Follmann, and Myron A. Waclawiw. Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48(4):pp. 1131–1143, 1992. ISSN 0006341X. URL `http://www.jstor.org/stable/2532704`.

Kyle D Rudser. *Variable importance in predictive models: separating borrowing information and forming contrasts*. PhD thesis, University of Washington, 2007.

Brittany J Sanchez. Evaluation of strategies for the Phase II to Phase III progression in treatment discovery. Master's thesis, University of Washington, 2014.

Helmut Schäfer and Hans-Helge Müller. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine*, 20

(24):3741–3751, 2001.

Daniel O Scharfstein, Anastasios A Tsiatis, and James M Robins. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association*, 92(440):1342–1350, 1997.

Norbert Schmitz. *Optimal Sequentially Planned Decision Procedures*, volume 79 of *Lecture Notes in Statistics*. Springer-Verlag New York, 1993.

David A Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, pages 499–503, 1983.

Yu Shen and Jianwen Cai. Sample size reestimation for clinical trials with censored survival data. *Journal of the American Statistical Association*, 98(462):418–426, 2003.

Yu Shen and Lloyd Fisher. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics*, 55(1):190–197, 1999.

Abigail Shoben, Kyle Rudser, and Scott S Emerson. Estimates of information growth in longitudinal clinical trials. 2010.

Abigail B Shoben. *Information Growth in Longitudinal Clinical Trials*. PhD thesis, University of Washington, 2010.

Abigail B Shoben and Scott S Emerson. Violations of the independent increment assumption when using generalized estimating equation in longitudinal group sequential trials. *Statistics in Medicine*, 33(29):5041–5056, 2014.

Galen Shorack. Biostat 512/513: Lecture notes material, 2010.

National Lung Screening Trial Research Team et al. The national lung screening trial: overview and study design1. *Radiology*, 2011.

The Women's Health Initiative Study Group. Design of the women's health initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1):61 – 109, 1998. ISSN 0197-2456. doi: http://dx.doi.org/10.1016/S0197-2456(97)00078-0. URL `http://www.sciencedirect.com/science/article/pii/S0197245697000780`.

Kanae Togo and Manabu Iwasaki. Sample size re-estimation for survival data in clinical trials with an adaptive design. *Pharmaceutical Statistics*, 10(4):325–331, 2011. ISSN 1539-1612.

doi: 10.1002/pst.469. URL http://dx.doi.org/10.1002/pst.469.

Anastasios A Tsiatis. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*, 77(380): 855–861, 1982.

Anastasios A Tsiatis and Cyrus Mehta. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90(2):367–378, 2003.

Anastasios A Tsiatis, Gary L Rosner, and Cyrus R Mehta. Exact confidence intervals following a group sequential test. *Biometrics*, pages 797–803, 1984.

Mark Van Der Laan. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Avy Violari, Mark F Cotton, Diana M Gibb, Abdel G Babiker, Jan Steyn, Shabir A Madhi, Patrick Jean-Philippe, and James A McIntyre. Early antiretroviral therapy and mortality among HIV-infected infants. *New England Journal of Medicine*, 359(21):2233–2244, 2008.

Samuel K Wang and Anastasios A Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43(1):193–199, 1987.

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

John Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.

John Whitehead, Anne Whitehead, Susan Todd, Kim Bolland, and M Roshini Sooriyarachchi. Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine*, 20 (2):165–176, 2001.

Janet Wittes and Erica Brittain. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–72, 1990.

Ronghui Xu and John O'Quigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439, 2000.

Appendix A

Time To Event Trials of Interests

A.1 National Lung Screening Trial

In the National Lung Screening Trial (NLST), the objective of the trial was to determine whether the use of screening procedure via low dose CT (computed tomography) as compared to the use of chest radiography as a screening measure for lung cancer can reduce lung cancer mortality among smokers or previous smokers with no history of lung cancer or cancers of other form [Team et al., 2011]. 53,454 participants were enrolled between the period of August 2002 through April 2004. A total of three annual screenings were to be conducted for all eligible participants. Compliance was high in both arms.

The objective of the NLST is to determine whether the use of low-dose CT as compared to the use of the chest radiography was effective in prevention of mortality from lung cancer. T0, T1, and T2 are the screening visits at baseline, year 1, and year 2. Following the screening visits, the participants are further followed for an additional 6 - 8 years with the duration of the trial to take place for a total study time of 10 years (including 3 years of screening). A weighted version of the logrank statistic, linear ramp was placed from time of randomization to year 4 with a full weighting after 4 years, was chosen with the scientific rational of emphasizing weights on later deaths relative to deaths within the first 4 years. A total of 6 analyses were performed such that interim analyses were conducted annually from 2006 to 2009, and two analyses were performed semiannually in 2010.

During the first few years of screening, the LDCT arm detected a higher proportion of early stage lung cancer relative to chest radiography, while chest radiography detected

a higher frequency of later stage cancer. A lower death rate observed during statistical monitoring raised study concerns. If the study was extended longer, the study would have addressed a totally different scientific question. In particular, the first three years of baseline screening would not have been sufficient to prevent lung cancer mortality for a longer period than what the study had planned for. By extending the trial, the study would not have been able to address whether LDCT was efficacious as a screening measure to detect early stages of lung cancer, and thus prevent lung cancer mortality.

A.2 HPTN052

HPTN052 was declared as one of the scientific breakthroughs of the year in 2011 [Cohen, 2011]. The primary objective in HPTN052 was to determine whether the early use of combination anti-retroviral therapy (ART) in infected patients among serodiscordant couples is effective in the prevention of HIV-1 transmission to uninfected partners [Cohen et al., 2011]. The trial was designed to provide at least 87% power to detect at least 39% reduction in the primary endpoint of HIV incidence. Based on various logistical constraints (18 months accrual, projected completion of follow-up at 6.5 years), an estimated accrual size of 1750 participants was computed assuming average 5-year placebo (13.2%) and treatment (8.3%) event rates, and an anticipated 188 events.

Six years after HPTN052 started, blinded analysis during a planned formal interim review showed 39 HIV infections among the 1,763 enrolled couples (877 on delayed vs 886 on early) with 28 of them being linked transmissions. Unblinded analysis showed that 27 of the linked transmissions arose on the delayed ART arm while only one came from the early ART arm yielding a hazard ratio of 0.04 (95% CI: 0.01 - 0.27; p-value < 0.0001). On the basis of this analysis, the DSMB recommended stopping further follow-up in the RCT due to demonstrated efficacy of the experimental treatment.

The pilot phase of the study Pilot phase started in April 2005, and enrollment took place from June 2007 through May 2010. The DSMB took place on April 2011 on the basis data collected up to Feb 2011. Composite monitoring of primary outcomes took place around

30% of the 340 events, i.e., at 105 events which result in crossing the efficacy boundary. In summary, no interim monitoring was planned solely for the primary outcome for only HIV infections.

A.3 Partners Pre-Exposure Prophylaxis (PrEP)

Partners PrEP is a Phase III, randomized control, double-blind, three arm trial of daily oral tenofovir (TDF) and emtricitabine/tenofovir (FTC/TDF) PrEP for the prevention of HIV as their primary endpoint among HIV serodiscordant partners [Baeten et al., 2012]. Based on a placebo event rate of 2.75 infections per 100 person-years (PY), 4747 HIV serodiscordant couples randomized with equal probability to the three arms followed for 36 months would be expected to provide the necessary number of events. Using a group sequential design with up to a maximum of four planned interim analyses, the trial was stopped early at the third interim analysis due to crossing the efficacy boundary. The observed placebo event rates were much smaller than what was used in planning with the observed treatment effects to be more extreme than had been anticipated.

A.4 Children with HIV Early Antiretroviral Therapy (CHER) Trial

We motivate the scientific and ethical deliberations of the DSMB charged with monitoring the CHER trial [Violari et al., 2008]. The primary objective of the study was to determine whether a limited course of ART administered to babies immediately (with interruptions) when their HIV statuses are known would have a long-term heath benefit, when compared to HIV-infected babies who are treated continuously with ART only after they developed symptoms of HIV, or weakened immune systems (referred to deferred ART). At design stage, a total of 375 children are projected to be enrolled over an 18 month accrual period, to provide a minimum of 3.5 years of followup data. Such a design would provide at least 80% power to reject the null hypothesis of no difference among the three groups assuming the global log-rank test with a two-sided alpha level of 0.05. Under their hypothesized alternative, they

postulated the possibility of crossing hazards.

The guiding statistical criterion for monitoring the trial is based on a difference of at least 3 SD in the log relative hazard (or nominal P<0.001) in any interim analysis (according to the Haybittle-Peto rule). During the course of the trial monitoring, a strong treatment effect was observed at the second interim analysis. However, the average follow-up of all participants was 40 weeks (Range: 24 - 58), far shorter than what was planned at the beginning of the trial. The DMC faced the dilemma of whether it was ethical to continue followup, and that if the hypothesized crossing of hazards was real, this early separation of survival curves may come back together when followup was extended. Despite these dilemmas, the DMC recommended stopping the trial with compelling strong evidence of difference in mortality (75% reduction in risk of mortality) being demonstrated. The DMS further recommended that babies in the deferred therapy group not currently receiving ART to possibly initiate ART, and continue followup for all three groups.

A.5 Autologous vs Allogenic Stem Cell Transplant

In both Logan et al. [2008] and Logan and Mo [2015]'s paper, they provided the bone marrow transplant example and a RCT in the disease setting of acute lymphoblastic leukemia (ALL) RCT as scientific motivations on identifying the treatment with better long term survival. The ALL RCT was designed to address two questions. One of which was to compare whether there was evidence of allogenic effect of transplantation between those who receive allogenic vs those who did not receive allogenic transplant. The other scientific question of interest was to determine whether among patients who did not receive allogenic transplant, does those who had autologous transplant compared to those on maintenance/chemotherapy be effective [Goldstone et al., 2008]. In this trial, the patients who had allogenic transplantation were observed to have higher risk of mortality, and have higher probability of survival at the end of 5 year relative to patients on the autologous transplantation arm. In this clinical trial example, as motivated in Logan et al. [2008], there was compelling evidence that the survival curves do indeed cross sometime after 2 years.

The scientific rationale of the potential of crossing hazards/survival using the ALL trial as an example can be argued as follows: Patients having an autologous bone marrow transplant may be at high risk of infections (organ related) after exposure to chemotherapy use. This suppress the patients' immune system in order to facilitate the stem cell transplant early on after the transplant. Upon surviving this period of weakened immune, the patients may recover over some period of time. On the other hand, it is plausible that had their stored stem not been completely purged of cancer cells, these patients' may have higher risk of relapse later in time. Thus, the prognosis for patients' whose stored stem cells were not free of malignant cancer cells, may have higher probability of relapse over time when their new immune system were unable to purge these cells.

For patients having allogenic bone marrow transplant in the ALL trial, their initial hazard may differ from patients on the autologous arm. In addition to the high risk of infections after exposure, there is additional hazard as a consequence of Graft-vs-Host disease (GVHD), or other organ related complications if the donor's HLA is not a perfect match. Thus, the initial risk of mortality may be higher for patients on the allogenic arm as compared to the autologous arm. This initial high risk or hazard rate may change greatly over time if either the donor's stem cells are a perfect match to the patient's. Therefore, the risk of GVHD becomes negligible, or the patient survives the complications of GVHD, and makes a near complete recovery if the host accepts the foreign cells over time, "curing" the disease. Thus, at any time, it is entirely possible that the hazards for each treatment group crosses after some time. However, depending on the risk of dying, it is entirely possible that some crossing in hazard rates for the two treatment group can potentially give rise to stochastically ordered, survival curves over the time frame of interest, with high probability of observing a spurious crossing.

Appendix B

Additional Results for Chapter 2

B.1 Blinded Repowering: Sepsis Example

We refer to the sepsis example described in Emerson et al. [2007] for illustration. Sepsis is a potentially life threatening complication of an infection. Typically, infection starts out in several stages. Sepsis occurs when chemicals released into bloodstream to fight an infection trigger an inflammatory response, resulting in a cascade of changes that may damage multiple organs, resulting in fatality. Often, this can induce septic shock. The current treatment is antibiotics. However, this works only if the infection was a direct result of Gram positive bacteria infection rather than non-bacteria/Gram negative bacteria infection.

The sepsis study was designed to compare the 28-day mortality probabilities between groups of patients receiving antibodies to endotoxin and groups of patients who receive placebo. Consider the following notation in comparing the difference in the probability of 28-day mortality, we define X_{ik} to be the indicator of 28 day mortality for the i^{th} patient on the k treatment group where $k = 0$ if randomly assigned to placebo and 1 if randomly assigned to treatment group for $i = 1, \cdots, n$. Then, our outcome $X_{ik} \sim \mathcal{B}(1, p_k)$. The target parameter, $\theta = -(p_1 - p_0)$, compares the difference in the probability of dying on the placebo arm relative to the treatment arm, parametrized such that positive values of θ correspond to a benefit in the experimental treatment (relative to the placebo). We are concerned in testing the null hypothesis $\mathbb{H}_0 : \theta \leq \theta_0$ vs $\mathbb{H}_A : \theta \geq \theta_{\text{Alt}}$ at some level α. We motivate this example when the background rate for the placebo group may be incorrect.

We can construct the estimator for $\widehat{\theta} = -(\widehat{p}_1 - \widehat{p}_0) = \sum_{i=1}^{n} X_{i0}/n - \sum_{i=1}^{n} X_{i1}/n$ for which n

subjects are recruited per arm. Our asymptotic distribution of $\widehat{\theta}$ is $\mathcal{N}\left(\theta, \frac{p_0(1-p_0)}{n} + \frac{p_1(1-p_1)}{n}\right)$. At level α, in order to detect the design alternative, θ_A, at power β, our sample size formula follows from the immediate setting with $V = p_0(1-p_0) + p_1(1-p_1)$.

Assuming the design in Emerson et al. [2007], we let our hypothesized background rate be 0.30, and our design alternative $\theta_A = 0.07$. Then, the incidence rate for the experimental treatment is 0.23. We want to test $H_0 : -(p_1 - p_0) \leq \theta_0$ vs $H_A : -(p_1 - p_0) \geq \theta_A$. Based on the sample size formula, using a one-sided level $\alpha = 2.5\%$, presuming the baseline incidence rate $p_0 = 0.30$, and the treatment arm to decrease the incidence rate to $p_1 = 0.23$, then a fixed sample design based on 1700 patients (i.e., $n_0 = n_1 = 850$ per arm) yield a statistical power of 90.7% to detect the design alternative $\theta_A = 0.07$.

We assume a GSD is planned with 4 equally spaced looks as according to the hybrid monitoring boundary that comprises of an OBF lower efficacy boundary, and a upper futility boundary corresponding to P = 0.8 in the unified family. This would correspond to the *Futility.8* stopping boundary in the sepsis example that was chosen by the sponsor in Emerson et al. [2007]. Based on this design, for a sample size of 1700, we would have 88.8% power to detect the design alternative θ_A. The boundary values are presented in Table B.1. We denote superscripts to refer to the interim analysis at which the test statistics or estimates were computed.

Table B.1: Summary of the *Futility.8* boundary values on either the Z statistic, sample mean (θ) scale, or fixed sample P-value (lower) scale.

Analyses (Sample Size)	Z		θ		P (lower)	
	Futility	Efficacy	Futility	Efficacy	Futility	Efficacy
Time 1 (n= 425)	-1.1082	3.9756	-0.0473	0.1697	0.00004	0.86611
Time 2 (n= 850)	0.3211	2.8112	0.0097	0.0848	0.00247	0.37408
Time 3 (n= 1275)	1.2577	2.2953	0.0310	0.0566	0.01086	0.10425
Time 4 (n= 1700)	1.9878	1.9878	0.0424	0.0424	0.02342	0.02342

Specifically, at any interim analysis, $\widehat{p}$ is used to estimate the common mortality proba-

bility $\widehat{p}$ under the null hypothesis of no treatment effect with the test statistic

$$Z = \frac{-\widehat{p}_1 + \widehat{p}_0}{\sqrt{\widehat{p}(1-\widehat{p})(1/n_1 + 1/n_0)}}$$

In particular, if the estimated variability of $\widehat{\theta}$ at the end of the trial matches the variability at design stage, i.e., $\widehat{Var}(\widehat{\theta}) = (0.3 \times 0.7 + 0.23 \times 0.77)/n = 0.3871/n$, the trial is said to have attained it's maximum statistical information. Otherwise, blinded revision of sample size may be required.

Suppose for now the event rate is as expected, i.e., $p_0 = 0.3$ but the true incidence rate of the treatment group to be 0.17. At the first interim analysis conducted based on 424 patients, 110 "successes" were observed (Table B.2). In a blinded setting, the overall estimate of $\overline{p}^1$ can be computed, and further solved via $\overline{p} = \frac{110}{424} = (n_1 p_0 + n_0 p_1)/(n_0 + n_1)$, giving rise to the estimate of $p_0^1 = 0.2944$. Since the computed $\widehat{Var}_{\overline{p}^1}$ based on the solved estimate of p_0^1, p_1^1 is $0.3818/n$, the deviation from the planned variance of $0.3871/n$ is less than 2%, the trial continues to the next interim analysis.

At the second interim analysis, $\widehat{V}_{\overline{p}^2} = 0.3587/n$ is slightly lower than anticipated. In this analysis, presuming that our final sample size is 1700, then our estimated quantity $\widehat{V}_2$ is 7% lower relative to $\widehat{V}_1$. However, the Z-statistic has crossed the efficacy boundary, and the trial may terminate for efficacy (if sufficient safety data has been obtained) even though the aggregate event rate is lower.

Suppose now our event rate is truly incorrect, i.e., $p_0 = 0.2$, we consider the second scenario in the Table B.2 where the true treatment effect is 0.13 but our design alternative remains the same. In this case, at the first interim analysis, an observed lower event rate is seen with $\overline{p} = 0.1368$. A simple calculation of $\widehat{Var}_{\overline{p}^1} = 0.2337/n$ indicates that this quantity now deviates by about 40% from the planned variance of $0.3871/n$. In order to maintain the same statistical information, we need to adjust n to $1.66n$, which is more than 50% increase in sample size. However, since this interim analysis is early, it is possible that our estimates are more variable. In a blinded setting, we may choose to stay the course and continue to

Table B.2: Interim estimates based on a realization of simulated data. Column θ and Z would not be shown in a blinded setting. Instead, only $\overline{p}$ and the estimated $\widehat{Var}_{\overline{p}}$ computed based on evaluating p_0 using overall estimate of $\overline{p}$. The reference $V = 0.3871/n$.

p_0	p_1	Analyses (Sample Size)	Success ($\overline{p}$)	$\widehat{V}_{\overline{p}}$	θ	Z
0.3	0.17	Time 1 (n= 424)	110 (0.2594)	0.3818	0.113	2.659
		Time 2 (n= 850)	201 (0.2365)	0.3587	0.125	4.278
		Time 3 (n= 1274)	305 (0.2394)	0.3617	0.118	4.924
		Time 4 (n= 1700)	399 (0.2347)	0.3568	0.126	6.123
0.2	0.13	Time 1 (n= 424)	58 (0.1368)	0.2337	0.028	0.848
		Time 2 (n= 850)	116 (0.1365)	0.2332	0.066	2.798
		Time 3 (n= 1274)	194 (0.1523)	0.2557	0.072	3.587
		Time 4 (n= 1700)	255 (0.1500)	0.2525	0.072	4.143

the next interim analysis to decide whether this "low event rate" persists.

At the second interim analysis, there may be cause for worry since the estimated event rate is now similar to the previous interim analysis. $\widehat{Var}_{\overline{p}^2} = 0.2332/n$ is still not too different from the first interim analysis. To maintain the planned statistical information, we are still required to adjust our sample size by $> 50\%$. Two strategies are likely, (1) one can choose to stay the course and wait for the third interim analysis to decide. Alternatively, one may revise the design at this time to consider expanding the trial by having 50% more subjects and keeping with a total of four analysis.

In this case, with 50% more subjects, the schedule of interim analysis should take place when roughly 638, 1275, 1913, and 2550 subjects are recruited. Since we already spend two analysis at 424 and 850, then by revising the design now, our next two analyses should take place at roughly 1913 and 2550. The deliberation here is that if our blinded estimates were truly transient and the event rate will ultimately pick up at the later analyses, then it is infeasible/unethical to conduct this next analysis at a larger sample size than what we initially had planned. We may choose to act according to FDA guidance [FDA, 2010], and stay the course, and prospectively plan to increase the sample size when the event rates at

the third analysis are still low.

At the third interim analysis, we are now 75% of the way through the study if our maximum sample size is 1700, $\widehat{Var}_{\bar{p}^2} = 0.2332/n$ is clearly lower than anticipated. Although there is concrete evidence that our event rate is truly lower than what we anticipated, we have crossed the efficacy boundary. Had a sample size revision been performed at the second analysis, we would have proceed beyond our planned sample size even though the difference between our treatment effect and placebo is hypothesized correctly.

Note that if a FSD is assumed, then in either of the above setting, we would only terminate the study when we complete accrual of all 1,700 subjects. This is less efficient under there is misspecification of design assumption, or when the treatment effect is more optimistic than anticipated. Using a GSD, we see the relative merits of adaptively revising the trial to a smaller sample size in a blinded fashion rather than continuing to the planned maximal sample size.

In both of these illustrations, the blinded adaptation only considers the use of the overall estimate of the event rate rather than the average of the individual rates from both arms. Additionally, both illustrations make use of the hypothesized treatment effect θ at planning stage to update design assumptions. When we revise the variance, the inverse of our Fisher's information, we are essentially revising N/V as described in section 2.4. At any point during the study, the use of the aggregate estimates do not unblind any aspects of the study. This preservation of the blinding through the use of blinded adaptations are considered well-understood by regulatory bodies so long as the procedures are clearly documented, and the parties involved in making these blinded adaptations have no knowledge of the treatment assignments or other aspects of the unblinded data in the trial.

Many of the adaptive procedures that we described in section 2.6 are however interested in making an increase in sample size at the penultimate stage based on unblinded data. While the blinded procedure appears justified to protecting the integrity of the trial, there may be ethical reasons to unblind the study in presence of low "background rates" during the course of monitoring. If the true baseline event rate was lower than anticipated, and the

interim estimated treatment effect is close to our design alternative, then increasing accrual may be important to improving precision. However, if extreme treatment efficacy is the sole reason leading to a lower than anticipated event rate, then there may not be a strong rationale to increase the sample size since the monitoring rule can potentially allow early stopping (See Chapter 5).

Appendix C

Additional Results for Chapter 3

C.1 Additional Results for Section 3.2

C.1.1 Relative Efficiency: Doubling the Original Sample Size

Consider the following design strategy. At some interim analyses, we double the sample size of the original design from n to $2n$ $(= \widetilde{n}^*)$. By parameterizing $\theta = \gamma + 2$, the relative efficiency can be re-expressed as $\frac{2(\gamma+1)}{\gamma+2}$. $\gamma = n_1/n_2^*$ can be interpreted as the fraction of the sample size at the interim analysis relative to the remaining sample size to be accrued based on the original design. Figure C.1 shows the behavior of the relative efficiency as a function of γ.

When unplanned adaptations are made during the trial, this loss of efficiency depends on when the adaptation is made to increase the number of subjects. The remaining weight, n_2^*/n, has to be re-distributed over a bigger pool of stage two subjects when the adaptation is chosen closer to n, i.e., $n_1 \approx n$, to double the total number of subjects from n to $2n$. The inefficiency of the use of the weighted statistics to control for the inflation of the overall Type 1 error is illustrated by the change in relative efficiency. We see that the inefficiency of the weighted statistics increases the variance of this weighted estimator relative to the variance of the optimal estimator when assuming the use of the pre-specified design.

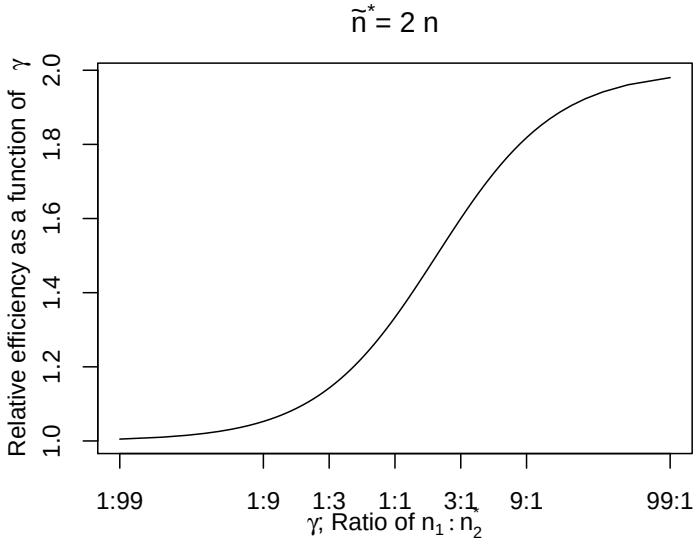

Figure C.1: Plot of relative efficiency vs γ under the simple setting of doubling the sample size of the original design after an interim analysis is made after accruing n_1 subjects.

C.1.2 Simulation Results for Adapting to A Larger Sample Size

We simulated 1,000,000 clinical trials, each with a sample size of $n = 100$ subjects, statistical power β to detect the design alternative of θ, known variance $\sigma^2 = 0.5$, and one-sided level $\alpha = 2.5\%$. Denote p to be the probability of adapting to a final sample size of $\tilde{n}$. Thus, $1 - p$ is the probability of staying the course with n respectively. The ASN is $(1 - p)n + p\tilde{n}$. We let an adaptation be conducted at an interim analysis corresponding to kn for discrete choices of $k = \{0.1, 0.2, \cdots, 0.9\}$. At the interim analysis, we do not stop the trial early for efficacy/futility. We considered values of $p \in (0.2, 0.3, 0.5, 0.7, 0.9)$.

We proceed to find the pre-specified adaptation that is best among all fully flexible designs with known probability p. We optimized the overall power of this fully, flexible design using a grid search for each known probability p of adapting under the design alternative. With this best fully adaptive rule, we find the empirical critical value that control the overall Type 1 error rate at 2.5% level based on a separate 1,000,000 simulations evaluated under the null

hypothesis using this best fully adaptive rule. We then evaluated the overall power for this best fully adaptive rule based on this empirical critical value, thus obtaining the power based on pre-specifying this adaptive rule. We then adjust this empirical critical value based on the approach by Cui et al. [1999] to control the overall Type 1 error in the fully adaptive setting when using the unblinded treatment results to make an adaptation. The results for these simulations are shown in Figure C.2, and presented in Table C.1 under the scenario for $\beta = 0.8$, and $\theta = 0.396204$.

When an earlier adaptation is conducted after $0.2n$ subjects are accumulated, the unplanned adaptation provides slight gain in terms of power relative to a fully prespecified procedure. We do not anticipate high efficiency gains at early adaptations since the estimated treatment effects at early interim analyses are less reliable as a consequence of possible random high bias. However, there is greater potential to be more "effective" in terms of determining the optimal sample size flexibly as we accumulate more statistical information (proportional to n_1). This flexibility at later adaptations does not enable us to be more efficient since we have already spend majority of the weights earlier on (Table C.1).

Specifically, we also considered the scenario where $\beta = 0.9$ and $\theta = 0.4584195$, with the objective of evaluating whether this adaptive strategy is consistent at higher power. The results based on adapting 50% of the time are presented in Table C.4. We observed negligible loss in overall power ($\sim 0.9\%$) from the use of reweighted statistics relative to having planned the design based on the minimum sufficient statistics (99.55% power).

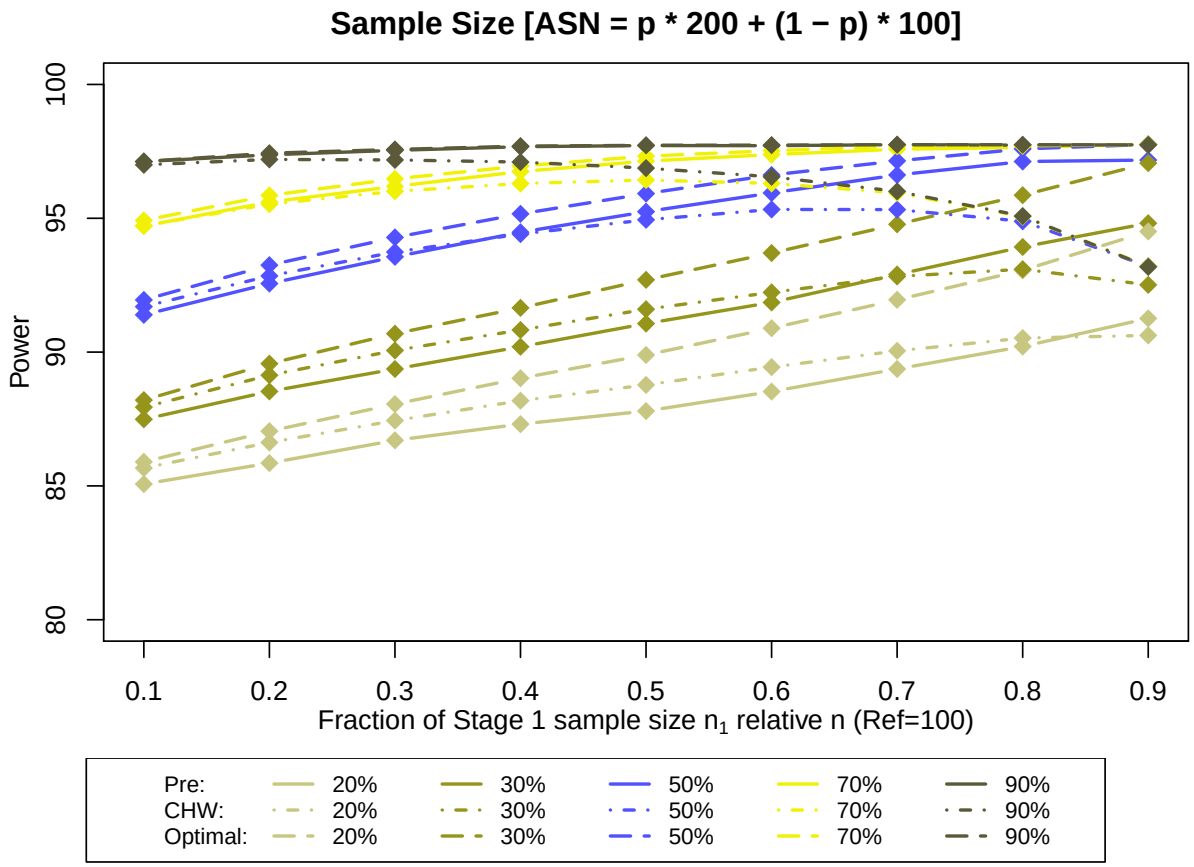

Figure C.2: Plot of overall power for adaptive design based on minimal sufficient statistics (Pre) vs the use of weighted statistics (CHW) when we considered various probabilities of increasing the sample size from $n = 100$ to 200. The prespecified adaptive design has higher power across various probability of increasing the final sample size. After adjusting for CHW, bigger loss of power is observed when late adaptations are made. At early adaptations, the loss of power is minimal relative to the prespecified adaptive design.

Table C.1: Simulation summary to double the original sample size n based on various probabilities, p, of adapting to $\widetilde{n} = 2n$. We examined the scenario when the original design has 80% power to detect an alternative $\theta = 0.396204$.

						n=100	$\widetilde{n} = 200$				ASN= $200p + 100(1-p)$				Sample size (FSD)		
	n_1	n_2^*	$\widetilde{n}_2^*$	γ	RE	Orig	CHW‡	Unadj	Adj$^\triangle$	$\triangle/\ddagger$	CHW‡	Unadj	Adj$^\triangle$	$\triangle/\ddagger$	SS‡	SS$^\triangle$	SS$^\triangle$/SS‡
$p = 20\%$	10	90	190	0.11	1.053	80.04	97.62	97.71	97.7	1.001	85.67	85.9	85.07	0.993	116.6	114.6	0.9828
	20	80	180	0.25	1.111	79.95	97.49	97.72	97.73	1.002	86.62	87.04	85.85	0.9911	120	117.3	0.9773
	30	70	170	0.43	1.176	79.98	97.35	97.72	97.74	1.004	87.44	88.06	86.7	0.9915	123	120.3	0.9775
	40	60	160	0.67	1.25	80.06	97.16	97.74	97.77	1.006	88.18	89.02	87.31	0.9901	126	122.5	0.9728
	50	50	150	1	1.333	79.98	96.91	97.72	97.72	1.008	88.77	89.84	87.8	0.989	128.4	124.4	0.9689
	60	40	140	1.5	1.429	80.06	96.54	97.72	97.73	1.012	89.44	90.76	88.52	0.9898	131.3	127.4	0.97
	70	30	130	2.33	1.538	80.05	96	97.73	97.72	1.018	90.05	91.63	89.37	0.9925	134.1	131	0.9768
	80	20	120	4	1.667	80.01	95.09	97.72	97.73	1.028	90.52	92.24	90.21	0.9965	136.4	134.9	0.9888
	90	10	110	9	1.818	79.96	93.12	97.72	97.72	1.049	90.62	92.96	91.26	1.007	136.9	140.2	1.024
$p = 30\%$	10	90	190	0.11	1.053	79.99	97.64	97.74	97.73	1.001	87.94	88.21	87.49	0.9949	125	123.2	0.9858
	20	80	180	0.25	1.111	80.04	97.52	97.75	97.75	1.002	89.14	89.57	88.53	0.9932	130	127.4	0.9802
	30	70	170	0.43	1.176	80.08	97.37	97.75	97.74	1.004	90.06	90.69	89.37	0.9923	134.2	131	0.9763
	40	60	160	0.67	1.25	79.97	97.17	97.73	97.74	1.006	90.83	91.65	90.2	0.9931	137.9	134.8	0.9776
	50	50	150	1	1.333	79.97	96.91	97.72	97.74	1.009	91.6	92.69	91.06	0.9942	142	139.1	0.9798
	60	40	140	1.5	1.429	80.01	96.54	97.75	97.72	1.012	92.23	93.67	91.86	0.9959	145.6	143.4	0.9851
	70	30	130	2.33	1.538	79.96	96.01	97.75	97.76	1.018	92.83	94.56	92.9	1.001	149.3	149.7	1.003
	80	20	120	4	1.667	80.04	95.07	97.72	97.73	1.028	93.1	95.33	93.93	1.009	151	156.9	1.039
	90	10	110	9	1.818	79.99	93.16	97.73	97.74	1.049	92.51	95.49	94.82	1.025	147.3	164	1.113
$p = 50\%$	10	90	190	0.11	1.053	79.94	97.64	97.74	97.72	1.001	91.7	91.95	91.39	0.9967	142.6	140.9	0.9882
	20	80	180	0.25	1.111	80	97.51	97.74	97.74	1.002	92.85	93.25	92.56	0.9969	149.4	147.6	0.988
	30	70	170	0.43	1.176	79.98	97.37	97.73	97.7	1.003	93.74	94.28	93.56	0.9981	155.5	154.2	0.9918
	40	60	160	0.67	1.25	80.04	97.14	97.73	97.74	1.006	94.41	95.16	94.48	1.001	160.6	161.1	1.003
	50	50	150	1	1.333	80.03	96.9	97.74	97.73	1.009	94.94	95.91	95.25	1.003	165.1	167.8	1.017
	60	40	140	1.5	1.429	80.02	96.58	97.73	97.71	1.012	95.33	96.59	95.95	1.006	168.6	174.8	1.037
	70	30	130	2.33	1.538	79.95	95.99	97.72	97.7	1.018	95.32	97.1	96.62	1.014	168.5	182.7	1.084
	80	20	120	4	1.667	79.99	95.1	97.75	97.72	1.027	94.89	97.46	97.12	1.024	164.6	189.7	1.153
	90	10	110	9	1.818	79.99	93.22	97.74	97.73	1.048	93.21	97.23	97.17	1.043	151.8	190.5	1.255

RE: (Conditional) Relative efficiency based on the relative variances of the flexible adaptive design to the pre-specified adaptive design.
CHW: Power after adjusting for the unplanned adaptation.
Unadj: Power computed based on naïve overall Type 1 error of $\alpha = 0.025$.
Adj: Adjusted power for the overall Type 1 error fixed at $\alpha = 0.025$.
SS‡: Sample size of a FSD based on the power obtained using CHW.
SS$^\triangle$: Sample size of a FSD based on the power obtained based on the (Adj)usted test for fixed overall Type 1 error of $\alpha = 0.025$.

Table C.2: Simulation summary to double the original sample size n based on various probabilities, p, of adapting to $\widetilde{n} = 2n$ (continued from previous table). We examined the scenario when the original design has 80% power to detect an alternative $\theta = 0.396204$.

						n=100	$\widetilde{n} = 200$				ASN= $200p + 100(1-p)$				Sample size (FSD)		
	n_1	n_2^*	$\widetilde{n}_2^*$	γ	RE	Orig	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	SS‡	SS$^{\triangle}$	SS$^{\triangle}$/SS‡
$p = 70\%$	10	90	190	0.11	1.053	80.03	97.64	97.74	97.77	1.001	94.72	94.92	94.71	0.9999	163.1	163	0.9995
	20	80	180	0.25	1.111	80	97.52	97.74	97.76	1.003	95.53	95.85	95.61	1.001	170.5	171.3	1.005
	30	70	170	0.43	1.176	80.03	97.36	97.73	97.73	1.004	96.01	96.47	96.19	1.002	175.5	177.6	1.012
	40	60	160	0.67	1.25	80.01	97.15	97.73	97.75	1.006	96.3	96.95	96.75	1.005	178.8	184.4	1.031
	50	50	150	1	1.333	80	96.9	97.73	97.74	1.009	96.43	97.32	97.14	1.007	180.4	189.9	1.053
	60	40	140	1.5	1.429	79.97	96.52	97.71	97.69	1.012	96.3	97.53	97.38	1.011	178.9	193.8	1.083
	70	30	130	2.33	1.538	80.02	96.02	97.74	97.71	1.018	95.95	97.69	97.58	1.017	174.9	197.2	1.127
	80	20	120	4	1.667	80.03	95.07	97.7	97.69	1.028	95.06	97.71	97.65	1.027	166.1	198.3	1.194
	90	10	110	9	1.818	80	93.21	97.75	97.78	1.049	93.21	97.75	97.77	1.049	151.8	200.6	1.322
$p = 90\%$	10	90	190	0.11	1.053	79.99	97.68	97.77	97.82	1.001	97	97.12	97.12	1.001	187.9	189.7	1.009
	20	80	180	0.25	1.111	80.04	97.56	97.77	97.76	1.002	97.2	97.43	97.37	1.002	190.9	193.6	1.014
	30	70	170	0.43	1.176	80.05	97.36	97.74	97.75	1.004	97.18	97.57	97.54	1.004	190.6	196.4	1.031
	40	60	160	0.67	1.25	79.99	97.18	97.76	97.77	1.006	97.1	97.7	97.67	1.006	189.4	198.8	1.05
	50	50	150	1	1.333	80.01	96.9	97.75	97.75	1.009	96.87	97.73	97.72	1.009	186.2	199.6	1.072
	60	40	140	1.5	1.429	80.04	96.56	97.74	97.72	1.012	96.55	97.73	97.71	1.012	181.9	199.6	1.097
	70	30	130	2.33	1.538	80.06	96.01	97.75	97.75	1.018	96.01	97.75	97.75	1.018	175.5	200.3	1.141
	80	20	120	4	1.667	79.99	95.08	97.75	97.73	1.028	95.08	97.75	97.73	1.028	166.3	199.9	1.202
	90	10	110	9	1.818	80	93.18	97.74	97.75	1.049	93.18	97.74	97.75	1.049	151.6	200.2	1.32

RE: (Conditional) Relative efficiency based on the relative variances of the flexible adaptive design to the pre-specified adaptive design.
CHW: Power after adjusting for the unplanned adaptation.
Unadj: Power computed based on naïve overall Type 1 error of $\alpha = 0.025$.
Adj: Adjusted power for the overall Type 1 error fixed at $\alpha = 0.025$.
SS‡: Sample size of a FSD based on the power obtained using CHW.
SS$^{\triangle}$: Sample size of a FSD based on the power obtained based on the (Adj)usted test for fixed overall Type 1 error of $\alpha = 0.025$.

Table C.3: Simulation summary to decrease the original sample size from n to $\widetilde{n} = 50$ with the probability of adaptation of 50% using CHW at different interim analyses for different power. We examined scenarios when the original design has either 80% power to detect the alternative of $\theta = 0.396204$, or 90% power to detect the alternative of 0.4584195 respectively.

						n=100	$\widetilde{n} = 50$				ASN = 75				Sample size (FSD)		
	n_1	n_2^*	$\widetilde{n}_2^*$	γ	RE	Orig	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	SS‡	SS$^{\triangle}$	SS$^{\triangle}$/SS‡
90% Power	5	95	45	0.05	1.028	89.98	62.6	62.98	63.08	1.008	78.16	79.78	78.73	1.007	71.33	72.35	1.014
	10	90	40	0.11	1.062	90.08	62.07	62.97	63.03	1.015	78.76	81.31	79.78	1.013	72.4	74.27	1.026
	15	85	35	0.18	1.107	89.99	61.53	62.97	62.91	1.022	79.12	82.47	80.7	1.02	73.05	76.06	1.041
	20	80	30	0.25	1.167	90.02	60.85	62.97	63.07	1.036	79.45	83.65	81.57	1.027	73.66	77.8	1.056
	25	75	25	0.33	1.25	90.03	60.04	63.04	62.97	1.049	79.67	84.79	82.28	1.033	74.07	79.26	1.07
	30	70	20	0.43	1.375	90.02	58.92	62.99	63.06	1.07	79.76	85.91	83.17	1.043	74.25	81.2	1.094
	35	65	15	0.54	1.583	89.95	57.37	62.98	62.87	1.096	79.58	87.03	84.07	1.056	73.91	83.24	1.126
	40	60	10	0.67	2	90.01	55.2	63.04	63.19	1.145	79.25	88.33	85.28	1.076	73.3	86.14	1.175
	45	55	5	0.82	3.25	90.01	51.4	62.96	63.06	1.227	78.08	89.69	86.48	1.108	71.19	89.23	1.253
80% Power	5	95	45	0.05	1.028	79.97	50.44	50.8	50.93	1.01	66.32	68.03	66.79	1.007	72.24	73.02	1.011
	10	90	40	0.11	1.062	80.04	50.06	50.86	51	1.019	66.68	69.32	67.4	1.011	72.84	74.06	1.017
	15	85	35	0.18	1.107	80.03	49.47	50.77	50.63	1.024	66.75	70.31	67.73	1.015	72.95	74.63	1.023
	20	80	30	0.25	1.167	80	48.96	50.86	50.76	1.037	66.8	71.29	68.1	1.019	73.05	75.26	1.03
	25	75	25	0.33	1.25	79.96	48.14	50.86	50.77	1.055	66.67	72.23	68.49	1.027	72.82	75.95	1.043
	30	70	20	0.43	1.375	79.99	47.22	50.88	50.83	1.076	66.63	73.35	68.97	1.035	72.75	76.79	1.056
	35	65	15	0.54	1.583	79.94	45.76	50.79	50.92	1.113	66.15	74.42	69.4	1.049	71.96	77.56	1.078
	40	60	10	0.67	2	79.97	43.97	50.82	50.93	1.158	65.57	75.83	70.11	1.069	71.01	78.85	1.11
	45	55	5	0.82	3.25	80.01	40.69	50.83	50.7	1.246	64.06	77.69	70.12	1.095	68.57	78.85	1.15

RE: (Conditional) Relative efficiency based on the relative variances of the flexible adaptive design to the pre-specified adaptive design.
CHW: Power after adjusting for the unplanned adaptation.
Unadj: Power computed based on naïve overall Type 1 error of $\alpha = 0.025$.
Adj: Adjusted power for the overall Type 1 error fixed at $\alpha = 0.025$.
SS‡: Sample size of a FSD based on the power obtained using CHW.
SS$^{\triangle}$: Sample size of a FSD based on the power obtained based on the (Adj)usted test for fixed overall Type 1 error of $\alpha = 0.025$.

Table C.4: Simulation results allowing 50% probability of adaptation to double the sample size of the original design. The original design has either 80% power to detect an alternative of $\theta = 0.396204$, or 90% power to detect an alternative of $\theta = 0.4584195$. Loss of power is observed after applying CHW when unplanned adaptation is made late during the study relative to the prespecified adaptive design with similar probability of adaptation timing.

						n=100	$\widetilde{n} = 200$				ASN = 150				Sample size (FSD)		
	n_1	n_2^*	$\widetilde{n}_2^*$	γ	RE	Orig	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	CHW‡	Unadj	Adj$^{\triangle}$	$\triangle/\ddagger$	SS‡	SS$^{\triangle}$	SS$^{\triangle}$/SS‡
90% Power	10	90	190	0.11	1.053	89.97	99.54	99.57	99.57	1	96.8	96.87	96.65	0.9984	138.4	136.8	0.9888
	20	80	180	0.25	1.111	89.95	99.51	99.57	99.57	1.001	97.6	97.69	97.46	0.9986	147.5	145.8	0.9884
	30	70	170	0.43	1.176	89.92	99.47	99.58	99.58	1.001	98.12	98.26	98.03	0.9991	155.2	153.7	0.9904
	40	60	160	0.67	1.25	89.96	99.41	99.57	99.56	1.002	98.54	98.72	98.54	1	163.1	163.1	0.9998
	50	50	150	1	1.333	90.06	99.32	99.56	99.57	1.003	98.84	99.11	98.99	1.002	170.4	174.6	1.025
	60	40	140	1.5	1.429	90.02	99.2	99.56	99.56	1.004	98.95	99.33	99.25	1.003	173.5	183.6	1.058
	70	30	130	2.3	1.538	90.07	99.02	99.56	99.56	1.006	98.93	99.49	99.45	1.005	172.9	192.7	1.115
	80	20	120	4	1.667	89.92	98.68	99.57	99.57	1.009	98.67	99.56	99.55	1.009	166	199	1.198
	90	10	110	9	1.818	89.99	97.92	99.58	99.58	1.017	97.92	99.58	99.58	1.017	152.1	200.6	1.319
80% Power	10	90	190	0.11	1.053	79.94	97.64	97.74	97.72	1.001	91.7	91.95	91.39	0.9967	142.6	140.9	0.9882
	20	80	180	0.25	1.111	80	97.51	97.74	97.74	1.002	92.85	93.25	92.56	0.9969	149.4	147.6	0.988
	30	70	170	0.43	1.176	79.98	97.37	97.73	97.7	1.003	93.74	94.28	93.56	0.9981	155.5	154.2	0.9918
	40	60	160	0.67	1.25	80.04	97.14	97.73	97.74	1.006	94.41	95.16	94.48	1.001	160.6	161.1	1.003
	50	50	150	1	1.333	80.03	96.9	97.74	97.73	1.009	94.94	95.91	95.25	1.003	165.1	167.8	1.017
	60	40	140	1.5	1.429	80.02	96.58	97.73	97.71	1.012	95.33	96.59	95.95	1.006	168.6	174.8	1.037
	70	30	130	2.33	1.538	79.95	95.99	97.72	97.7	1.018	95.32	97.1	96.62	1.014	168.5	182.7	1.084
	80	20	120	4	1.667	79.99	95.1	97.75	97.72	1.027	94.89	97.46	97.12	1.024	164.6	189.7	1.153
	90	10	110	9	1.818	79.99	93.22	97.74	97.73	1.048	93.21	97.23	97.17	1.043	151.8	190.5	1.255

RE: (Conditional) Relative efficiency based on the relative variances of the flexible adaptive design to the pre-specified adaptive design.
CHW: Power after adjusting for the unplanned adaptation.
Unadj: Power computed based on naïve overall Type 1 error of $\alpha = 0.025$.
Adj: Adjusted power for the overall Type 1 error fixed at $\alpha = 0.025$.
SS‡: Sample size of a FSD based on the power obtained using CHW.
SS$^{\triangle}$: Sample size of a FSD based on the power obtained based on the (Adj)usted test for fixed overall Type 1 error of $\alpha = 0.025$.

C.2 Additional Results for Section 3.3

C.2.1 Optimal Three Stage Designs

In this section, we evaluate and describe the operating characteristics of the GSDs when $J = 3$. We evaluate numerically the operating characteristics by considering the class of one-sided symmetric designs, two-sided symmetric designs, and one-sided asymmetric designs. Contour plots for the minimum ASN based on a grid of possible combinations of schedule of interim analyses are shown for specific fixed P of interest (Figure C.4, C.5 and C.6 respectively). We first characterized the ASN for each fixed value of P when presuming an equally spaced analyses that is most common at design stage, as described by the blue line in Figure C.3. We then characterized the minimum ASN among all possible spacings of the interim analyses for each fixed value of P as shown by the black lines in Figure C.3.

For each fixed P, we note that the schedule of interim analyses for each optimal design, among the class of one-sided symmetric designs, is typically conducted earlier than an equally spaced design. This is observed similarly for the class of two-sided symmetric designs. In Table C.5, the best one-sided symmetric designs tends to be optimal when the schedule of interim analyses is conducted earlier (27.5% and 53% of the maximum statistical information) relative to an equally spaced design (with interim analyses conducted at 33.3% and 67% of the maximum statistical information). This best optimal design has a minimum ASN of 3685, a 40% reduction in ASN relative to the FSD, and has a P parameter defined in the unified family that is close to the Pocock class of designs. Results for the class of two-sided symmetric designs are similar.

The class of three-stage hybrid designs, with a fixed OBF efficacy parameter, tends to have the first interim analysis conducted at more than 1/3 of the way through the study. The second interim analysis is then conducted close to 2/3 of the way through the study. These schedules of interim analysis are rather similar across fixed values of P. The inflation of the maximum statistical information relative to the FSD is at most 1.14% among the choices explored. Additionally, relative to the best one-sided symmetric three-stage OBF design, the

schedules of analyses obtained for each P within these choices of hybrid design are similar.

Table C.5: Optimal spacing of analysis for common designs P for the one-sided, two-sided, and asymmetric group sequential designs under the unified family with a total of three analyses, assuming $\sigma = 1$. A fixed sample design requires $N = 6146$ under the same alternative of $\theta = 0.1$.

		Equally Spaced			By P					By Interim $\mathcal{I}_1^P$			
	P	ASN	Max	$\frac{\text{Max}}{N}$	ASN^P	Max^P	$\frac{\text{Max}^P}{N}$	$\mathcal{I}_1^P$	$\mathcal{I}_2^P$	P_{Opt}	ASN_{Opt}	Max_{Opt}	$\frac{\text{Max}_{\text{Opt}}}{N}$
One-Sided	0.5	3801	8123	1.32	3685	8315	1.35	270	525	0.50	3685	8315	1.35
	0.6	3799	7412	1.21	3713	7427	1.21	305	550	0.53	3694	7962	1.30
	0.7	3880	6936	1.13	3790	6924	1.13	350	580	0.58	3734	7546	1.23
	0.8	4023	6633	1.08	3893	6636	1.08	390	610	0.63	3793	7252	1.18
	0.9	4202	6448	1.05	4006	6466	1.05	435	635	0.69	3877	6999	1.14
	1	4382	6337	1.03	4118	6361	1.03	470	660	0.74	3956	6848	1.11
	Opt				3685	8199	1.33	275	530	0.51			
Two-Sided	0.5	3648	6953	1.13	3593	7021	1.14	300	565	0.45	3583	7229	1.18
	0.6	3713	6677	1.09	3656	6695	1.09	325	575	0.46	3595	7147	1.16
	0.7	3834	6487	1.06	3756	6492	1.06	360	595	0.48	3630	7026	1.14
	0.8	3999	6362	1.04	3873	6371	1.04	405	620	0.52	3703	6861	1.12
	0.9	4187	6283	1.02	3994	6296	1.02	445	645	0.56	3788	6739	1.10
	1	4372	6235	1.01	4112	6250	1.02	480	670	0.60	3876	6646	1.08
	Opt				3582	7294	1.19	290	560	0.44			
Hybrid (Fut, Eff)	(0.5, 1)	4584	7115	1.16	4244	7025	1.14	460	645	1.28	4115	6264	1.02
	(0.6, 1)	4502	6820	1.11	4199	6807	1.11	460	650	1.28	4114	6267	1.02
	(0.7, 1)	4447	6614	1.08	4166	6640	1.08	465	655	1.32	4111	6265	1.02
	(0.8, 1)	4413	6478	1.05	4143	6514	1.06	465	655	1.32	4111	6265	1.02
	(0.9, 1)	4584	7115	1.16	4127	6426	1.05	470	660	1.37	4110	6262	1.02
	Opt				4109	6267	1.02	480	670	1.38			

Equally spaced: GSD with equally spaced analysis obtained for each parameter P.
By ^{P} : For each P, the GSD with the schedule of analyses that attains the minimum ASN is obtained. Using this favorable interim analysis, we then search for the design parameter P that can minimize this ASN^P.
Opt: Optimum GSD with schedule of interim analyses chosen to minimize ASN under class $P > 0$

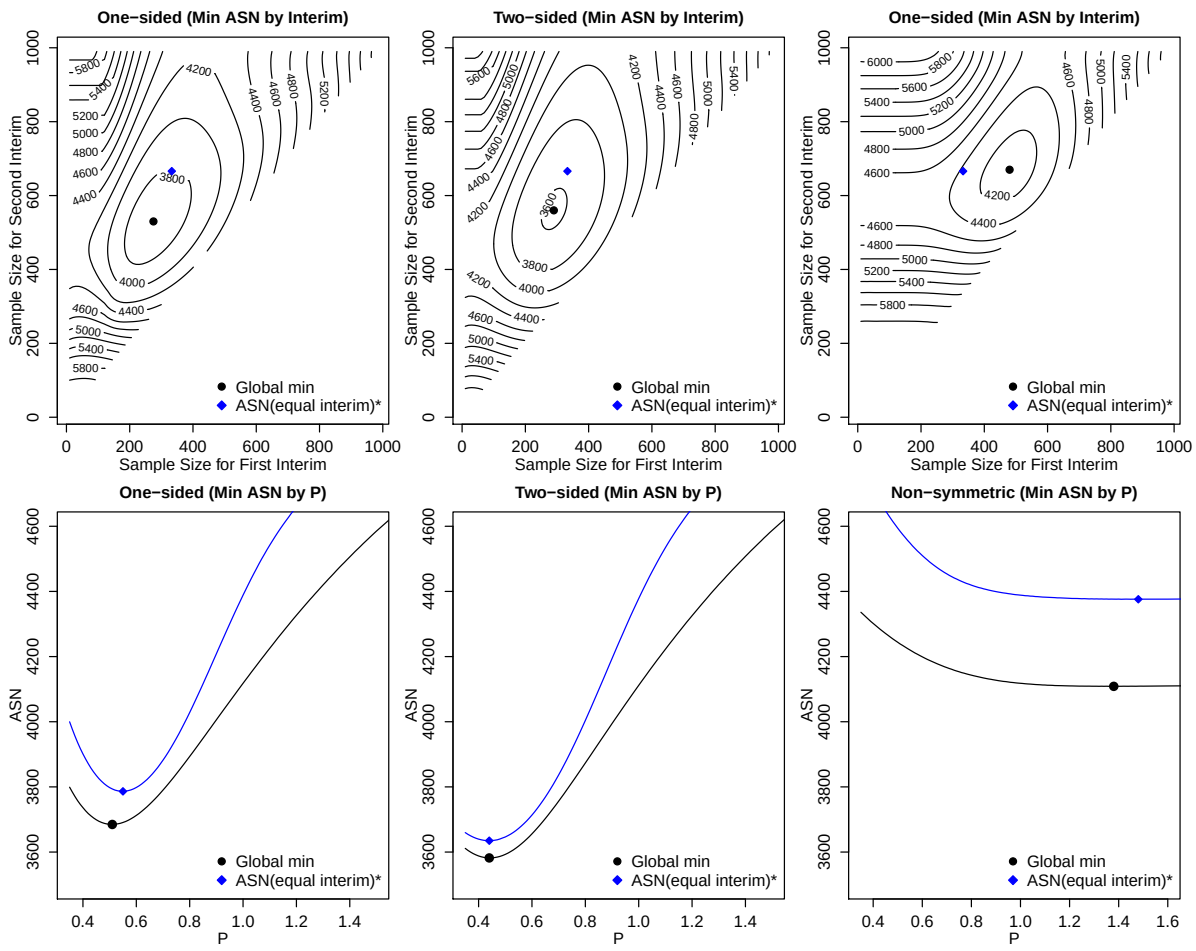

Figure C.3: Top row: Contour plot of minimum ASN for the first vs the second interim analysis. The black point corresponds to the minimum ASN similar to the adjacent plot on the left. The blue point corresponds to the minimum ASN for a monitoring rule with equal information. This point corresponds to the P for the adjacent plot on the left. Bottom row: Plot of best ASN for each P (black line). The blue line corresponds to the ASN line when assuming an equally spaced statistical information monitoring for the various P.

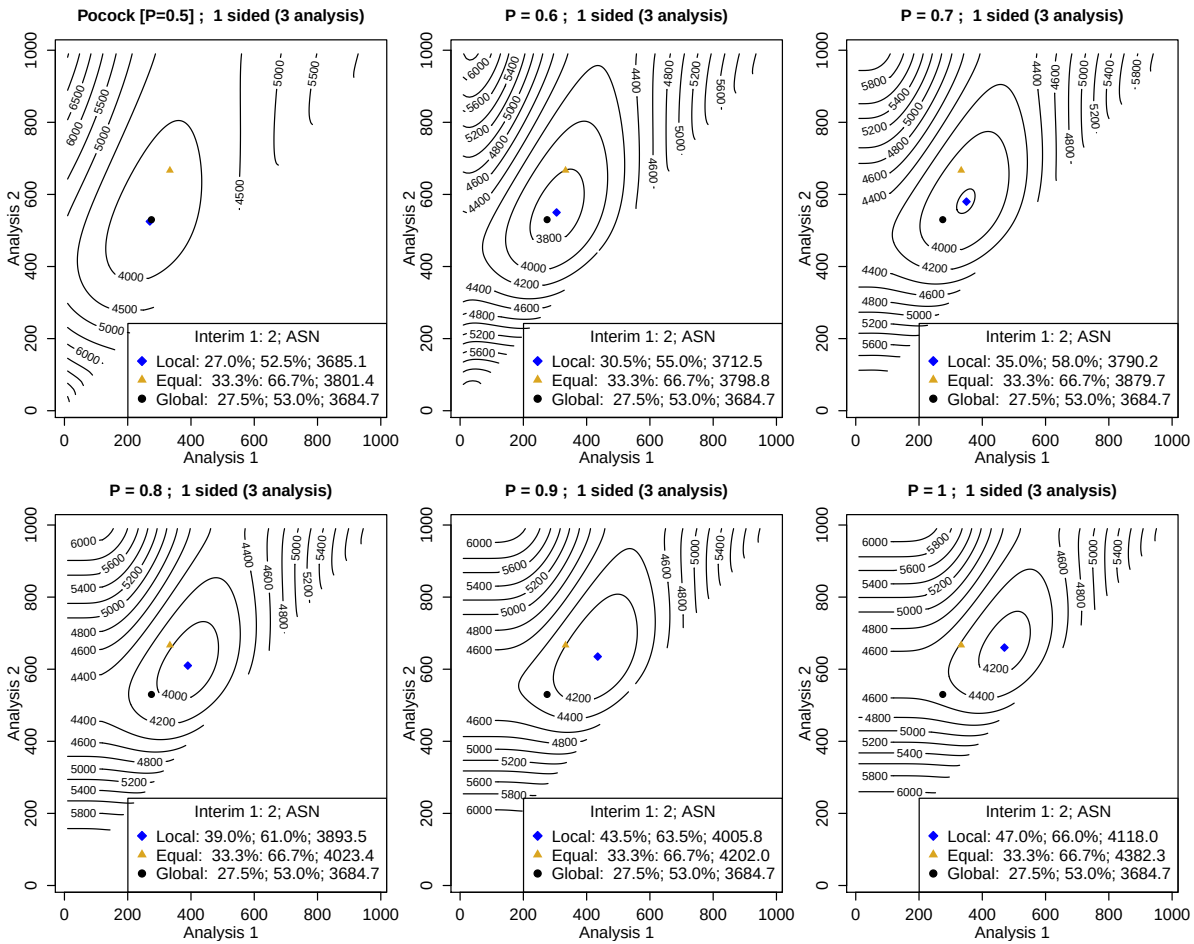

Figure C.4: Optimal ASN for three-stage, one-sided, symmetric boundaries at level $\alpha = 0.025$, and power of 97.5% to detect the design alternative of 0.1, and known variance 1. The global minimum ASN is close to the local minimum ASN when we choose a Pocock boundary. This global minimum moves further away from this local minimum as we becoming increasing conservative at earlier interim analyses, i.e., P increases. The designs with equally spaced analyses tend to have higher ASN relative to the designs with optimized schedule of interim analyses either globally or locally for the respective P.
Global optimum in black circle for $P = 0.51$; Local optimum for P is shown in blue diamond; Equal information monitoring for the P is shown as a gold triangle.

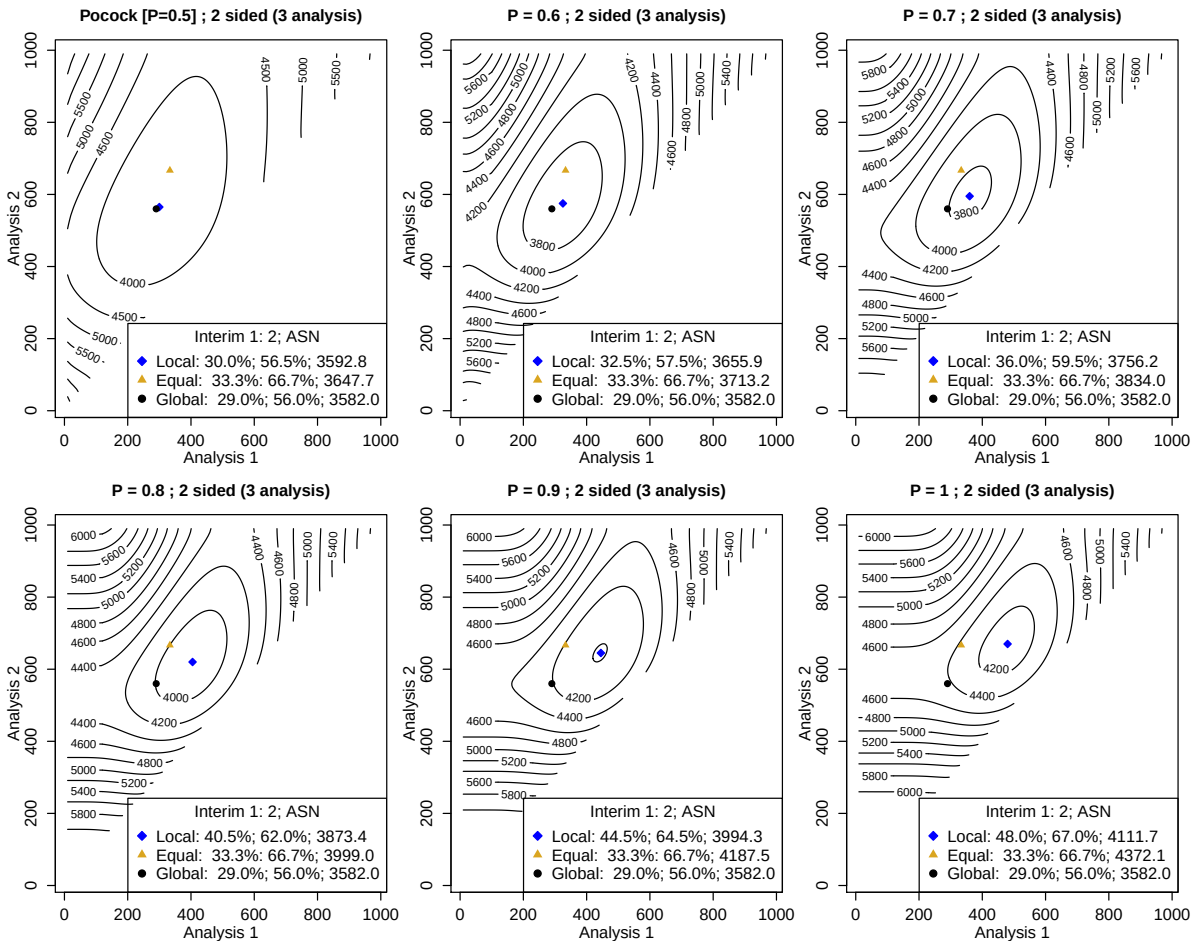

Figure C.5: Optimal ASN for three-stage, two-sided, symmetric boundaries at level $\alpha = 0.025$, and power of 97.5% to detect the design alternative of 0.1, and known variance 1. The global minimum ASN is close to the local minimum ASN when we choose a Pocock boundary. This global minimum moves further away from this local minimum as we becoming increasing conservative at earlier interim analyses, i.e., P increases. The designs with equally spaced analyses tend to have higher ASN relative to the designs with optimized schedule of interim analyses either globally or locally for the respective P.
Global optimum in black circle for $P = 0.44$; Local optimum for P is shown in blue diamond; Equal information monitoring for the P is shown as a gold triangle.

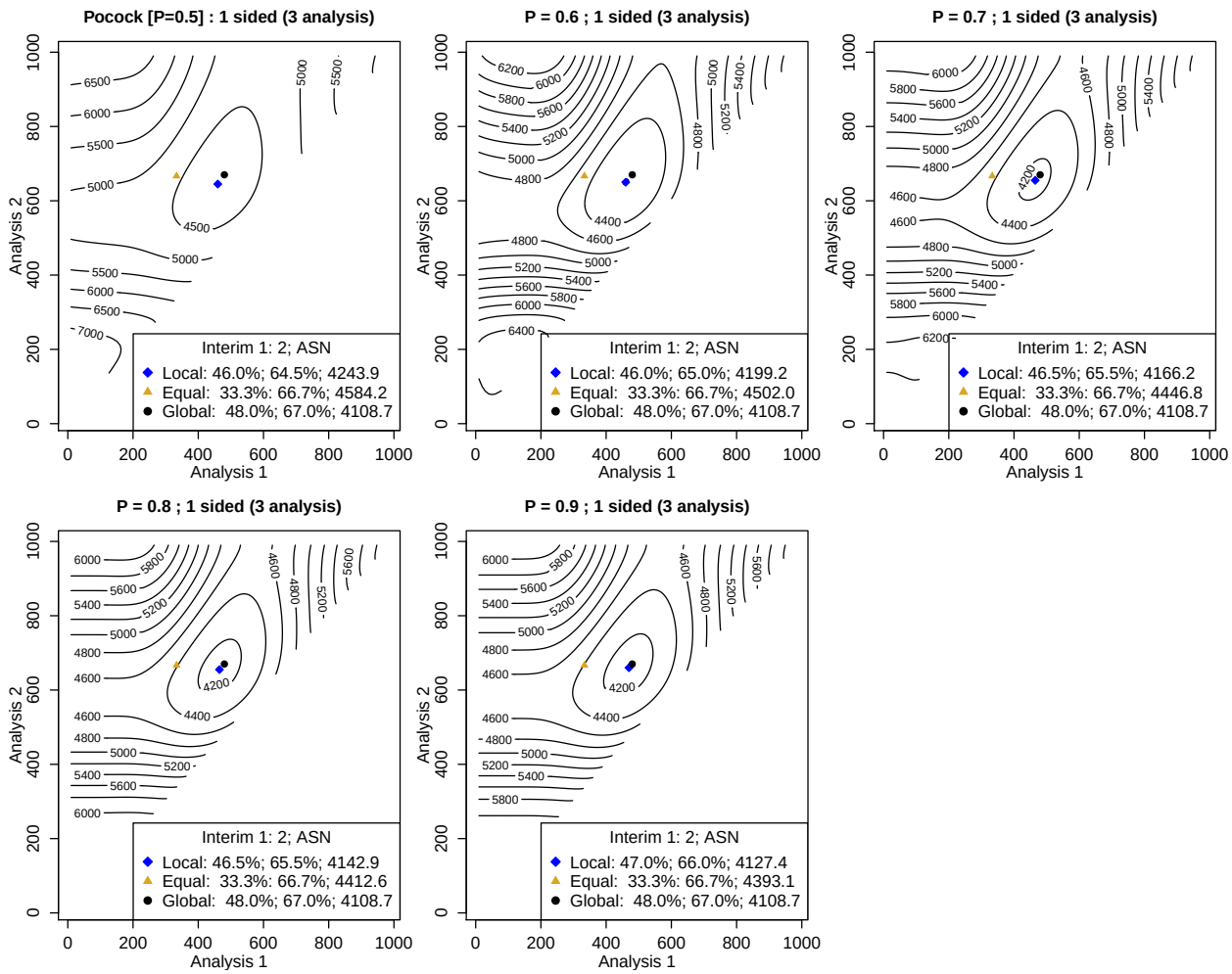

Figure C.6: Optimal ASN for three stage, one-sided, asymmetric designs with OBF efficacy boundaries at level $\alpha = 0.025$, and power of 97.5% to detect the design alternative of 0.1, and known variance 1. There is little difference between the global ASN and local ASN. However, designs with equally spaced analyses averaged higher ASN.
Global optimum in black circle for $P = 1.38$; Local optimum for the P is shown in blue diamond; Equal information monitoring for the P is shown as a gold triangle.

Appendix D

Additional Results for Chapter 5

D.1 Miscellaneous Results

Several competing well-understood designs have been considered to provide references so that we can compare the operating characteristics with the adaptive strategies. We evaluated the operating characteristics (average calendar time of stopping, average event size, average accrual size, and overall power) of the fixed sample design (*FSDInf*) for various combinations of event rates, and design alternative without the maximum calendar time restriction. We evaluated the ideal operating characteristics of the competing group sequential designs based on either the O'Brien Fleming (*GSDOBFInf*) design, or hybrid design (*GSDHYBInf*) similarly without imposing the maximum calendar time restriction.

We then impose calendar time as a constraint for stopping the study. This allows us to evaluate the operating characteristics of the fixed sample designs (*FSD078* and *FSD117*), group sequential designs presuming original sample size (*GSDOBF078*, *GSDHYB078*, *GSDOBG117*, *GSDHYB117*), and group sequential designs incorporating the strategy of blinded adaptation (either continue or restart accrual), and escape clause. These operating characteristics obtained from the FSDs (*FSD078* and *FSD117*), and GSDs (*GSDOBF078*, *GSDHYB078*, *GSDOBG117*, *GSDHYB117*) provide some form of reference to enable us to understand the best power one can obtain in the ideal setting. Note that additional group sequential designs that are planned with 3,500 subjects from the start of the study with accrual patterns similar to continuing accrual, or restarting accrual later in the study can also provide benchmarks on the best possible power one can obtain.

Table D.1: Table of the overall power for the various fixed sample designs (FSDs) and group sequential designs (OBF and Hybrid) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		N = 1750									N = 3500							
		No additional accrual									Cont		Restart		Cont		Restart	
		∞			78			117			78				117			
		FSD	OBF	HYB	FSD	OBF	HYB	FSD	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB
$\theta = 0.04$	$\lambda_0/8$	100	100	100	99.72	99.72	99.72	100	100	100	100	100	99.96	99.96	100	100	100	100
	$\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.1$	$\lambda_0/8$	100	100	100	98.15	98.15	98.15	99.91	99.91	99.91	99.99	99.99	99.63	99.63	100	100	100	100
	$\lambda_0/4$	100	100	100	99.99	99.99	99.99	100	100	100	100	100	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.25$	$\lambda_0/8$	100	100	100	84.59	84.59	84.59	96.04	96.04	96.04	97.33	97.33	92.8	92.8	99.94	99.94	99.68	99.68
	$\lambda_0/4$	100	100	100	98.75	98.75	98.75	99.98	99.98	99.98	100	100	99.8	99.8	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.5$	$\lambda_0/8$	99.93	99.81	99.71	41.39	41.39	41.39	58.44	58.44	58.44	62.97	62.97	51.34	51.34	83.89	83.89	77.29	77.29
	$\lambda_0/4$	99.93	99.8	99.73	68.69	68.69	68.69	87.09	87.09	87.09	90.08	90.08	80.27	80.27	98.75	98.73	96.95	96.96
	$\lambda_0/2$	99.96	99.82	99.75	93.5	93.5	93.5	99.16	99.14	99.14	99.49	99.45	97.84	97.83	99.77	99.68	99.79	99.68
	$3\lambda_0/4$	99.96	99.82	99.72	98.92	98.9	98.89	99.92	99.81	99.71	99.79	99.71	99.72	99.62	99.79	99.71	99.84	99.74
	λ_0	99.95	99.81	99.71	99.83	99.75	99.66	99.95	99.81	99.71	99.82	99.74	99.81	99.73	99.82	99.74	99.81	99.73
$\theta = \theta_A$	$\lambda_0/8$	91.85	89.66	88.91	22.53	22.53	22.53	32.63	32.63	32.63	35.89	35.89	28.19	28.19	53.45	53.45	46.85	46.85
	$\lambda_0/4$	91.86	89.74	89.1	39.88	39.88	39.88	57.41	57.41	57.41	61.95	61.95	49.68	49.68	82.82	82.62	76.38	76.38
	$\lambda_0/2$	92	89.7	88.99	67.61	67.55	67.54	84.74	84.31	84.07	88.4	87.82	79.47	79.41	90.22	89.28	90.41	89.53
	$3\lambda_0/4$	91.96	89.8	89.01	83.15	82.89	82.74	91.95	89.81	89.01	89.83	88.93	90.06	89.41	89.83	88.93	90.4	89.6
	λ_0	92	89.76	89.04	91.27	89.54	88.87	92	89.76	89.04	89.96	89.19	89.88	89.12	89.96	89.19	89.88	89.12
$\theta = 0.75$	$\lambda_0/8$	57.14	53.18	52.66	12.15	12.15	12.15	16.85	16.85	16.85	18.04	18.04	14.92	14.92	27.34	27.34	23.57	23.57
	$\lambda_0/4$	57.13	53.33	52.44	19.91	19.91	19.91	29.09	29.08	29.08	32.05	32.05	24.98	24.98	48.15	47.9	42.04	41.98
	$\lambda_0/2$	57.36	53.45	52.7	35.76	35.73	35.73	50.68	49.85	49.44	53.08	52.27	44	43.95	54.15	53.2	54.37	53.02
	$3\lambda_0/4$	57.35	53.5	52.65	49.18	48.65	48.32	57.35	53.5	52.65	53.85	52.97	53.89	53.13	53.85	52.97	53.94	53.16
	λ_0	57.2	53.54	52.81	56.93	53.54	52.81	57.2	53.54	52.81	54.03	52.96	53.57	52.54	54.03	52.96	53.57	52.54
$\theta = 1$	$\lambda_0/8$	2.58	2.54	2.51	2.52	2.52	2.52	2.64	2.64	2.64	2.6	2.6	2.75	2.75	2.67	2.67	2.59	2.59
	$\lambda_0/4$	2.58	2.52	2.48	2.36	2.36	2.36	2.52	2.52	2.52	2.42	2.42	2.25	2.25	2.73	2.64	2.71	2.73
	$\lambda_0/2$	2.59	2.59	2.59	2.42	2.41	2.4	2.47	2.51	2.47	2.75	2.7	2.48	2.49	2.76	2.71	2.69	2.63
	$3\lambda_0/4$	2.73	2.54	2.58	2.56	2.58	2.6	2.73	2.54	2.58	2.75	2.71	2.54	2.49	2.75	2.71	2.54	2.49
	λ_0	2.67	2.56	2.56	2.67	2.56	2.56	2.67	2.56	2.56	2.84	2.78	2.55	2.53	2.84	2.78	2.55	2.53

Table D.2: Table of the average number of events for the various fixed sample designs (FSDs) and group sequential designs (OBF and Hybrid) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		N = 1750									N = 3500							
		No additional accrual									Cont		Restart		Cont		Restart	
		∞			78			117			78				117			
		FSD	OBF	HYB	FSD	OBF	HYB	FSD	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB
$\theta = 0.04$	$\lambda_0/8$	220	44	44	19	19	19	29	29	29	32	32	24	24	44	44	43	43
	$\lambda_0/4$	220	44	44	37	37	37	57	44	44	44	44	44	44	44	44	44	44
	$\lambda_0/2$	220	44	44	72	44	44	111	44	44	44	44	44	44	44	44	44	44
	$3\lambda_0/4$	220	44	44	106	44	44	161	44	44	44	44	44	44	44	44	44	44
	λ_0	220	44	44	139	44	44	207	44	44	44	44	44	44	44	44	44	44
$\theta = 0.1$	$\lambda_0/8$	220	45	45	20	20	20	31	31	31	34	34	26	26	46	45	44	44
	$\lambda_0/4$	220	45	45	39	39	39	61	45	45	46	45	45	44	46	45	46	45
	$\lambda_0/2$	220	45	45	77	45	45	118	45	45	46	45	45	45	46	45	45	45
	$3\lambda_0/4$	220	46	45	113	45	45	171	45	45	46	45	45	45	46	45	46	45
	λ_0	220	46	45	148	46	45	216	46	45	46	45	46	45	46	45	46	45
$\theta = 0.25$	$\lambda_0/8$	220	61	61	22	22	22	35	35	35	39	39	29	29	59	57	52	51
	$\lambda_0/4$	220	62	61	44	44	44	69	59	59	62	60	54	54	63	61	62	61
	$\lambda_0/2$	220	62	61	87	62	61	134	62	61	62	61	62	61	63	61	62	61
	$3\lambda_0/4$	220	62	61	129	62	61	196	62	61	62	61	62	61	62	61	62	61
	λ_0	220	62	61	169	62	61	220	62	61	62	61	62	61	62	61	62	61
$\theta = 0.5$	$\lambda_0/8$	220	104	103	27	27	27	42	42	42	47	47	35	35	75	75	65	65
	$\lambda_0/4$	220	104	103	53	53	53	83	80	79	87	86	69	69	103	102	100	99
	$\lambda_0/2$	220	104	104	105	93	92	161	103	103	104	103	102	101	105	103	104	104
	$3\lambda_0/4$	220	104	104	155	103	102	219	104	104	105	104	104	103	105	104	104	104
	λ_0	220	104	104	202	104	104	220	104	104	105	104	104	104	105	104	104	104
$\theta = \theta_A$	$\lambda_0/8$	220	143	141	29	29	29	46	46	46	51	51	38	38	83	83	71	71
	$\lambda_0/4$	220	143	141	58	58	58	90	89	89	99	98	76	76	135	132	125	124
	$\lambda_0/2$	220	143	141	114	109	108	175	138	136	141	139	129	128	143	140	143	141
	$3\lambda_0/4$	220	143	141	168	136	134	220	143	141	143	140	143	140	143	140	143	141
	λ_0	220	143	141	215	143	141	220	143	141	144	141	143	140	144	141	143	140
$\theta = 0.75$	$\lambda_0/8$	220	163	156	31	31	31	49	49	49	55	55	41	41	89	89	76	76
	$\lambda_0/4$	220	163	156	62	62	62	96	95	95	107	105	81	81	152	146	139	135
	$\lambda_0/2$	220	164	156	122	119	116	187	157	151	161	154	145	140	163	155	163	156
	$3\lambda_0/4$	220	163	156	180	154	148	220	163	156	163	155	163	156	163	155	163	156
	λ_0	220	163	156	219	163	156	220	163	156	163	156	163	156	163	156	163	156
$\theta = 1$	$\lambda_0/8$	220	125	111	36	36	36	56	56	55	62	61	47	47	97	90	85	81
	$\lambda_0/4$	220	125	111	71	70	69	110	101	94	109	99	90	84	125	110	122	109
	$\lambda_0/2$	220	125	111	139	115	104	210	125	111	125	111	123	110	125	111	125	111
	$3\lambda_0/4$	220	125	111	203	125	111	220	125	111	125	111	125	111	125	111	125	111
	λ_0	220	125	111	220	125	111	220	125	111	125	111	125	111	125	111	125	111

Table D.3: Table of the average calendar time of stopping for the various fixed sample designs (FSDs), and group sequential designs (OBF and Hybrid) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		N = 1750									N = 3500							
		No additional accrual									Cont		Restart		Cont		Restart	
		∞			78			117			78				117			
		FSD	OBF	HYB	FSD	OBF	HYB	FSD	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB
$\theta = 0.04$	$\lambda_0/8$	926	174	174	78	78	78	117	117	117	78	78	78	78	101	101	113	113
	$\lambda_0/4$	468	92	92	78	78	78	117	93	93	59	59	74	74	59	59	74	74
	$\lambda_0/2$	238	50	50	78	50	50	117	50	50	38	38	50	50	38	38	50	50
	$3\lambda_0/4$	162	37	37	78	37	37	117	37	37	31	31	37	37	31	31	37	37
	λ_0	124	30	30	78	30	30	116	30	30	27	27	30	30	27	27	30	30
$\theta = 0.1$	$\lambda_0/8$	864	169	169	78	78	78	117	117	117	78	78	78	78	98	97	111	110
	$\lambda_0/4$	437	89	89	78	77	77	117	89	89	58	58	73	72	58	58	73	73
	$\lambda_0/2$	223	49	49	78	49	49	117	49	49	38	38	49	48	38	38	49	48
	$3\lambda_0/4$	151	36	36	78	36	36	117	36	36	31	31	36	36	31	31	36	36
	λ_0	116	29	29	78	29	29	114	29	29	27	27	29	29	27	27	29	29
$\theta = 0.25$	$\lambda_0/8$	744	201	199	78	78	78	117	117	117	78	78	78	78	109	106	114	112
	$\lambda_0/4$	377	105	104	78	77	77	117	102	101	66	65	75	75	67	65	81	80
	$\lambda_0/2$	193	57	57	78	57	57	117	57	57	42	42	55	54	42	42	54	54
	$3\lambda_0/4$	132	41	41	78	41	41	117	41	41	34	34	41	41	34	34	41	41
	λ_0	101	33	33	78	33	33	101	33	33	29	29	33	33	29	29	33	33
$\theta = 0.5$	$\lambda_0/8$	611	282	282	78	78	78	117	117	117	78	78	78	78	115	114	117	116
	$\lambda_0/4$	310	146	145	78	78	78	117	113	113	74	74	77	77	85	84	98	98
	$\lambda_0/2$	160	78	77	78	70	70	117	77	77	52	51	66	65	52	52	66	66
	$3\lambda_0/4$	109	55	55	78	54	54	109	55	55	41	40	52	52	41	40	52	52
	λ_0	84	44	43	78	43	43	84	44	43	35	35	43	43	35	35	43	43
$\theta = \theta_A$	$\lambda_0/8$	560	360	354	78	78	78	117	117	117	78	78	78	78	116	116	117	117
	$\lambda_0/4$	284	185	182	78	78	78	117	116	115	77	76	78	78	98	97	108	107
	$\lambda_0/2$	147	97	95	78	75	74	117	93	92	60	59	72	71	61	60	76	75
	$3\lambda_0/4$	101	68	66	78	64	64	101	68	66	47	46	60	60	47	46	60	60
	λ_0	78	53	52	76	53	52	78	53	52	39	39	51	50	39	39	51	50
$\theta = 0.75$	$\lambda_0/8$	522	384	367	78	78	78	117	117	117	78	78	78	78	117	116	117	117
	$\lambda_0/4$	266	197	188	78	78	78	117	116	115	77	76	78	78	103	100	111	108
	$\lambda_0/2$	137	103	98	78	76	75	117	99	95	63	61	74	72	64	61	79	76
	$3\lambda_0/4$	95	72	69	78	68	66	95	72	69	48	47	63	61	48	47	63	61
	λ_0	73	56	54	73	56	54	73	56	54	41	40	53	51	41	40	53	51
$\theta = 1$	$\lambda_0/8$	458	258	229	78	78	78	117	117	116	78	77	78	78	111	105	115	111
	$\lambda_0/4$	233	134	119	78	78	76	117	109	101	71	66	76	74	79	72	93	86
	$\lambda_0/2$	121	71	64	78	66	60	116	71	64	48	45	62	57	49	45	62	57
	$3\lambda_0/4$	84	51	46	78	50	46	84	51	46	38	36	49	44	38	36	49	44
	λ_0	65	40	37	65	40	37	65	40	37	33	31	40	36	33	31	40	36

Table D.4: Table of the average sample/accrual size for the various fixed sample designs (FSDs), and group sequential designs (OBF and Hybrid) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		N = 1750									N = 3500							
		No additional accrual									Cont		Restart		Cont		Restart	
		∞			78			117			78				117			
		FSD	OBF	HYB	FSD	OBF	HYB	FSD	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB	OBF	HYB
$\theta = 0.04$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3479	3479	3500	3500	3479	3479
	$\lambda_0/2$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3467	3467	2053	2052	3467	3467	2053	2052
	$3\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3043	3043	1751	1751	3043	3043	1751	1751
	λ_0	1750	1750	1750	1750	1750	1750	1750	1750	1750	2641	2641	1750	1750	2641	2641	1750	1750
$\theta = 0.1$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3460	3459	3500	3500	3460	3459
	$\lambda_0/2$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3446	3445	2019	2001	3446	3445	2015	2000
	$3\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	2996	2987	1763	1761	2998	2988	1764	1761
	λ_0	1750	1750	1750	1750	1750	1750	1750	1750	1750	2610	2600	1750	1750	2610	2600	1750	1750
$\theta = 0.25$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3459	3453	3500	3500	3459	3453
	$\lambda_0/2$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3459	3453	2500	2467	3460	3454	2497	2467
	$3\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3224	3199	1813	1810	3224	3200	1814	1811
	λ_0	1750	1750	1750	1750	1750	1750	1750	1750	1750	2847	2826	1752	1752	2847	2826	1752	1752
$\theta = 0.5$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3497	3495	3500	3500	3496	3495
	$\lambda_0/2$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3497	3496	3120	3110	3497	3496	3121	3111
	$3\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3428	3422	2337	2328	3428	3422	2339	2328
	λ_0	1750	1750	1750	1750	1750	1750	1750	1750	1750	3234	3223	1924	1918	3234	3223	1924	1918
$\theta = \theta_A$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3499	3498	3500	3500	3499	3498
	$\lambda_0/2$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3499	3498	3350	3336	3499	3498	3351	3336
	$3\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3475	3469	2842	2807	3475	3469	2843	2807
	λ_0	1750	1750	1750	1750	1750	1750	1750	1750	1750	3388	3375	2301	2267	3388	3375	2301	2267
$\theta = 0.75$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3499	3494	3500	3500	3499	3494
	$\lambda_0/2$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3499	3494	3410	3346	3499	3494	3411	3346
	$3\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3484	3462	3025	2920	3484	3462	3025	2920
	λ_0	1750	1750	1749	1750	1750	1749	1750	1750	1749	3431	3384	2437	2355	3431	3384	2437	2355
$\theta = 1$	$\lambda_0/8$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3500	3500	3497	3500	3500	3500	3497
	$\lambda_0/4$	1750	1750	1750	1750	1750	1750	1750	1750	1750	3500	3497	3480	3394	3500	3497	3479	3394
	$\lambda_0/2$	1750	1750	1749	1750	1750	1749	1750	1750	1749	3482	3418	2919	2659	3482	3418	2921	2659
	$3\lambda_0/4$	1750	1750	1747	1750	1750	1747	1750	1750	1747	3350	3193	2200	2083	3350	3193	2200	2083
	λ_0	1750	1749	1734	1750	1749	1734	1750	1749	1734	3105	2920	1864	1833	3105	2920	1864	1833

Table D.5: Table of the percentage of adaptations for the various monitoring rules (OBF and HYB for O'Brien Fleming and Hybrid design respectively) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates. Stay means continue the design with only 1,750 subjects while Adapt means increase accrual to 3,500 subjects

		78 Months								117 Months							
		Continue				Restart				Continue				Restart			
		OBF		Hybrid		OBF		Hybrid		OBF		Hybrid		OBF		Hybrid	
		Stay	Adapt	Stay	Adapt	Stay	Adapt	Stay	Adapt	Stay	Adapt	Stay	Adapt	Stay	Adapt	Stay	Adapt
		1750	3500	1750	3500	1750	3500	1750	3500	1750	3500	1750	3500	1750	3500	1750	3500
$\theta = 0.04$	$\lambda_0/8$	0	91.6	0	91.6	0	100	0	100	0	91.6	0	91.6	0	100	0	100
	$\lambda_0/4$	0	99.3	0	99.3	0	100	0	100	0	99.3	0	99.3	0	100	0	100
	$\lambda_0/2$	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100
	$3\lambda_0/4$	2.4	97.6	2.4	97.6	0	100	0	100	2.4	97.6	2.4	97.6	0	100	0	100
	λ_0	23.2	76.8	23.2	76.8	0	100	0	100	23.2	76.8	23.2	76.8	0	100	0	100
$\theta = 0.1$	$\lambda_0/8$	0	92.8	0	92.8	0	100	0	100	0	92.8	0	92.8	0	100	0	100
	$\lambda_0/4$	0	99.4	0	99.4	0	100	0	100	0	99.4	0	99.4	0	100	0	100
	$\lambda_0/2$	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100
	$3\lambda_0/4$	4.2	95.8	4.2	95.8	0	100	0	100	4.2	95.8	4.2	95.8	0	100	0	100
	λ_0	32.1	67.9	32.1	67.9	0.4	99.6	0.4	99.6	32.1	67.9	32.1	67.9	0.4	99.6	0.4	99.6
$\theta = 0.25$	$\lambda_0/8$	0	94.9	0	94.9	0	100	0	100	0	94.9	0	94.9	0	100	0	100
	$\lambda_0/4$	0	99.8	0	99.8	0	100	0	100	0	99.8	0	99.8	0	100	0	100
	$\lambda_0/2$	0.2	99.8	0.2	99.8	0	100	0	100	0.2	99.8	0.2	99.8	0	100	0	100
	$3\lambda_0/4$	11.6	88.4	11.6	88.4	0	100	0	100	11.6	88.4	11.6	88.4	0	100	0	100
	λ_0	55.8	44.2	55.8	44.2	10.4	89.6	10.4	89.6	55.8	44.2	55.8	44.2	10.4	89.6	10.4	89.6
$\theta = 0.5$	$\lambda_0/8$	0	97.2	0	97.2	0	100	0	100	0	97.2	0	97.2	0	100	0	100
	$\lambda_0/4$	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100
	$\lambda_0/2$	1.8	98.2	1.8	98.2	0	100	0	100	1.8	98.2	1.8	98.2	0	100	0	100
	$3\lambda_0/4$	36.1	63.9	36.1	63.9	1.1	98.9	1.1	98.9	36.1	63.9	36.1	63.9	1.1	98.9	1.1	98.9
	λ_0	86	14	86	14	75.9	24.1	75.9	24.1	86	14	86	14	75.9	24.1	75.9	24.1
$\theta = \theta_A$	$\lambda_0/8$	0	98	0	98	0	100	0	100	0	98	0	98	0	100	0	100
	$\lambda_0/4$	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100
	$\lambda_0/2$	4	96	4	96	0	100	0	100	4	96	4	96	0	100	0	100
	$3\lambda_0/4$	52.6	47.4	52.6	47.4	8.1	91.9	8.1	91.9	52.6	47.4	52.6	47.4	8.1	91.9	8.1	91.9
	λ_0	93.4	6.6	93.4	6.6	95.2	4.8	95.2	4.8	93.4	6.6	93.4	6.6	95.2	4.8	95.2	4.8
$\theta = 0.75$	$\lambda_0/8$	0	98.3	0	98.3	0	100	0	100	0	98.3	0	98.3	0	100	0	100
	$\lambda_0/4$	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100
	$\lambda_0/2$	7.3	92.7	7.3	92.7	0	100	0	100	7.3	92.7	7.3	92.7	0	100	0	100
	$3\lambda_0/4$	65.8	34.2	65.8	34.2	25.4	74.6	25.4	74.6	65.8	34.2	65.8	34.2	25.4	74.6	25.4	74.6
	λ_0	97	3	97	3	99.2	0.8	99.2	0.8	97	3	97	3	99.2	0.8	99.2	0.8
$\theta = 1$	$\lambda_0/8$	0	99.1	0	99.1	0	100	0	100	0	99.1	0	99.1	0	100	0	100
	$\lambda_0/4$	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100
	$\lambda_0/2$	19.1	80.9	19.1	80.9	0	100	0	100	19.1	80.9	19.1	80.9	0	100	0	100
	$3\lambda_0/4$	85.7	14.3	85.7	14.3	77.2	22.8	77.2	22.8	85.7	14.3	85.7	14.3	77.2	22.8	77.2	22.8
	λ_0	99.4	0.6	99.4	0.6	100	0	100	0	99.4	0.6	99.4	0.6	100	0	100	0

Table D.6: Table of the overall power(%) for the hybrid monitoring rule based on a O'Brien Fleming efficacy rule, and a futility rule that is intermediate between the O'Brien Fleming and Pocock rule ($P = 0.8$ using the unified family design in Kittelson and Emerson [1999]) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		78 Months						117 Months					
		Continue			Restart			Continue			Restart		
		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*	
		Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj
$\theta = 0.04$	$\lambda_0/8$	100	100	100	99.96	99.96	99.75	100	100	100	100	100	100
	$\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.1$	$\lambda_0/8$	99.99	99.99	99.97	99.63	99.63	98.5	100	100	100	100	100	100
	$\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.25$	$\lambda_0/8$	97.33	97.33	97.2	92.8	92.8	87.01	99.94	99.71	99.7	99.68	99.68	99.5
	$\lambda_0/4$	100	100	100	99.8	99.8	99.56	100	100	100	100	100	100
	$\lambda_0/2$	100	100	100	100	100	100	100	100	100	100	100	100
	$3\lambda_0/4$	100	100	100	100	100	100	100	100	100	100	100	100
	λ_0	100	100	100	100	100	100	100	100	100	100	100	100
$\theta = 0.5$	$\lambda_0/8$	62.97	62.97	62.03	51.34	51.34	44.88	83.89	83.08	82.88	77.29	77.29	75.75
	$\lambda_0/4$	90.08	90.08	89.73	80.27	80.27	77.08	98.73	98.73	98.73	96.96	96.95	96.95
	$\lambda_0/2$	99.45	99.37	99.37	97.83	97.82	97.82	99.68	99.6	99.6	99.68	99.67	99.67
	$3\lambda_0/4$	99.71	99.42	99.41	99.62	99.62	99.62	99.71	99.42	99.41	99.74	99.74	99.74
	λ_0	99.74	99.71	99.71	99.73	99.71	99.71	99.74	99.71	99.71	99.73	99.71	99.71
$\theta = \theta_A$	$\lambda_0/8$	35.89	35.89	35.33	28.19	28.19	24.88	53.45	52.99	52.72	46.85	46.85	45.78
	$\lambda_0/4$	61.95	61.95	61.87	49.68	49.68	47.62	82.62	82.65	82.59	76.38	76.39	76.45
	$\lambda_0/2$	87.82	87.1	87.1	79.41	79.42	79.5	89.28	88.48	88.45	89.53	89.54	89.54
	$3\lambda_0/4$	88.93	86.12	86.13	89.41	89.21	89.2	88.93	86.12	86.13	89.6	89.38	89.38
	λ_0	89.19	88.36	88.35	89.12	88.68	88.68	89.19	88.36	88.35	89.12	88.68	88.68
$\theta = 0.75$	$\lambda_0/8$	18.04	18.04	18.11	14.92	14.92	13.29	27.34	27.19	27.29	23.57	23.57	23.04
	$\lambda_0/4$	32.05	32.05	31.91	24.98	24.98	23.78	47.9	47.91	47.93	41.98	41.98	42.14
	$\lambda_0/2$	52.27	51.12	51.13	43.95	43.97	43.93	53.2	52.05	52.06	53.02	53.11	53.11
	$3\lambda_0/4$	52.97	50.33	50.29	53.13	52.28	52.29	52.97	50.33	50.29	53.16	52.3	52.31
	λ_0	52.96	52.42	52.42	52.54	52.77	52.77	52.96	52.42	52.42	52.54	52.77	52.77
$\theta = 1$	$\lambda_0/8$	2.6	2.6	2.49	2.75	2.75	2.49	2.67	2.69	2.66	2.59	2.59	2.62
	$\lambda_0/4$	2.42	2.42	2.37	2.25	2.25	2.28	2.64	2.64	2.63	2.73	2.74	2.75
	$\lambda_0/2$	2.7	2.61	2.6	2.49	2.5	2.49	2.71	2.62	2.61	2.63	2.63	2.63
	$3\lambda_0/4$	2.71	2.56	2.56	2.49	2.57	2.57	2.71	2.56	2.56	2.49	2.57	2.57
	λ_0	2.78	2.56	2.56	2.53	2.67	2.61	2.78	2.56	2.56	2.53	2.67	2.61

GSD Ref refers to the GSD with 3500 planned at the beginning.
GSDMod refers the strategy of blinded sample size adaptation conducted at either 18 months (continue accrual), or 48 months (restart accrual). The *GSDMod* has both the interpretation of the prespecified adaptive design (Pres) and fully adaptive design(Adj).

Table D.7: Table of the average event size for the hybrid monitoring rule based on a O'Brien Fleming efficacy rule, and a futility rule that is intermediate between the O'Brien Fleming and Pocock rule ($P = 0.8$ using the unified family design in Kittelson and Emerson [1999]) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		78 Months						117 Months					
		Continue			Restart			Continue			Restart		
		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*	
		Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj
$\theta = 0.04$	$\lambda_0/8$	32	32	32	24	24	24	44	43	43	43	42	42
	$\lambda_0/4$	44	44	44	44	43	43	44	44	44	44	44	44
	$\lambda_0/2$	44	44	44	44	42	42	44	44	44	44	42	42
	$3\lambda_0/4$	44	44	44	44	44	44	44	44	44	44	44	44
	λ_0	44	53	53	44	44	44	44	53	53	44	44	44
$\theta = 0.1$	$\lambda_0/8$	34	34	34	26	26	26	45	44	44	44	44	44
	$\lambda_0/4$	45	45	45	44	44	44	45	45	45	45	45	45
	$\lambda_0/2$	45	45	45	45	46	46	45	45	45	45	46	46
	$3\lambda_0/4$	45	46	46	45	45	45	45	46	46	45	45	45
	λ_0	45	56	56	45	45	45	45	56	56	45	45	45
$\theta = 0.25$	$\lambda_0/8$	39	39	39	29	29	29	57	56	56	51	51	51
	$\lambda_0/4$	60	60	60	54	54	54	61	61	61	61	61	61
	$\lambda_0/2$	61	61	61	61	61	61	61	61	61	61	62	62
	$3\lambda_0/4$	61	62	62	61	62	62	61	62	62	61	62	62
	λ_0	61	66	66	61	61	61	61	66	66	61	61	61
$\theta = 0.5$	$\lambda_0/8$	47	47	47	35	35	35	75	74	74	65	65	65
	$\lambda_0/4$	86	86	86	69	69	69	102	102	102	99	99	99
	$\lambda_0/2$	103	103	103	101	103	103	103	103	103	104	105	105
	$3\lambda_0/4$	104	104	104	103	107	107	104	104	104	104	107	107
	λ_0	104	104	104	104	106	106	104	104	104	104	106	106
$\theta = \theta_A$	$\lambda_0/8$	51	51	51	38	38	38	83	82	82	71	71	71
	$\lambda_0/4$	98	98	98	76	76	76	132	132	132	124	124	124
	$\lambda_0/2$	139	138	138	128	129	129	140	139	139	141	141	141
	$3\lambda_0/4$	140	138	138	140	143	143	140	138	138	141	144	144
	λ_0	141	141	141	140	144	144	141	141	141	140	144	144
$\theta = 0.75$	$\lambda_0/8$	55	55	55	41	41	41	89	88	88	76	76	76
	$\lambda_0/4$	105	105	105	81	81	81	146	146	146	135	135	135
	$\lambda_0/2$	154	151	151	140	141	141	155	152	152	156	156	156
	$3\lambda_0/4$	155	152	152	156	157	157	155	152	152	156	157	157
	λ_0	156	156	156	156	159	159	156	156	156	156	159	159
$\theta = 1$	$\lambda_0/8$	61	61	61	47	47	47	90	90	90	81	81	81
	$\lambda_0/4$	99	99	99	84	84	84	110	110	110	109	109	109
	$\lambda_0/2$	111	110	110	110	112	112	111	110	110	111	113	113
	$3\lambda_0/4$	111	112	112	111	113	113	111	112	112	111	113	113
	λ_0	111	115	115	111	113	113	111	115	115	111	113	113

GSD Ref refers to the GSD with 3500 planned at the beginning.
GSDMod refers the strategy of blinded sample size adaptation conducted at either 18 months (continue accrual), or 48 months (restart accrual). The *GSDMod* has both the interpretation of the prespecified adaptive design (Pres) and fully adaptive design(Adj).

Table D.8: Table of the average calendar time for the hybrid monitoring rule based on a O'Brien Fleming efficacy rule, and a futility rule that is intermediate between the O'Brien Fleming and Pocock rule ($P = 0.8$ using the unified family design in Kittelson and Emerson [1999]) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		78 Months						117 Months					
		Continue			Restart			Continue			Restart		
		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*	
		Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj
$\theta = 0.04$	$\lambda_0/8$	78	78	78	78	78	78	101	98	98	113	112	112
	$\lambda_0/4$	59	59	59	74	73	73	59	59	59	74	74	74
	$\lambda_0/2$	38	38	38	50	48	48	38	38	38	50	48	48
	$3\lambda_0/4$	31	32	32	37	37	37	31	32	32	37	37	37
	λ_0	27	31	31	30	30	30	27	31	31	30	30	30
$\theta = 0.1$	$\lambda_0/8$	78	78	78	78	78	78	97	95	95	110	110	110
	$\lambda_0/4$	58	58	58	72	72	72	58	58	58	73	73	73
	$\lambda_0/2$	38	38	38	48	49	49	38	38	38	48	49	49
	$3\lambda_0/4$	31	31	31	36	36	36	31	31	31	36	36	36
	λ_0	27	31	31	29	29	29	27	31	31	29	29	29
$\theta = 0.25$	$\lambda_0/8$	78	78	78	78	78	78	106	104	104	112	112	112
	$\lambda_0/4$	65	65	65	75	75	75	65	65	65	80	81	81
	$\lambda_0/2$	42	42	42	54	54	54	42	42	42	54	54	54
	$3\lambda_0/4$	34	35	35	41	41	41	34	35	35	41	41	41
	λ_0	29	33	33	33	33	33	29	33	33	33	33	33
$\theta = 0.5$	$\lambda_0/8$	78	78	78	78	78	78	114	113	113	116	116	116
	$\lambda_0/4$	74	74	74	77	77	77	84	84	84	98	98	98
	$\lambda_0/2$	51	52	52	65	66	66	52	52	52	66	67	67
	$3\lambda_0/4$	40	45	45	52	53	53	40	45	45	52	53	53
	λ_0	35	42	42	43	44	44	35	42	42	43	44	44
$\theta = \theta_A$	$\lambda_0/8$	78	78	78	78	78	78	116	115	115	117	117	117
	$\lambda_0/4$	76	76	76	78	78	78	97	97	97	107	107	107
	$\lambda_0/2$	59	60	60	71	71	71	60	61	61	75	75	75
	$3\lambda_0/4$	46	55	55	60	61	61	46	55	55	60	61	61
	λ_0	39	51	51	50	53	53	39	51	51	50	53	53
$\theta = 0.75$	$\lambda_0/8$	78	78	78	78	78	78	116	115	115	117	116	116
	$\lambda_0/4$	76	76	76	78	78	78	100	100	100	108	108	108
	$\lambda_0/2$	61	62	62	72	72	72	61	62	62	76	77	77
	$3\lambda_0/4$	47	59	59	61	63	63	47	59	59	61	63	63
	λ_0	40	53	53	51	55	55	40	53	53	51	55	55
$\theta = 1$	$\lambda_0/8$	77	77	77	78	78	78	105	105	105	111	111	111
	$\lambda_0/4$	66	66	66	74	74	74	72	72	72	86	86	86
	$\lambda_0/2$	45	48	48	57	58	58	45	48	48	57	58	58
	$3\lambda_0/4$	36	44	44	44	46	46	36	44	44	44	46	46
	λ_0	31	38	38	36	37	37	31	38	38	36	37	37

GSD Ref refers to the GSD with 3500 planned at the beginning.
GSDMod refers the strategy of blinded sample size adaptation conducted at either 18 months (continue accrual), or 48 months (restart accrual). The *GSDMod* has both the interpretation of the prespecified adaptive design (Pres) and fully adaptive design(Adj).

Table D.9: Table of the average sample size for the hybrid monitoring rule based on a O'Brien Fleming efficacy rule, and a futility rule that is intermediate between the O'Brien Fleming and Pocock rule ($P = 0.8$ using the unified family design in Kittelson and Emerson [1999]) evaluated under different maximum calendar time of stopping, and different combinations of hazard ratios θ, and baseline event rates.

		78 Months						117 Months					
		Continue			Restart			Continue			Restart		
		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*		GSD	*GSDMod*	
		Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj	Ref	Pres	Adj
$\theta = 0.04$	$\lambda_0/8$	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	3500	3500	3500	3479	3456	3456	3500	3500	3500	3479	3456	3456
	$\lambda_0/2$	3467	3467	3467	2052	1860	1860	3467	3467	3467	2052	1860	1860
	$3\lambda_0/4$	3043	3036	3036	1751	1750	1750	3043	3036	3036	1751	1750	1750
	λ_0	2641	2649	2649	1750	1750	1750	2641	2649	2649	1750	1750	1750
$\theta = 0.1$	$\lambda_0/8$	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	3500	3500	3500	3459	3450	3450	3500	3500	3500	3459	3450	3450
	$\lambda_0/2$	3445	3445	3445	2001	2001	2001	3445	3445	3445	2000	2001	2001
	$3\lambda_0/4$	2987	2980	2980	1761	1751	1751	2988	2980	2980	1761	1751	1751
	λ_0	2600	2561	2561	1750	1750	1750	2600	2561	2561	1750	1750	1750
$\theta = 0.25$	$\lambda_0/8$	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	3500	3500	3500	3453	3458	3458	3500	3500	3500	3453	3458	3458
	$\lambda_0/2$	3453	3451	3451	2467	2430	2430	3454	3451	3451	2467	2460	2460
	$3\lambda_0/4$	3199	3082	3082	1810	1800	1800	3200	3082	3082	1811	1800	1800
	λ_0	2826	2320	2320	1752	1751	1751	2826	2320	2320	1752	1751	1751
$\theta = 0.5$	$\lambda_0/8$	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	3500	3500	3500	3495	3497	3497	3500	3500	3500	3495	3497	3497
	$\lambda_0/2$	3496	3465	3465	3110	3164	3164	3496	3465	3465	3111	3170	3170
	$3\lambda_0/4$	3422	2830	2830	2328	2408	2408	3422	2830	2830	2328	2408	2408
	λ_0	3223	1970	1970	1918	1835	1835	3223	1970	1970	1918	1835	1835
$\theta = \theta_A$	$\lambda_0/8$	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	3500	3500	3500	3498	3498	3498	3500	3500	3500	3498	3498	3498
	$\lambda_0/2$	3498	3428	3428	3336	3359	3359	3498	3428	3428	3336	3361	3361
	$3\lambda_0/4$	3469	2571	2571	2807	2811	2811	3469	2571	2571	2807	2811	2811
	λ_0	3375	1862	1862	2267	1795	1795	3375	1862	1862	2267	1795	1795
$\theta = 0.75$	$\lambda_0/8$	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500	3500
	$\lambda_0/4$	3500	3500	3500	3494	3494	3494	3500	3500	3500	3494	3494	3494
	$\lambda_0/2$	3494	3367	3367	3346	3367	3367	3494	3367	3367	3346	3368	3368
	$3\lambda_0/4$	3462	2340	2340	2920	2707	2707	3462	2340	2340	2920	2707	2707
	λ_0	3384	1801	1801	2355	1759	1759	3384	1801	1801	2355	1759	1759
$\theta = 1$	$\lambda_0/8$	3500	3500	3500	3497	3496	3496	3500	3500	3500	3497	3496	3496
	$\lambda_0/4$	3497	3499	3499	3394	3389	3389	3497	3499	3499	3394	3390	3390
	$\lambda_0/2$	3418	3117	3117	2659	2710	2710	3418	3117	3117	2659	2711	2711
	$3\lambda_0/4$	3193	1972	1972	2083	1871	1871	3193	1972	1972	2083	1871	1871
	λ_0	2920	1758	1758	1833	1749	1749	2920	1758	1758	1833	1749	1749

GSD Ref refers to the GSD with 3500 planned at the beginning.
GSDMod refers the strategy of blinded sample size adaptation conducted at either 18 months (continue accrual), or 48 months (restart accrual). The *GSDMod* has both the interpretation of the prespecified adaptive design (Pres) and fully adaptive design(Adj).

D.2 Results for Setting A1: No Extension of Accrual Size

We considered the setting where the accrual size of the trial is fixed at 1750 and the calendar time of the study to be 78 months. The use of a GSD with the prespecified opportunity of making an adaptive resizing of the maximum statistical information in a blinded manner based on our "escape clause" results in significant loss of statistical power under our design alternative when the baseline event rate is misspecified. At event rates that are markedly lower, i.e., $\lambda \in (\lambda_0/4, \lambda_0/2)$, the overall power under the design alternative decreases to 68%. When the treatment effect is extreme, i.e., $\theta \leq 0.5$, the use of a GSD has high statistical power to stop prior to 78 months. At more extreme treatment effects, the average calendar time of stopping based on the GSD is markedly shorter as compared to a FSD. Under combinations of extreme efficacy and extreme low baseline rate, the overall power of the GSD is on average similar to a FSD.

D.3 Results for Setting A2: Additional Results for Fully Blinded Adaptations

The use of the GSD with the "escape clause" strategy, allowing a blinded revision of sample size when event rates are low, generally improves the overall power under the design alternative. Compared to the optimal design with an accrual size of 3500, the overall power curve is matched at event rates lower than a quarter while there is generally a slight loss in overall power when the baseline event rates are greater than a quarter. This loss of power ($\approx 1\%$) is compensated by an overwhelmingly smaller expected accrual size at $\lambda > \lambda_0/4$, and longer average calendar time of stopping.

At extreme event rates, i.e., $\lambda \leq \lambda_0/8$, we have at least 90% power to detect the extreme treatment effect if the true hazard ratio is less than 0.25. Overall power loss under the design alternative is observed when the hazard ratio is greater than 0.25 across all settings. In presence of extreme efficacy, i.e., when the hazard ratio ≤ 0.5, we have high statistical power ($> 99\%$) to declare efficacy with negligible difference in power when the blinded

adaptation is made later during the study. In cases when the accrual is restarted and the baseline event rate is markedly lower, our statistical power under the design alternative is reduced to 80%.

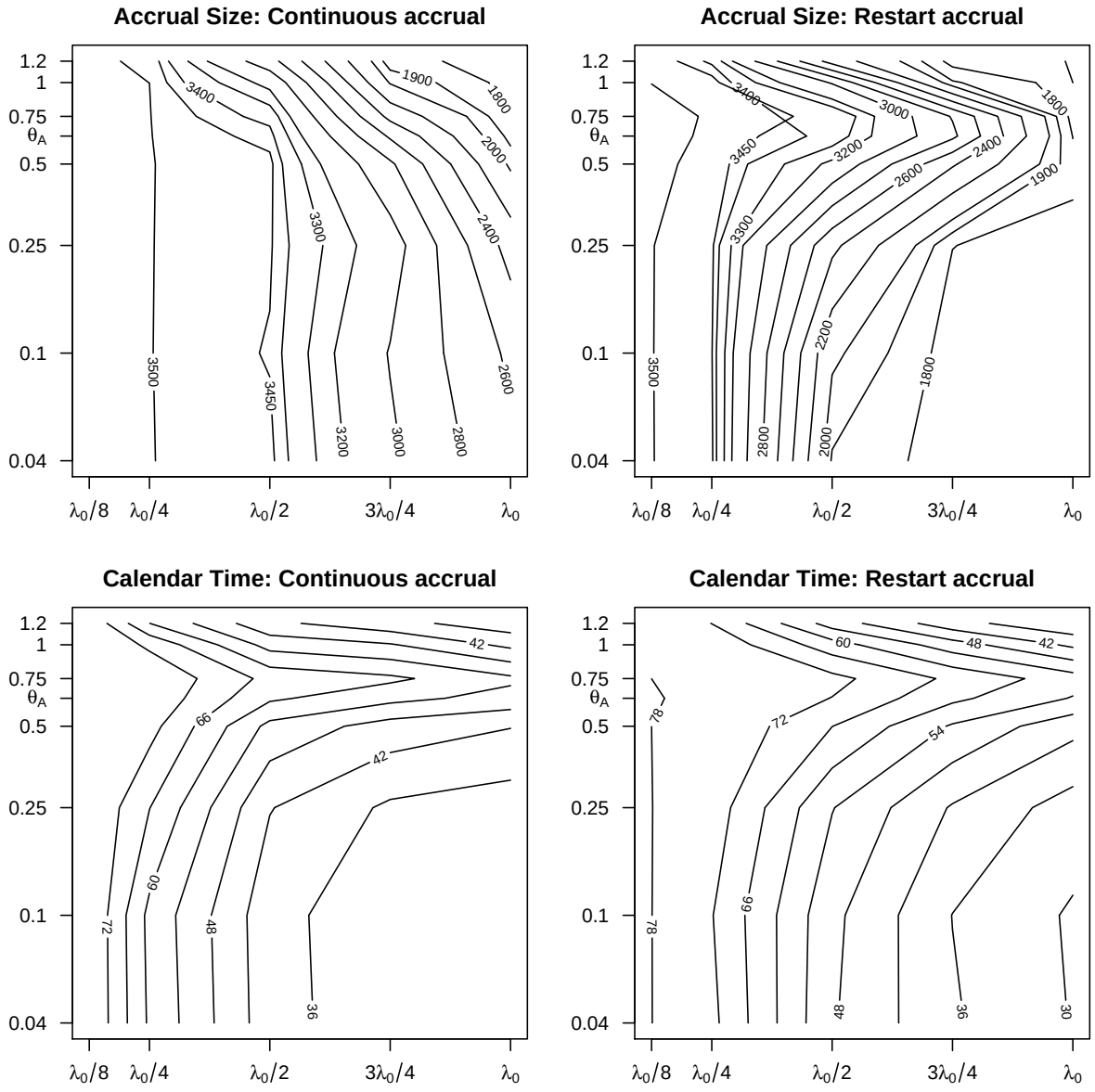

Figure D.1: Contour plots of average accrual size and average calendar time of stopping for hazard ratios vs baseline event rates for continuing or restarting accrual.

We compare the average accrual size required when the sponsor either continue accrual early, or restart accrual later. In summary, by restarting accrual or keeping accrual open as recommended in the statistical literature or guidance material from FDA [2010], and a blinded adaptation at later interim analysis is often more cost effective, in terms of ASN, relative to continuing accrual early. The reduction in cost saving can be substantial when treatment effect is extreme as the monitoring boundaries enable early stopping. When θ is closer to our design alternative, a slightly bigger accrual size is required when the event rate is not far off than planned, i.e., $\lambda \in (\lambda_0/2, \lambda_0)$, as seen in Figure D.1. This in turn leads to a higher probability of a blinded adaptation. One of the disadvantages of keeping accrual open is that this may not be logistically feasible in many clinical or prevention setting.

D.4 Results for Setting B1: No Extension of Accrual Size and Extension of Maximum Calendar Time

Results when allowing the maximum follow-up to be extended to 117 months were similar compared to results in setting A1. The average calendar time of stopping with the use of monitoring boundaries, under combinations of extreme treatment effects, and lower than anticipated event rate, is typically shorter relative to a FSD. In summary, the GSD with the "escape clause" is clearly preferred over FSD when $\theta \leq \theta_A$, and the event rate λ is much lower than anticipated. Relative to the FSD, the GSD is almost matched in terms of average power and total sample size. In addition, a smaller average event size, shorter average calendar time of completion is required to obtain similar average statistical power. We note that this result is similar when the "escape clause" is applied at either 78 or 117 months without additional accrual of patients.

The simulation results, based on a planned, original calendar time of 78 months, provide a comparison of the operating characteristics if we choose to extend the study to 117 months via blinded or unblinded adaptations. In addition, when the extension of the calendar time is allowed such that this extension does not affect our ability to address the primary scientific hypothesis, the average calendar time of stopping is not considerably longer than 78 months

when the baseline event rate $\lambda > \lambda_0/4$. In fact, even with extreme treatment effects, the monitoring rule offers protection by adaptively stopping the trial early, and concluding with high probability in favor of the treatment over the placebo.

D.5 Results for Setting B2: Increase in Accrual Size and Extension of Maximum Calendar Time

When we are allowed either a blinded/unblinded increase in our sample size, and at most 50% extension of the calendar time, we can consider the option of making an adaptation early by extension of accrual when the very last patient enters or restart accrual later in the study. When allowing for a blinded increase in accrual size, a smaller average accrual size is observed when we increased patient accrual later. This translates to a lower operational cost when patient costs dominate the budget of the trial.

When the hazard ratio is more extreme than our design alternative (HR $\leq \theta_A$), and the event rate is not far off than planned, the average patient size is smaller if we restart accrual since the group sequential monitoring boundaries have enabled early stopping before additional accrual is necessary in the presence of extreme efficacy. When event rate is lower than planned ($\lambda \in (1/8, 1/2]\lambda_0$), the benefit of a smaller average patient size still persists since the monitoring rule also enables early stopping prior to completion of additional accrual with extreme efficacy. The consequence of restarting accrual, and allowing an extension of the calendar time means that a longer follow-up is obtained on average.

When our hazard ratio is between θ_A and the null, the average patient size is larger, the average calendar time is shorter when accrual was continued relative to the strategy of restarting accrual. The strategy to adapt accrual later, however, provides a more reliable estimate of the event rate.

Under our treatment effect is moderately effective such that $\theta \in (0.5, \theta_A)$, there is little difference in overall power between continuing accrual or restarting accrual when event rates are markedly lower than planned. Restarting accrual decreases the average sample size, at the cost of a slightly longer average calendar time relative to continuing accrual. When

the baseline event rates are close to planned, i.e., $\lambda \geq \lambda_0/2$, there is minimal advantage in choosing to either continue accrual or restart accrual in terms of overall power. On average, a bigger sample size is observed, with a difference between the average calendar time anywhere between 2 to 10 months of more follow-up when restarting accrual. The ambiguity of either strategy occurs when the event rate is close to $3/4\lambda_0$. The strategy to continue accrual surprisingly beats restarting accrual later with a smaller average sample size, a shorter average calendar time at the cost of 4% loss of power.

Under the extreme event rate of $\lambda_0/8$, either strategy of continuing or restarting accrual appear similar in terms of average sample size with the expected calendar time terminating close to the maximum extended time. The loss of power under this setting is minimal relative to the strategy of designing the trial with twice the original sample size.

Table D.10: Results for the various operating characteristics using the best sampling rule obtained based on $\theta = 0.5$ and $\lambda_0/4$ is prespecified and applied across other values of θ under continuous accrual. Generally, the overall Type 1 error is protected, the power under the hypothesized alternatives are higher relative to fully flexible adaptive design.

			Blinded *GSDMod*†		Blinded (C1) *GSDMod*‡		Adaptive (D0) Rate Diff		Adaptive (D1) HR		Sample Size			Calendar Time			Events			Percentage of adaptation		
	HR	Rate	Pres	Cond	Pres	Cond	Pres	Cond	Pres	Cond	C1	D0	D1	C1	D0	D1	C1	D0	D1	C1	D0	D1
Continuous Accrual	$\theta = 0.5$	$\lambda_0/8$	62.97	62.01	62.90	61.92	62.48	61.15	56.92	54.43	3489	3428	2929	78	78	78	47	46	40	99.4	95.9	67.3
		$\lambda_0/4$	90.08	89.72	86.33	85.74	88.09	86.52	87.55	86.31	3150	3150	3150	75	75	75	80	81	81	80.0	80.0	80.0
		$\lambda_0/2$	99.41	99.41	94.16	94.04	97.37	97.07	95.17	95.04	1902	2481	2123	68	62	66	94	100	96	8.7	41.8	21.3
		$3\lambda_0/4$	99.47	99.46	98.91	98.91	99.04	99.04	98.92	98.92	1753	1921	1770	54	53	54	103	103	103	0.2	9.8	1.1
		λ_0	99.80	99.80	99.75	99.75	99.75	99.75	99.75	99.75	1750	1760	1750	44	44	44	105	105	105	0.0	0.6	0.0
	$\theta = \theta_A$	$\lambda_0/8$	35.89	35.33	35.75	35.17	35.62	34.65	33.06	31.50	3483	3441	3079	78	78	78	51	50	46	99.0	96.6	75.9
		$\lambda_0/4$	61.95	61.83	55.92	55.71	59.94	58.83	59.03	58.25	3014	3232	3199	77	77	77	87	93	92	72.2	84.7	82.8
		$\lambda_0/2$	87.61	87.60	68.59	68.31	79.33	78.59	70.73	70.49	1837	2550	2005	74	69	73	111	126	114	5.0	45.7	14.6
		$3\lambda_0/4$	86.60	86.62	82.89	82.95	83.73	83.76	82.93	82.99	1751	1877	1758	64	63	64	136	137	136	0.1	7.3	0.4
		λ_0	89.05	89.05	89.52	89.52	89.46	89.46	89.52	89.52	1750	1756	1750	53	53	53	143	143	143	0.0	0.3	0.0
	$\theta = 0.75$	$\lambda_0/8$	18.04	18.09	17.92	17.96	17.93	17.70	17.35	16.79	3475	3449	3174	78	78	78	54	54	50	98.6	97.1	81.3
		$\lambda_0/4$	32.05	31.92	27.73	27.38	30.73	29.92	30.05	29.50	2883	3288	3182	77	77	77	91	101	99	64.8	87.9	81.8
		$\lambda_0/2$	51.75	51.73	36.42	36.14	44.14	43.55	37.50	37.16	1803	2539	1920	76	71	75	120	139	123	3.0	45.1	9.7
		$3\lambda_0/4$	50.90	50.85	48.67	48.62	48.97	48.90	48.66	48.61	1750	1834	1753	68	67	68	154	155	154	0.0	4.8	0.2
		λ_0	53.16	53.16	53.55	53.55	53.51	53.51	53.55	53.55	1750	1752	1750	56	56	56	163	163	163	0.0	0.1	0.0
	$\theta = 1$	$\lambda_0/8$	2.60	2.49	2.61	2.49	2.66	2.50	2.69	2.51	3450	3464	3294	78	78	78	61	62	59	97.2	97.9	88.2
		$\lambda_0/4$	2.42	2.38	2.35	2.31	2.37	2.23	2.39	2.31	2618	3352	3041	74	71	73	90	105	99	49.6	91.5	73.8
		$\lambda_0/2$	2.63	2.62	2.43	2.47	2.50	2.49	2.47	2.51	1765	2317	1815	66	60	65	115	118	115	0.8	32.6	3.7
		$3\lambda_0/4$	2.56	2.56	2.57	2.58	2.59	2.60	2.57	2.58	1750	1775	1750	50	50	50	125	125	125	0.0	1.5	0.0
		λ_0	2.57	2.57	2.56	2.56	2.56	2.56	2.56	2.56	1750	1750	1750	41	41	41	127	127	127	0.0	0.0	0.0

†: Corresponds to the group sequential design with "escape clause" and blinded adaptations conducted at 100% of the time under the setting when $\theta = 0.5$ and $\lambda_0/4$.

‡: Corresponds to the group sequential design with "escape clause" and blinded adaptations at 80% of the time.

Table D.11: Results for the various operating characteristics using the best sampling rule obtained based on $\theta = 0.5$ and $\lambda_0/4$ is prespecified and applied across other values of θ when accrual is restarted. Generally, the overall Type 1 error is protected, the power under the hypothesized alternatives are higher relative to fully flexible adaptive design.

			Blinded *GSDMod*†		Blinded (C1) *GSDMod*‡		Adaptive (D0) Rate Diff		Adaptive (D1) HR		Sample Size			Calendar Time			Events			Percentage of adaptation		
	HR	Rate	Pres	Cond	Pres	Cond	Pres	Cond	Pres	Cond	C1	D0	D1	C1	D0	D1	C1	D0	D1	C1	D0	D1
Restart Accrual	$\theta = 0.5$	$\lambda_0/8$	51.34	44.65	51.34	44.65	51.33	44.59	50.36	40.57	3500	3486	2992	78	78	78	35	35	33	100.0	99.2	71.0
		$\lambda_0/4$	80.27	76.88	78.27	73.91	80.27	75.25	80.10	75.07	3149	3150	3150	78	78	78	66	66	66	80.0	80.0	80.0
		$\lambda_0/2$	97.84	97.78	93.50	92.50	94.81	93.89	94.76	94.22	1750	1909	2035	70	70	70	94	95	96	0.0	9.2	17.0
		$3\lambda_0/4$	99.73	99.73	98.92	98.84	98.92	98.84	98.92	98.84	1750	1750	1750	55	55	55	105	105	105	0.0	0.0	0.0
		λ_0	99.81	99.81	99.75	99.74	99.75	99.74	99.75	99.74	1750	1750	1750	44	44	44	106	106	106	0.0	0.0	0.0
	$\theta = \theta_A$	$\lambda_0/8$	28.19	24.80	28.19	24.80	28.21	24.74	28.04	23.56	3500	3486	3219	78	78	78	38	38	37	100.0	99.2	84.0
		$\lambda_0/4$	49.68	47.37	46.01	42.94	49.66	46.32	49.73	46.85	2847	3310	3383	78	78	78	69	74	75	62.7	89.1	93.3
		$\lambda_0/2$	79.47	79.51	67.56	66.01	68.65	67.15	68.58	67.18	1750	1854	1856	75	75	75	109	111	111	0.0	5.9	6.2
		$3\lambda_0/4$	89.76	89.75	82.90	82.82	82.90	82.82	82.90	82.82	1750	1750	1750	65	65	65	138	138	138	0.0	0.0	0.0
		λ_0	89.17	89.17	89.65	89.64	89.65	89.64	89.65	89.64	1750	1750	1750	54	54	54	146	146	146	0.0	0.0	0.0
	$\theta = 0.75$	$\lambda_0/8$	14.92	13.21	14.92	13.21	14.92	13.21	15.03	12.89	3499	3477	3315	78	78	78	41	41	40	100.0	98.7	89.4
		$\lambda_0/4$	24.98	23.64	22.31	20.63	25.11	23.18	25.07	23.43	2569	3367	3453	78	78	78	71	80	81	46.8	92.4	97.3
		$\lambda_0/2$	44.01	43.93	35.74	34.79	35.85	34.99	35.86	34.99	1750	1790	1786	76	76	76	119	120	119	0.0	2.3	2.1
		$3\lambda_0/4$	52.90	52.88	48.68	48.52	48.68	48.52	48.68	48.52	1750	1750	1750	69	69	69	156	156	156	0.0	0.0	0.0
		λ_0	53.17	53.16	53.76	53.76	53.76	53.76	53.76	53.76	1750	1750	1750	57	57	57	167	167	167	0.0	0.0	0.0
	$\theta = 1$	$\lambda_0/8$	2.75	2.46	2.75	2.46	2.77	2.46	2.89	2.59	3497	3387	3352	78	78	78	47	46	46	99.8	93.5	91.6
		$\lambda_0/4$	2.25	2.29	2.36	2.28	2.49	2.26	2.28	2.23	2080	3208	3422	77	77	77	74	87	89	18.9	83.4	96.2
		$\lambda_0/2$	2.49	2.50	2.40	2.47	2.40	2.47	2.40	2.47	1750	1751	1751	67	67	67	116	116	116	0.0	0.0	0.0
		$3\lambda_0/4$	2.58	2.57	2.62	2.63	2.62	2.63	2.62	2.63	1750	1750	1750	51	51	51	127	127	127	0.0	0.0	0.0
		λ_0	2.69	2.64	2.70	2.65	2.70	2.65	2.70	2.65	1750	1750	1750	41	41	41	127	127	127	0.0	0.0	0.0

†: Corresponds to the group sequential design with "escape clause" and blinded adaptations conducted at 100% of the time under the setting when $\theta = 0.5$ and $\lambda_0/4$.

‡: Corresponds to the group sequential design with "escape clause" and blinded adaptations at 80% of the time.

Table D.12: Results for the various operating characteristics using the best sampling rule obtained based on θ_A and $\lambda_0/2$ is prespecified and applied across other values of θ under continuous accrual. Generally, the overall Type 1 error is protected, the power under the hypothesized alternatives are higher relative to fully flexible adaptive design.

			Blinded *GSDMod*†		Blinded (C1) *GSDMod*‡		Adaptive (D0) Rate Diff		Adaptive (D1) HR		Sample Size			Calendar Time			Events			Percentage of adaptation		
	HR	Rate	Pres	Cond	Pres	Cond	Pres	Cond	Pres	Cond	C1	D0	D1	C1	D0	D1	C1	D0	D1	C1	D0	D1
Continuous Accrual	$\theta = 0.5$	$\lambda_0/8$	62.97	62.01	62.97	62.01	62.93	61.92	52.93	50.08	3500	3494	2583	78	78	78	47	47	36	100.0	99.6	47.6
		$\lambda_0/4$	90.08	89.72	90.08	89.71	89.84	89.10	84.01	81.81	3499	3410	2784	74	75	76	87	86	74	100.0	94.9	59.1
		$\lambda_0/2$	99.41	99.41	98.95	98.95	98.93	98.68	98.61	98.33	3289	2994	2890	54	57	58	104	103	102	88.1	71.1	65.2
		$3\lambda_0/4$	99.47	99.46	99.20	99.20	99.56	99.52	99.67	99.64	2255	2524	2778	50	47	45	104	104	105	29.5	44.7	59.8
		λ_0	99.80	99.80	99.76	99.76	99.83	99.83	99.83	99.83	1795	2013	2173	43	42	41	105	105	105	2.9	15.9	26.3
	$\theta = \theta_A$	$\lambda_0/8$	35.89	35.33	35.89	35.33	35.86	35.28	30.39	28.72	3500	3495	2735	78	78	78	51	51	42	100.0	99.7	56.3
		$\lambda_0/4$	61.95	61.83	61.94	61.82	61.63	61.18	57.09	55.45	3499	3433	2995	77	77	77	99	97	87	100.0	96.2	71.2
		$\lambda_0/2$	87.61	87.60	84.63	84.59	86.21	85.69	86.10	85.58	3149	3150	3150	63	64	64	136	137	137	80.0	80.0	80.0
		$3\lambda_0/4$	86.60	86.62	84.56	84.58	87.48	87.44	88.21	88.20	2057	2647	2872	61	55	52	138	141	142	17.6	51.4	64.5
		λ_0	89.05	89.05	89.47	89.47	89.66	89.66	89.65	89.65	1767	1979	2066	53	51	50	143	144	144	1.0	13.4	18.6
	$\theta = 0.75$	$\lambda_0/8$	18.04	18.09	18.04	18.09	18.02	18.04	16.22	15.35	3500	3495	2845	78	78	78	55	55	46	100.0	99.7	62.5
		$\lambda_0/4$	32.05	31.92	32.02	31.89	31.94	31.49	29.67	28.68	3498	3451	3135	77	77	77	107	105	98	99.9	97.2	79.1
		$\lambda_0/2$	51.75	51.73	48.34	48.22	50.96	50.53	51.25	50.86	2990	3249	3287	67	65	65	150	156	157	70.9	85.7	87.9
		$3\lambda_0/4$	50.90	50.85	49.41	49.35	51.67	51.54	52.01	51.91	1935	2636	2785	66	58	56	156	160	160	10.6	50.9	59.5
		λ_0	53.16	53.16	53.52	53.52	53.81	53.81	53.81	53.81	1757	1914	1949	56	54	54	163	163	163	0.4	9.5	11.6
	$\theta = 1$	$\lambda_0/8$	2.60	2.49	2.60	2.49	2.59	2.48	2.75	2.55	3500	3496	3017	78	78	78	62	62	55	100.0	99.8	72.4
		$\lambda_0/4$	2.42	2.38	2.42	2.39	2.40	2.31	2.50	2.35	3493	3472	3316	71	71	71	109	108	104	99.6	98.4	89.5
		$\lambda_0/2$	2.63	2.62	2.51	2.49	2.73	2.66	2.73	2.69	2603	3299	3341	57	50	50	120	123	124	49.0	89.4	91.8
		$3\lambda_0/4$	2.56	2.56	2.62	2.63	2.55	2.55	2.61	2.61	1800	2306	2347	50	46	46	125	125	125	3.0	34.0	36.4
		λ_0	2.57	2.57	2.56	2.56	2.58	2.58	2.58	2.58	1750	1791	1794	41	40	40	127	127	127	0.0	2.8	2.9

†: Corresponds to the group sequential design with "escape clause" and blinded adaptations conducted at 100% of the time under the setting when θ_A and $\lambda_0/2$.

‡: Corresponds to the group sequential design with "escape clause" and blinded adaptations at 80% of the time.

Table D.13: Results for the various operating characteristics using the best sampling rule obtained based on θ_A and $\lambda_0/2$ is prespecified and applied across other values of θ when accrual is restarted. Generally, the overall Type 1 error is protected, the power under the hypothesized alternatives are higher relative to fully flexible adaptive design.

			Blinded *GSDMod*†		Blinded (C1) *GSDMod*‡		Adaptive (D0) Rate Diff		Adaptive (D1) HR		Sample Size			Calendar Time			Events			Percentage of adaptation		
	HR	Rate	Pres	Cond	Pres	Cond	Pres	Cond	Pres	Cond	C1	D0	D1	C1	D0	D1	C1	D0	D1	C1	D0	D1
Restart Accrual	$\theta = 0.5$	$\lambda/8$	51.34	44.65	51.34	44.65	51.34	44.59	47.33	36.81	3500	3490	2460	78	78	78	35	35	30	100.0	99.4	40.6
		$\lambda/4$	80.27	76.88	80.27	76.88	80.27	76.78	78.29	70.18	3498	3459	2562	77	78	78	69	69	61	100.0	97.7	46.4
		$\lambda/2$	97.84	97.78	97.70	97.64	97.77	97.43	97.72	97.00	3108	2728	2543	66	68	68	103	102	101	86.1	58.7	47.1
		$3\lambda/4$	99.73	99.73	98.95	98.88	99.44	99.36	99.67	99.67	1784	1943	2260	55	55	54	105	106	107	2.4	13.0	37.4
		λ	99.81	99.81	99.75	99.74	99.75	99.74	99.75	99.74	1750	1750	1763	44	44	44	106	106	106	0.0	0.0	1.1
	$\theta = \theta_A$	$\lambda/8$	28.19	24.80	28.19	24.80	28.18	24.74	26.29	21.07	3500	3466	2680	78	78	78	38	38	34	100.0	98.1	53.1
		$\lambda/4$	49.68	47.37	49.68	47.37	49.63	47.31	48.71	43.37	3499	3456	2956	78	78	78	76	75	70	100.0	97.5	68.9
		$\lambda/2$	79.47	79.51	77.55	77.36	79.31	78.84	79.35	78.77	3094	3121	3119	73	73	73	127	128	128	80.0	80.0	80.0
		$3\lambda/4$	89.76	89.75	82.93	82.85	84.68	84.55	87.33	87.30	1755	1965	2404	65	64	63	138	140	143	0.3	13.4	43.4
		λ	89.17	89.17	89.65	89.64	89.65	89.64	89.66	89.65	1750	1750	1751	54	54	54	146	146	146	0.0	0.0	0.1
	$\theta = 0.75$	$\lambda/8$	14.92	13.21	14.92	13.21	14.92	13.18	14.49	11.80	3500	3425	2776	78	78	78	41	41	37	100.0	95.7	58.6
		$\lambda/4$	24.98	23.64	24.98	23.64	24.98	23.61	24.82	22.25	3499	3406	3113	78	78	78	81	80	77	100.0	94.6	77.9
		$\lambda/2$	44.01	43.93	41.05	40.53	43.99	43.44	44.06	43.70	2812	3259	3314	75	74	74	136	144	145	62.1	87.5	91.0
		$3\lambda/4$	52.90	52.88	48.69	48.53	49.37	49.24	50.55	50.46	1750	1846	2137	69	68	67	156	157	160	0.0	5.9	24.8
		λ	53.17	53.16	53.76	53.76	53.76	53.76	53.76	53.76	1750	1750	1750	57	57	57	167	167	167	0.0	0.0	0.0
	$\theta = 1$	$\lambda/8$	2.75	2.46	2.75	2.46	2.75	2.52	2.98	2.62	3500	3235	2753	78	78	78	47	45	42	100.0	84.8	57.3
		$\lambda/4$	2.25	2.29	2.25	2.29	2.26	2.27	2.46	2.21	3481	3091	2957	76	77	77	90	86	85	99.7	76.7	69.0
		$\lambda/2$	2.49	2.50	2.40	2.42	2.48	2.40	2.49	2.48	2033	2699	2896	66	64	63	119	124	125	18.2	60.8	74.8
		$3\lambda/4$	2.58	2.57	2.62	2.63	2.62	2.63	2.61	2.62	1750	1752	1768	51	51	51	127	127	127	0.0	0.2	1.4
		λ	2.69	2.64	2.70	2.65	2.70	2.65	2.70	2.65	1750	1750	1750	41	41	41	127	127	127	0.0	0.0	0.0

†: Corresponds to the group sequential design with "escape clause" and blinded adaptations conducted at 100% of the time under the setting when θ_A and $\lambda_0/2$.

‡: Corresponds to the group sequential design with "escape clause" and blinded adaptations at 80% of the time.

Appendix E

Additional Results for Chapter 6

E.1 Additional Results for Blinded Adaptations

Table E.1: Overall Type 1 error rate when we increase the accrual size in a blinded fashion at interim analyses conducted at either 1/3, 1/2, or 2/3 of the final event size.

		Immediate				Early				Uniform				Delayed			
	N_{Final}	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
Events=255	1000	2.493	2.476	2.504	2.487	2.496	2.476	2.498	2.484	2.495	2.477	2.502	2.482	2.495	2.475	2.502	2.484
	1500	2.483	2.474	2.472	2.469	2.483	2.475	2.480	2.484	2.481	2.473	2.476	2.475	2.479	2.472	2.477	2.472
	2000	2.482	2.473	2.502	2.511	2.483	2.475	2.500	2.494	2.482	2.476	2.492	2.490	2.481	2.477	2.502	2.497
	3000	2.478	2.479	2.484	2.477	2.476	2.473	2.489	2.475	2.479	2.487	2.478	2.483	2.473	2.476	2.485	2.486
	5000	2.502	2.506	2.473	2.466	2.472	2.472	2.462	2.461	2.465	2.468	2.467	2.457	2.478	2.471	2.471	2.467
Events=382	1000	2.493	2.476	2.504	2.487	2.496	2.476	2.498	2.484	2.495	2.477	2.502	2.482	2.495	2.475	2.502	2.484
	1500	2.456	2.467	2.474	2.475	2.469	2.472	2.472	2.480	2.466	2.472	2.468	2.475	2.462	2.468	2.473	2.473
	2000	2.471	2.485	2.479	2.478	2.485	2.476	2.492	2.480	2.482	2.476	2.484	2.473	2.476	2.473	2.489	2.479
	3000	2.461	2.456	2.467	2.462	2.473	2.467	2.468	2.470	2.472	2.463	2.466	2.458	2.468	2.463	2.464	2.458
	5000	2.488	2.493	2.482	2.481	2.490	2.476	2.471	2.470	2.495	2.482	2.480	2.474	2.486	2.486	2.468	2.472
Events=510	1000	2.493	2.476	2.504	2.487	2.496	2.476	2.498	2.484	2.495	2.477	2.502	2.482	2.495	2.475	2.502	2.484
	1500	2.485	2.482	2.483	2.486	2.481	2.474	2.485	2.490	2.474	2.475	2.479	2.492	2.478	2.474	2.483	2.492
	2000	2.480	2.488	2.491	2.480	2.470	2.494	2.493	2.494	2.474	2.493	2.496	2.499	2.476	2.491	2.495	2.488
	3000	2.487	2.466	2.473	2.458	2.471	2.484	2.487	2.487	2.463	2.481	2.482	2.483	2.469	2.484	2.480	2.487
	5000	2.475	2.472	2.478	2.481	2.472	2.479	2.483	2.484	2.464	2.482	2.487	2.487	2.472	2.484	2.479	2.491

Error rate in blue denotes that the 1,000,000 simulations results are not within the 95% CI of what would have been a typical Type 1 error rate of 2.5% (2.4694 2.5306).

Table E.2: Overall Type 1 error rate of $\alpha = 5\%$ when we increase the accrual size in a blinded fashion at interim analyses conducted at either 1/3, 1/2, or 2/3 of the final event size.

		Immediate				Early				Uniform				Delayed			
	N_{Final}	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
Events=255	1000	4.980	4.976	5.018	4.997	4.980	4.976	5.023	4.999	4.982	4.976	5.026	4.998	4.981	4.977	5.020	4.998
	1500	4.981	4.988	4.993	4.993	4.988	4.985	4.987	4.980	4.984	4.983	4.999	4.976	4.980	4.989	4.998	4.980
	2000	4.967	4.961	4.994	4.984	4.971	4.959	5.005	4.986	4.974	4.956	4.997	4.982	4.972	4.958	4.989	4.984
	3000	4.972	4.965	4.967	4.963	4.960	4.969	4.979	4.949	4.963	4.966	4.966	4.954	4.967	4.965	4.967	4.958
	5000	4.986	4.993	4.949	4.959	4.951	4.952	4.934	4.929	4.961	4.963	4.935	4.938	4.956	4.953	4.949	4.951
Events=382	1000	4.980	4.976	5.018	4.997	4.980	4.976	5.023	4.999	4.982	4.976	5.026	4.998	4.981	4.977	5.020	4.998
	1500	4.969	4.970	4.980	4.964	4.965	4.969	4.992	4.968	4.961	4.967	4.993	4.972	4.955	4.969	4.984	4.969
	2000	4.948	4.957	4.961	4.956	4.946	4.954	4.961	4.964	4.955	4.958	4.957	4.961	4.958	4.957	4.958	4.961
	3000	4.960	4.979	4.955	4.957	4.967	4.981	4.963	4.946	4.966	4.976	4.956	4.947	4.970	4.976	4.950	4.943
	5000	4.991	5.005	4.962	4.980	4.967	4.961	4.958	4.946	4.974	4.968	4.965	4.954	4.980	4.964	4.976	4.950
Events=510	1000	4.980	4.976	5.018	4.997	4.980	4.976	5.023	4.999	4.982	4.976	5.026	4.998	4.981	4.977	5.020	4.998
	1500	4.969	4.980	4.966	4.973	4.971	4.979	4.969	4.972	4.975	4.975	4.964	4.980	4.970	4.971	4.961	4.968
	2000	4.998	4.989	4.966	4.963	4.990	4.970	4.968	4.976	4.986	4.980	4.969	4.973	4.991	4.983	4.979	4.980
	3000	4.979	4.982	4.966	4.958	4.986	4.974	4.982	4.969	4.988	4.969	4.979	4.977	4.982	4.968	4.978	4.973
	5000	4.969	4.979	4.977	4.973	4.986	4.970	4.987	4.965	4.979	4.963	4.983	4.979	4.982	4.974	4.985	4.973

Error rate in blue denotes that the 1,000,000 simulations results are not within the 95% CI of what would have been a typical Type 1 error rate of 5% (4.9573, 5.0428).

Table E.3: Overall power when we increase the accrual size in a blinded fashion at interim analysis conducted at 1/3 of the final event size using a one sided level α. Power is computed by assuming the alternative distribution from Weibull with mean 120.5 and shape parameter of 0.5. In addition, the alternative is calibrated based on the $G^{1,0}$ statistic assuming a accrual size of 1000 subjects accrued assuming the above accrual patterns with final event size of 765. Blinded accrual size increase is performed for all simulations after 1/3 the total number of events is observed.

		Immediate				Early				Uniform				Delayed			
	N_{Final}	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
2.5%	1000	93.15	90.02	90.06	84.85	93.15	90.02	90.06	84.90	93.16	90.01	90.06	84.87	93.16	90.02	90.06	84.89
	1500	92.99	92.06	88.13	84.62	92.99	92.06	88.15	84.56	92.98	92.08	88.19	84.62	92.99	92.07	88.17	84.56
	2000	93.01	92.61	87.05	84.31	92.99	92.59	87.04	84.23	92.98	92.58	87.06	84.27	93.00	92.60	87.02	84.24
	3000	93.04	92.85	85.76	83.43	93.12	92.89	85.55	82.89	93.11	92.87	85.52	82.92	93.10	92.86	85.49	82.91
	5000	93.12	92.98	81.92	78.69	93.00	92.80	83.52	80.33	93.02	92.80	83.48	80.31	93.04	92.83	83.48	80.32
5%	1000	92.86	90.00	89.99	85.45	92.86	89.98	90.02	85.43	92.87	89.98	90.03	85.42	92.88	89.98	90.03	85.42
	1500	92.69	91.90	88.31	85.17	92.70	91.89	88.30	85.22	92.70	91.92	88.31	85.22	92.72	91.90	88.30	85.16
	2000	92.79	92.41	87.37	84.93	92.72	92.32	87.30	84.77	92.78	92.34	87.30	84.82	92.75	92.33	87.30	84.83
	3000	92.79	92.58	86.30	84.24	92.87	92.62	85.99	83.68	92.86	92.62	86.00	83.76	92.86	92.64	86.05	83.81
	5000	92.84	92.75	82.96	80.10	92.81	92.59	84.19	81.41	92.79	92.58	84.08	81.43	92.79	92.58	84.16	81.44

Table E.4: Overall power when we increase the accrual size in a blinded fashion at interim analysis conducted at 1/2 of the final event size using a one sided level α. Power is computed by assuming the alternative distribution from Weibull with mean 120.5 and shape parameter of 0.5. In addition, the alternative is calibrated based on the $G^{1,0}$ statistic assuming a accrual size of 1000 subjects accrued assuming the above accrual patterns with final event size of 765. Blinded accrual size increase is performed for all simulations after 1/2 the total number of events is observed.

		Immediate				Early				Uniform				Delayed			
	N_{Final}	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
2.5%	1000	93.15	90.02	90.06	84.85	93.15	90.02	90.06	84.90	93.16	90.01	90.06	84.87	93.16	90.02	90.06	84.89
	1500	93.04	92.08	88.15	84.42	93.04	92.06	88.17	84.38	93.05	92.06	88.19	84.40	93.05	92.05	88.22	84.40
	2000	93.08	92.50	86.82	83.38	93.10	92.50	86.86	83.27	93.06	92.51	86.80	83.23	93.08	92.52	86.86	83.28
	3000	93.07	92.71	84.14	79.63	93.05	92.62	83.90	79.16	93.06	92.66	83.89	79.17	93.04	92.63	83.90	79.17
	5000	93.02	92.67	78.04	72.20	93.02	92.66	82.72	77.74	93.03	92.63	82.76	77.69	93.03	92.61	82.70	77.69
5%	1000	92.86	90.00	89.99	85.45	92.86	89.98	90.02	85.43	92.87	89.98	90.03	85.42	92.88	89.98	90.03	85.42
	1500	92.78	91.86	88.29	85.06	92.76	91.84	88.30	85.07	92.76	91.83	88.29	85.09	92.74	91.83	88.32	85.09
	2000	92.79	92.26	87.19	84.13	92.86	92.33	87.12	84.02	92.84	92.30	87.19	84.03	92.85	92.28	87.21	84.06
	3000	92.80	92.45	84.78	80.77	92.75	92.38	84.57	80.37	92.77	92.43	84.58	80.39	92.76	92.38	84.61	80.41
	5000	92.74	92.42	79.44	74.37	92.76	92.43	83.53	79.11	92.75	92.42	83.52	79.12	92.78	92.46	83.57	79.10

Table E.5: Overall power when we increase accrual size in a blinded fashion at interim analysis conducted at 2/3 of the final event size using a one sided level α. Power is computed by assuming the alternative distribution from Weibull with mean 120.5 and shape parameter of 0.5. In addition, the alternative is calibrated based on the $G^{1,0}$ statistic assuming a accrual size of 1000 subjects accrued assuming the above accrual patterns with final event size of 765. Blinded accrual size increase is performed for all simulations after 2/3 the total number of events is observed.

		Immediate				Early				Uniform				Delayed			
	N_{Final}	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
2.5%	1000	93.15	90.02	90.06	84.85	93.15	90.02	90.06	84.90	93.16	90.01	90.06	84.87	93.16	90.02	90.06	84.89
	1500	93.08	91.93	88.28	83.81	93.12	91.92	88.29	83.85	93.09	91.91	88.28	83.85	93.10	91.93	88.26	83.91
	2000	93.09	92.17	86.49	81.27	93.10	92.22	86.60	81.22	93.13	92.20	86.59	81.24	93.12	92.20	86.57	81.29
	3000	93.09	92.21	83.09	76.92	93.07	92.13	83.82	77.78	93.08	92.13	83.80	77.70	93.08	92.15	83.86	77.79
	5000	93.00	92.19	79.34	73.11	93.07	92.14	83.82	77.75	93.08	92.13	83.79	77.68	93.08	92.15	83.84	77.79
5%	1000	92.86	90.00	89.99	85.45	92.86	89.98	90.02	85.43	92.87	89.98	90.03	85.42	92.88	89.98	90.03	85.42
	1500	92.85	91.75	88.45	84.55	92.82	91.77	88.48	84.52	92.84	91.75	88.49	84.51	92.83	91.75	88.41	84.54
	2000	92.77	91.97	86.86	82.21	92.83	91.98	86.77	82.15	92.80	91.97	86.79	82.13	92.77	91.96	86.80	82.13
	3000	92.78	92.06	83.86	78.37	92.83	91.98	84.50	79.24	92.82	91.97	84.53	79.23	92.82	91.97	84.49	79.20
	5000	92.76	91.96	80.50	75.09	92.83	91.99	84.49	79.23	92.81	91.97	84.51	79.23	92.82	91.96	84.48	79.20

E.2 Additional Results for Unblinded Adaptations

Table E.6: Maximum overall Type 1 error rate with unblinded adaptation conducted at an interim analysis 1/3 of the total event size. 95% CI for 5% error based on 100,000 simulations should be within 4.9573 and 5.0428.

	Immediate				Early				Uniform				Delayed			
Rule	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
1a	4.999	5.847	4.897	4.968	5.012	5.857	4.901	4.970	5.003	5.855	4.907	4.969	5.002	5.850	4.901	4.970
1b	5.012	6.059	4.878	4.957	5.003	6.056	4.883	4.958	5.000	6.054	4.886	4.956	5.001	6.053	4.880	4.957
1c	5.003	6.181	4.875	4.955	5.003	6.112	4.879	4.955	4.996	6.118	4.883	4.953	4.994	6.104	4.877	4.955
2a	5.016	6.055	4.841	4.939	5.000	6.053	4.849	4.940	5.001	6.051	4.849	4.940	5.005	6.052	4.843	4.940
2b	5.007	6.177	4.838	4.937	5.000	6.110	4.846	4.937	5.000	6.113	4.846	4.937	4.997	6.101	4.840	4.937
3	5.009	6.177	4.802	4.913	5.000	6.112	4.811	4.915	5.005	6.112	4.813	4.917	5.000	6.104	4.808	4.917
4a	4.992	5.600	4.926	4.981	5.002	5.606	4.927	4.982	4.992	5.606	4.932	4.981	4.993	5.601	4.927	4.982
4b	5.001	5.843	4.859	4.950	4.999	5.852	4.868	4.952	4.999	5.848	4.870	4.952	4.995	5.845	4.864	4.952
4c	5.009	6.050	4.805	4.916	4.996	6.054	4.815	4.917	4.993	6.044	4.816	4.920	4.995	6.048	4.811	4.920
4d	5.000	6.181	4.792	4.905	4.984	6.105	4.803	4.909	4.990	6.107	4.806	4.910	4.986	6.098	4.800	4.910
5a	4.993	5.600	5.239	5.056	5.002	5.606	5.250	5.060	4.993	5.606	5.258	5.059	4.995	5.601	5.254	5.067
5b	5.000	5.843	5.441	5.173	4.999	5.851	5.438	5.174	4.998	5.847	5.436	5.170	4.995	5.845	5.429	5.177
5c	5.008	6.049	5.711	5.390	4.999	6.053	5.689	5.347	5.000	6.043	5.700	5.370	4.999	6.046	5.675	5.360
5d	5.012	6.178	5.899	5.613	4.988	6.103	5.799	5.480	4.992	6.105	5.807	5.478	4.991	6.096	5.790	5.483

Table E.7: Maximum overall Type 1 error rate with unblinded adaptation conducted at an interim analysis 1/2 of the total event size. 95% CI for 5% error based on 100,000 simulations should be within 4.9573 and 5.0428.

	Immediate				Early				Uniform				Delayed			
Rule	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
1a	5.004	6.164	4.947	4.991	5.007	6.143	4.949	4.994	5.007	6.141	4.952	4.993	5.010	6.151	4.946	4.992
1b	5.003	6.390	4.947	4.991	4.998	6.351	4.949	4.994	4.994	6.341	4.952	4.993	5.004	6.358	4.946	4.991
1c	5.000	6.501	4.947	4.991	4.997	6.369	4.949	4.994	4.992	6.368	4.952	4.993	4.997	6.374	4.946	4.991
2a	5.005	6.384	4.936	4.988	5.004	6.347	4.938	4.991	5.003	6.344	4.941	4.989	5.007	6.353	4.935	4.988
2b	5.008	6.495	4.936	4.988	5.003	6.365	4.938	4.991	5.004	6.370	4.941	4.989	5.008	6.370	4.935	4.988
3	5.013	6.494	4.935	4.988	5.000	6.364	4.937	4.990	4.998	6.363	4.940	4.988	5.011	6.365	4.934	4.988
4a	4.994	5.841	4.951	4.992	4.987	5.829	4.953	4.995	4.988	5.830	4.957	4.994	4.987	5.838	4.950	4.992
4b	5.004	6.157	4.936	4.989	5.003	6.137	4.938	4.991	5.003	6.139	4.941	4.989	5.006	6.144	4.936	4.988
4c	5.000	6.372	4.935	4.988	4.999	6.344	4.937	4.990	4.994	6.335	4.940	4.988	5.009	6.347	4.934	4.988
4d	4.999	6.480	4.935	4.988	4.998	6.362	4.937	4.990	4.995	6.364	4.940	4.988	5.003	6.367	4.934	4.988
5a	4.995	5.841	5.315	5.067	4.987	5.829	5.328	5.062	4.992	5.830	5.321	5.059	4.990	5.838	5.314	5.056
5b	5.005	6.157	5.525	5.174	5.003	6.136	5.503	5.168	5.004	6.139	5.506	5.155	5.007	6.144	5.495	5.159
5c	5.004	6.372	5.589	5.284	5.003	6.344	5.607	5.276	5.000	6.335	5.599	5.262	5.011	6.346	5.594	5.271
5d	4.999	6.480	5.460	5.288	5.006	6.361	5.596	5.293	5.002	6.364	5.592	5.291	5.008	6.367	5.594	5.288

Table E.8: Maximum overall Type 1 error rate with unblinded adaptation conducted at an interim analysis 2/3 of the total event size. 95% CI for 5% error based on 100,000 simulations should be within 4.9573 and 5.0428.

	Immediate				Early				Uniform				Delayed			
Rule	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$	LR	$G^{1,0}$	$G^{0,1}$	$G^{1,1}$
1a	5.003	6.388	5.006	4.997	4.996	6.370	5.012	4.999	5.001	6.370	5.014	4.998	4.999	6.373	5.008	4.997
1b	4.988	6.550	5.006	4.997	4.994	6.516	5.012	4.999	4.996	6.510	5.014	4.998	4.990	6.511	5.008	4.997
1c	4.989	6.616	5.006	4.997	4.997	6.511	5.012	4.999	4.995	6.508	5.014	4.998	4.993	6.512	5.008	4.997
2a	5.003	6.533	5.006	4.997	5.001	6.510	5.012	4.999	5.000	6.507	5.014	4.998	4.997	6.505	5.008	4.997
2b	5.007	6.600	5.006	4.997	5.003	6.504	5.012	4.999	5.000	6.501	5.014	4.998	4.999	6.506	5.008	4.997
3	4.995	6.595	5.006	4.997	4.996	6.500	5.012	4.999	4.997	6.494	5.014	4.998	4.995	6.497	5.008	4.997
4a	4.987	6.099	5.006	4.997	4.987	6.098	5.012	4.999	4.985	6.092	5.014	4.998	4.986	6.097	5.008	4.997
4b	5.003	6.370	5.006	4.997	4.995	6.356	5.012	4.999	4.997	6.359	5.014	4.998	4.997	6.363	5.008	4.997
4c	4.990	6.524	5.006	4.997	4.993	6.498	5.012	4.999	4.995	6.492	5.014	4.998	4.994	6.491	5.008	4.997
4d	4.983	6.584	5.006	4.997	4.994	6.492	5.012	4.999	4.994	6.489	5.014	4.998	4.994	6.494	5.008	4.997
5a	4.989	6.099	5.336	5.069	4.987	6.098	5.333	5.066	4.990	6.092	5.333	5.067	4.988	6.097	5.328	5.063
5b	5.005	6.370	5.351	5.116	4.996	6.356	5.359	5.123	4.998	6.359	5.364	5.119	4.997	6.363	5.366	5.130
5c	4.992	6.524	5.267	5.122	4.993	6.498	5.307	5.124	4.995	6.492	5.315	5.121	4.995	6.491	5.301	5.121
5d	4.992	6.584	5.172	5.085	4.994	6.492	5.310	5.123	4.994	6.489	5.316	5.118	4.994	6.494	5.307	5.116

E.3 Additional Results Based on Naïve Assumption that Statistical Information is Related to the Number of Events

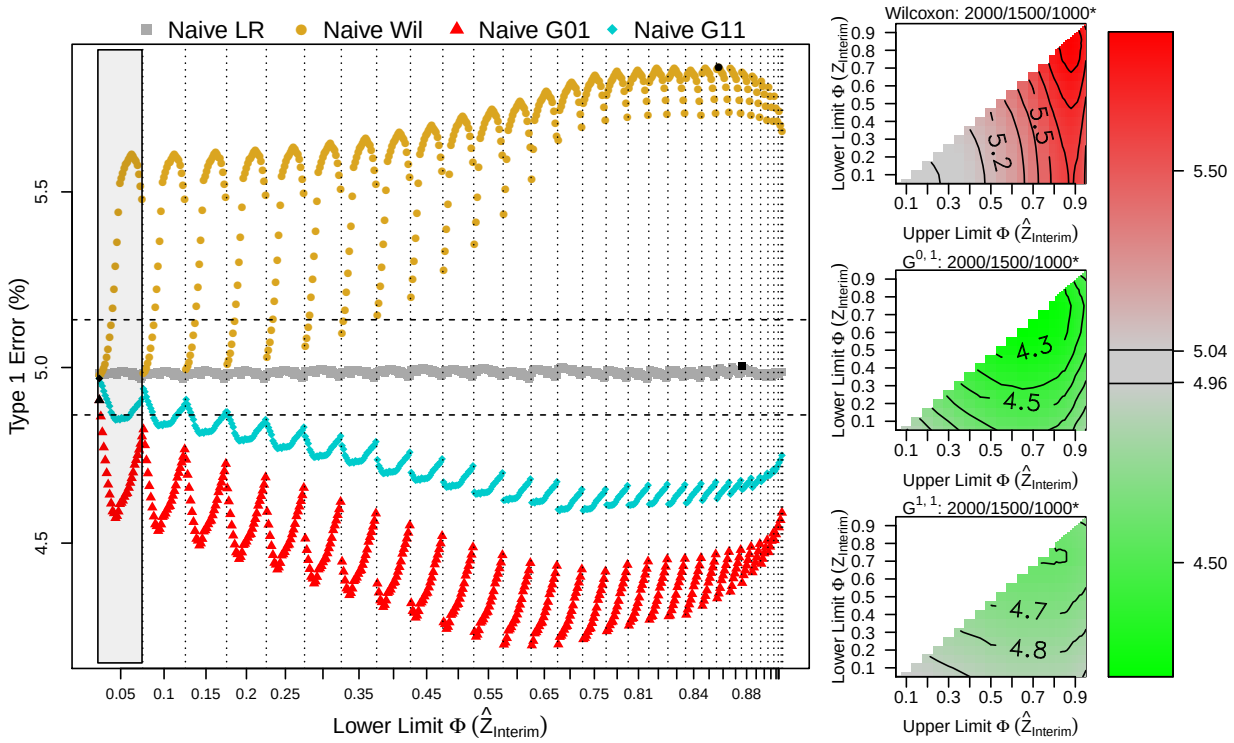

Figure E.1: Degree of inflation of overall Type 1 error using the adaptive rule to increase accrual size to 1500 in the promising zone, and 2000 in the favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

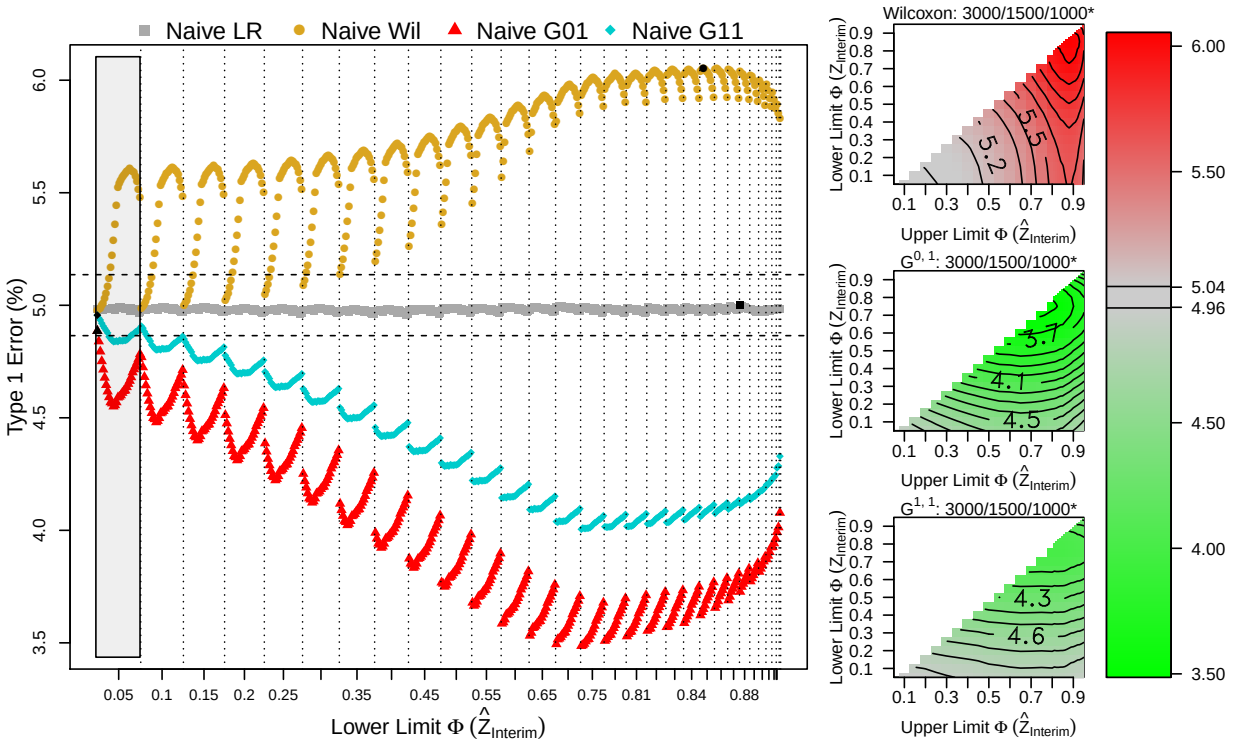

Figure E.2: Degree of inflation of overall Type 1 error using the adaptive rule to increase accrual size to 1500 in the promising zone, and 3000 in the favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

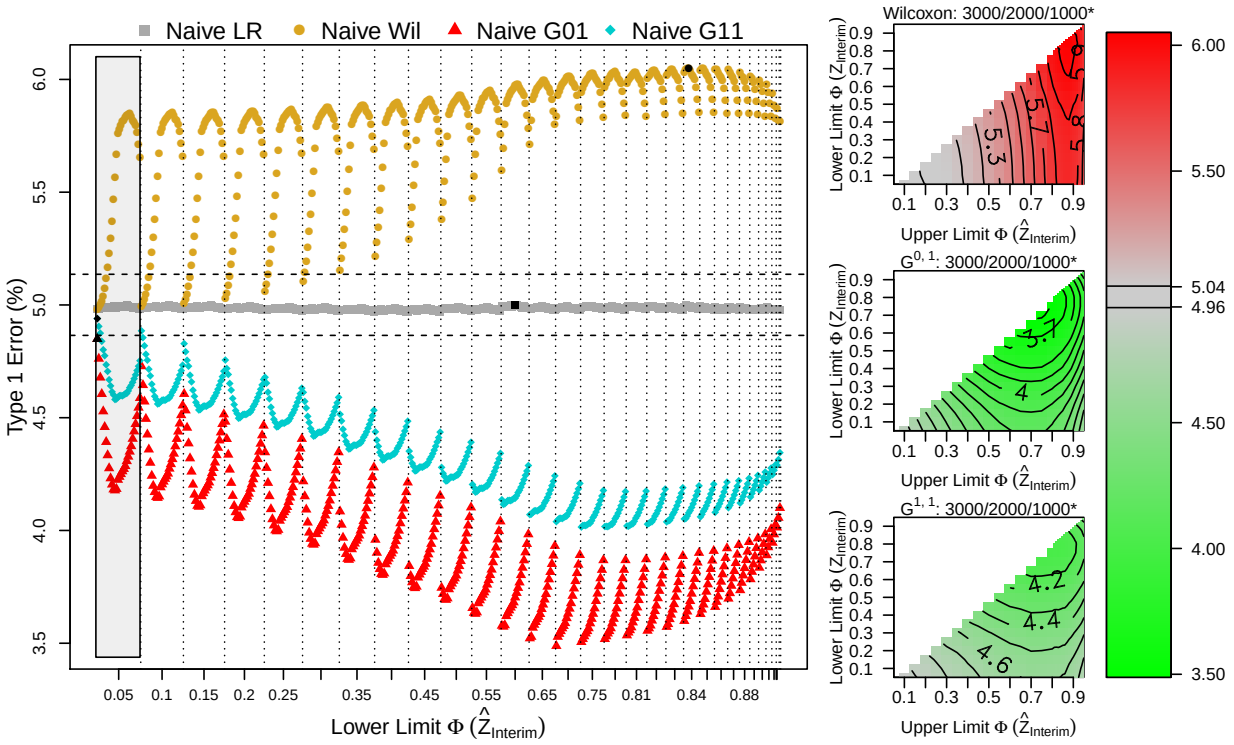

Figure E.3: Degree of inflation of overall Type 1 error using the adaptive rule to increase accrual size to 2000 in the promising zone, and 3000 in the favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

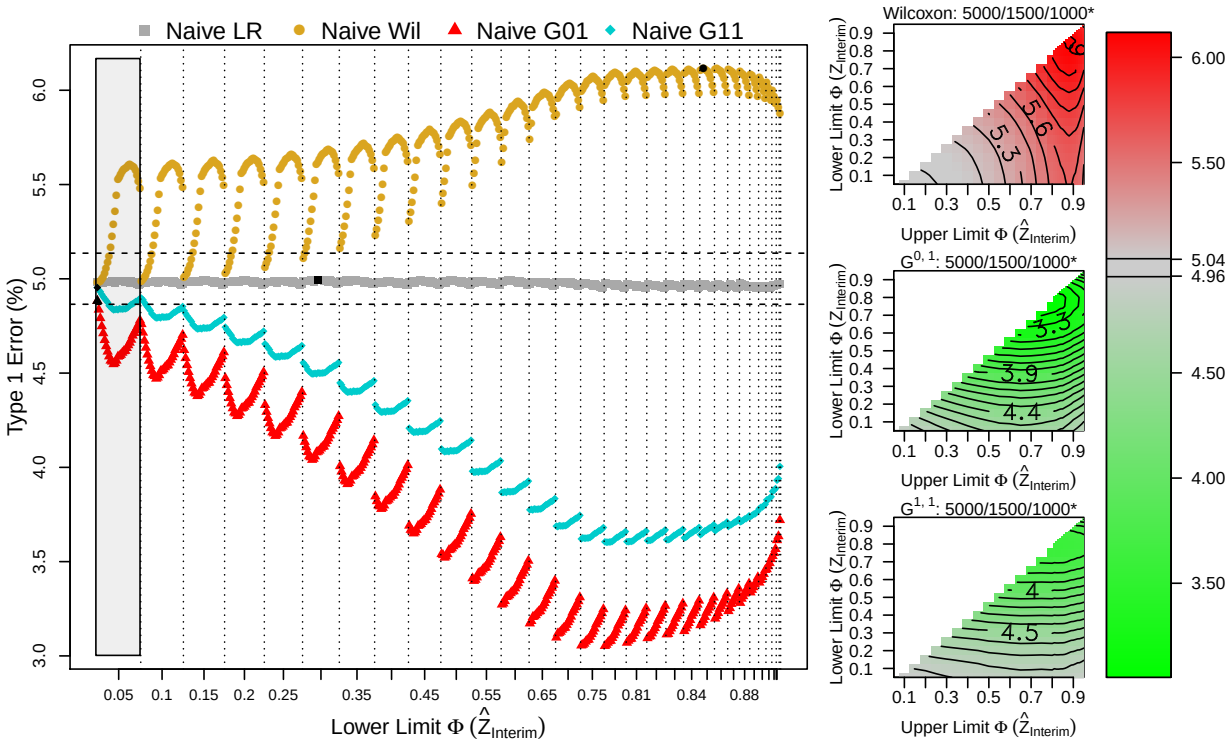

Figure E.4: Degree of inflation of overall Type 1 error using the adaptive rule to increase accrual size to 1500 in the promising zone, and 5000 in the favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

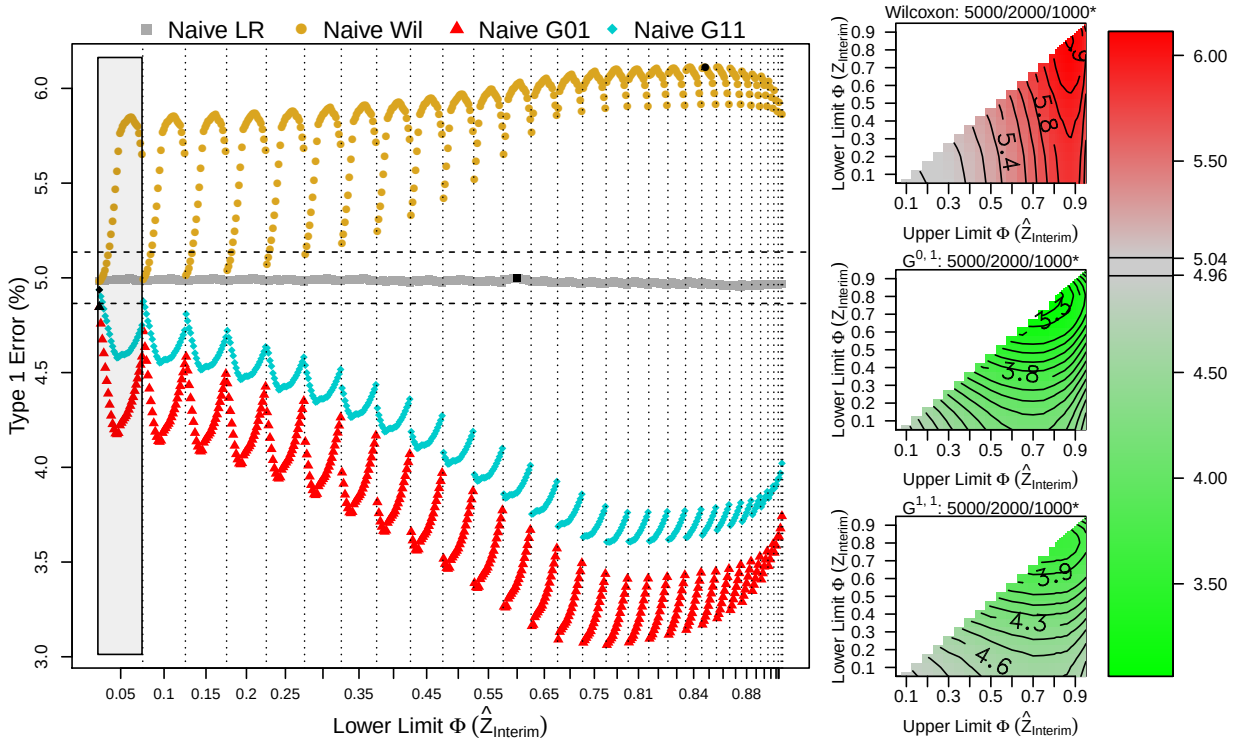

Figure E.5: Degree of inflation of overall Type 1 error using the adaptive rule to increase accrual size to 2000 in the promising zone, and 5000 in the favorable zone under uniform accrual when an adaptation is made at 1/3 of the total event size.

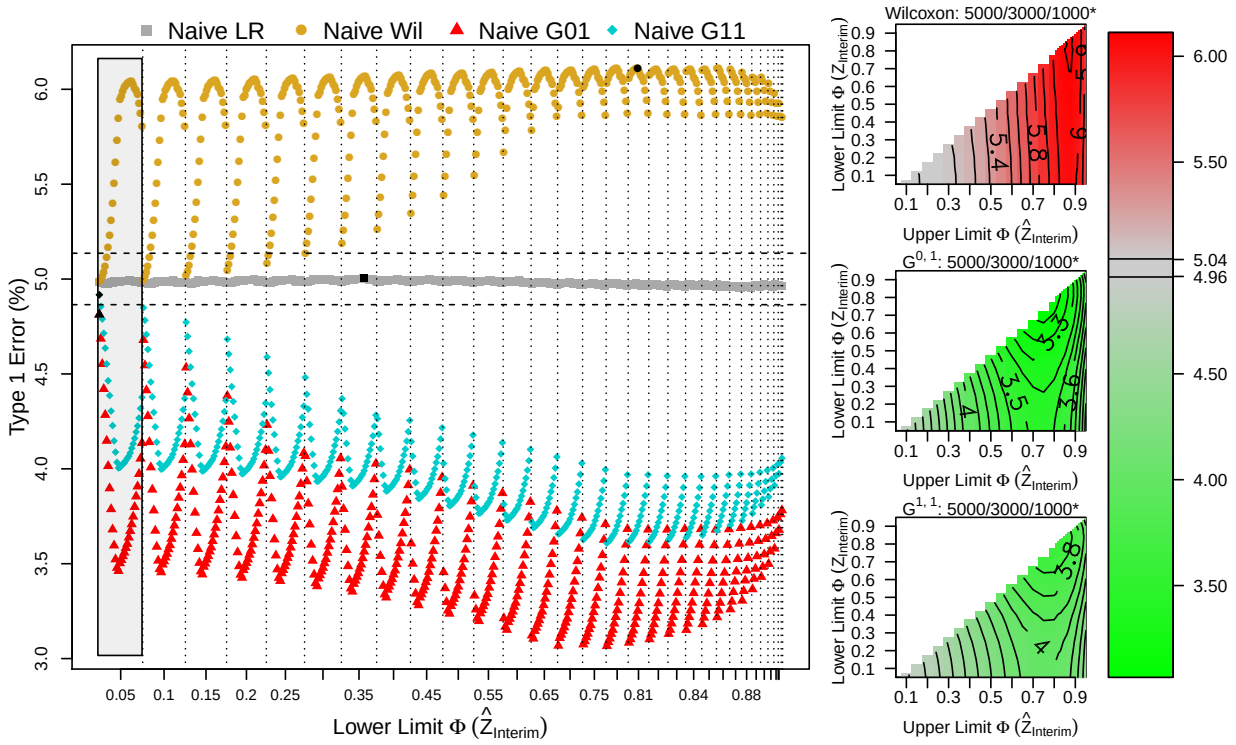

Figure E.6: Degree of inflation of overall Type 1 error using the adaptive rule to increase accrual size to 3000 in the promising zone, and 5000 in the favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

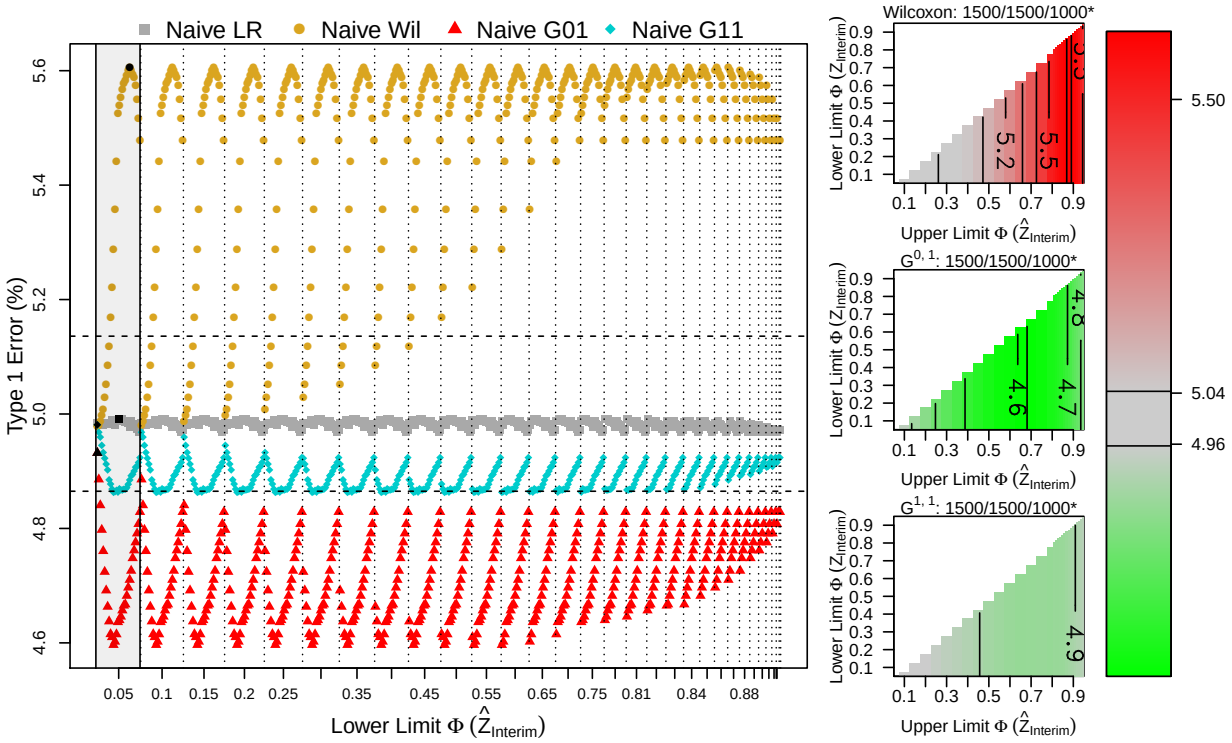

Figure E.7: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 1500 in the promising/favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

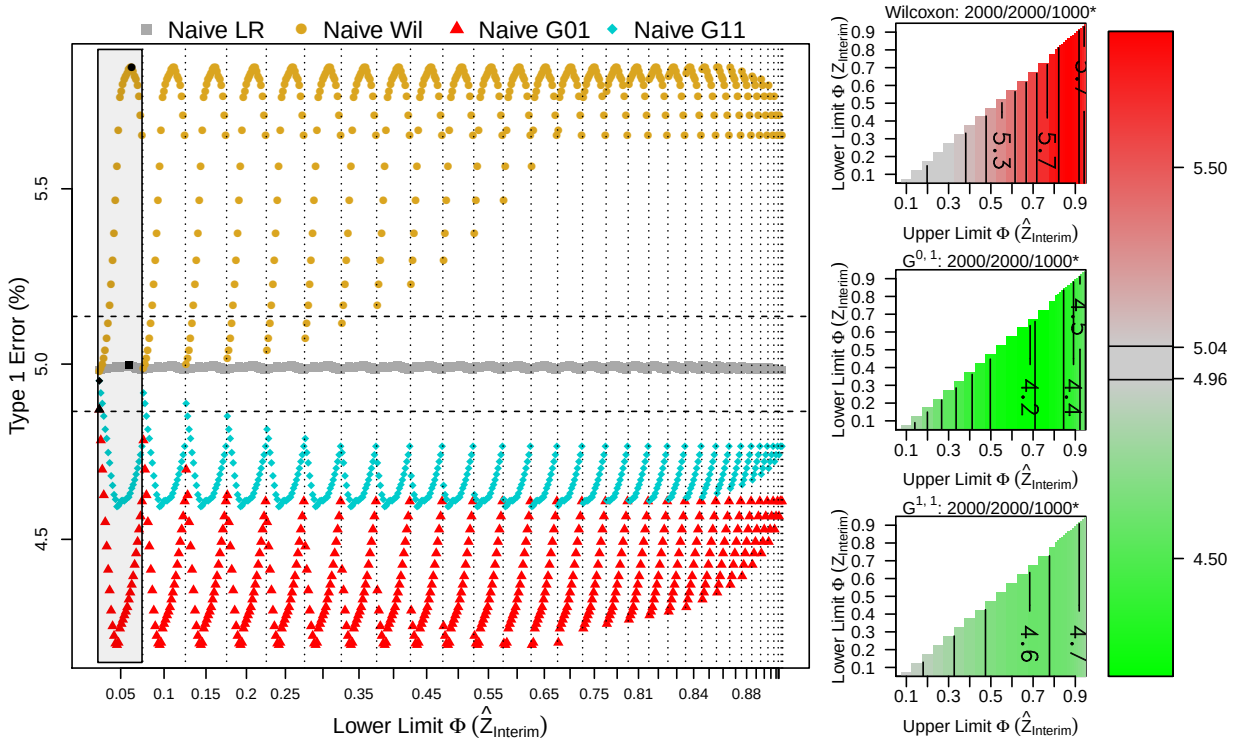

Figure E.8: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 2000 in the promising/favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

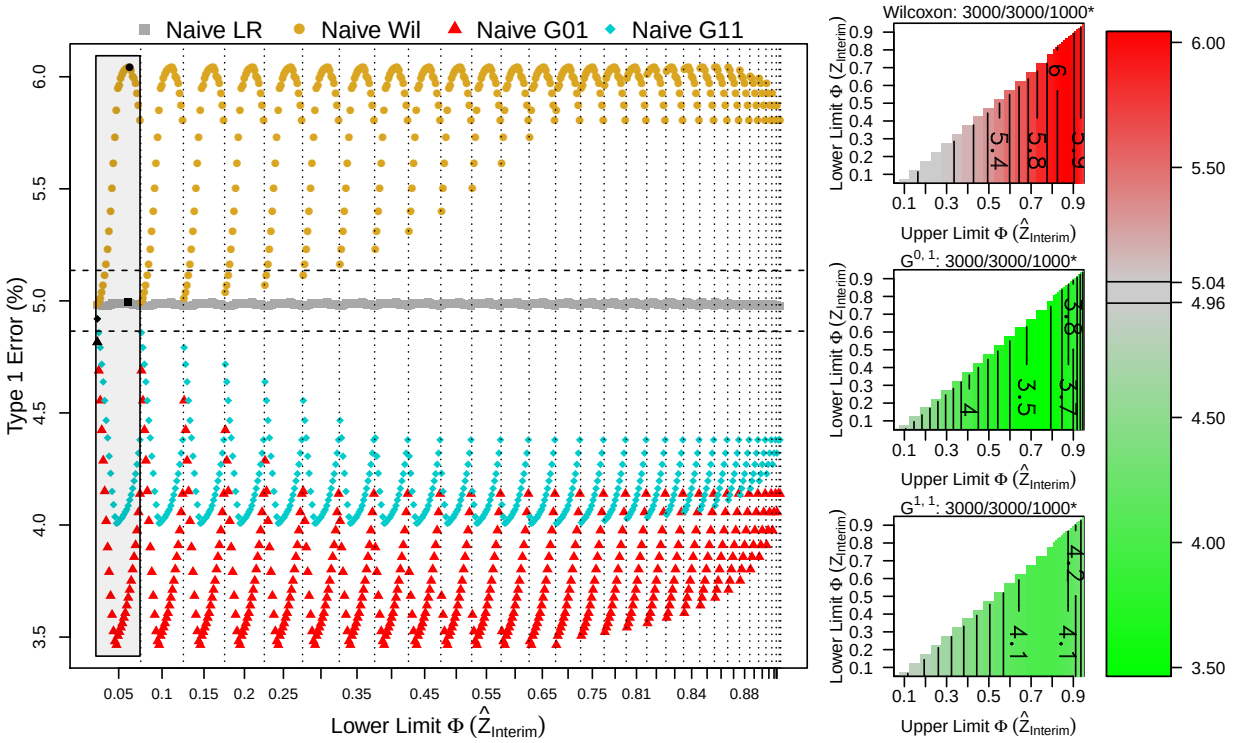

Figure E.9: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 3000 in the promising/favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

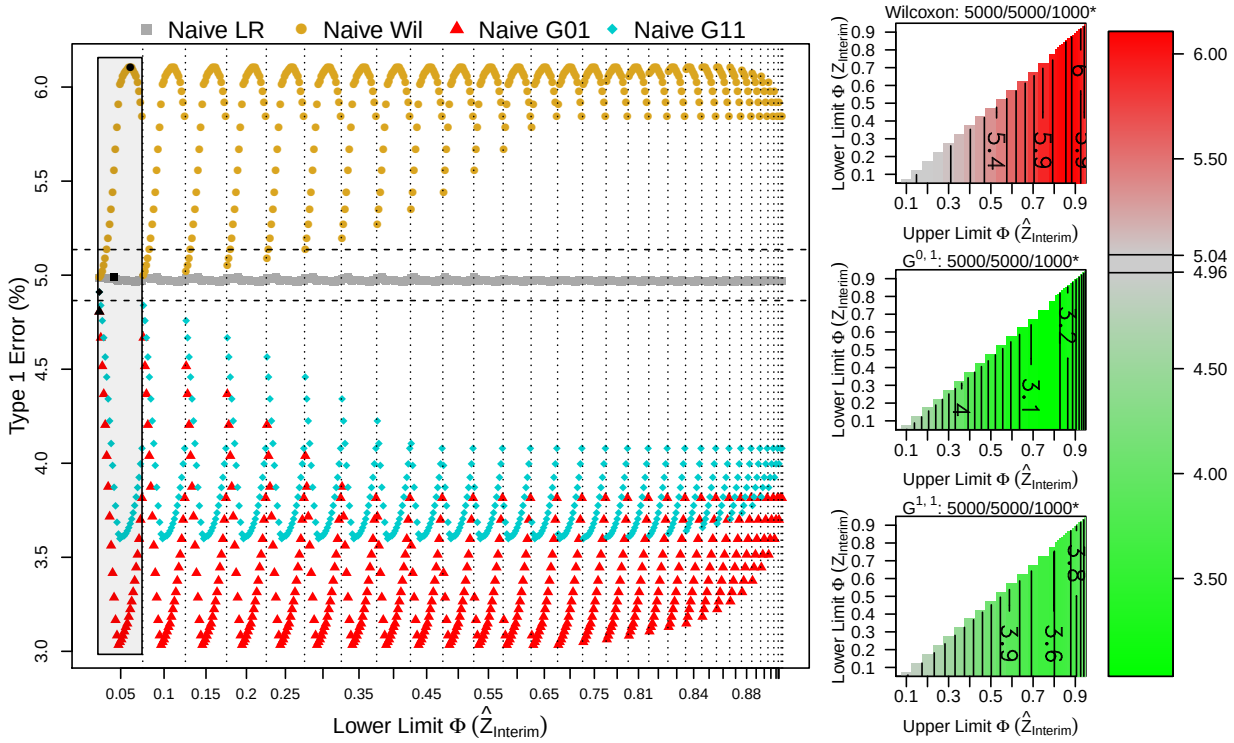

Figure E.10: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 5000 in the promising/favorable zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

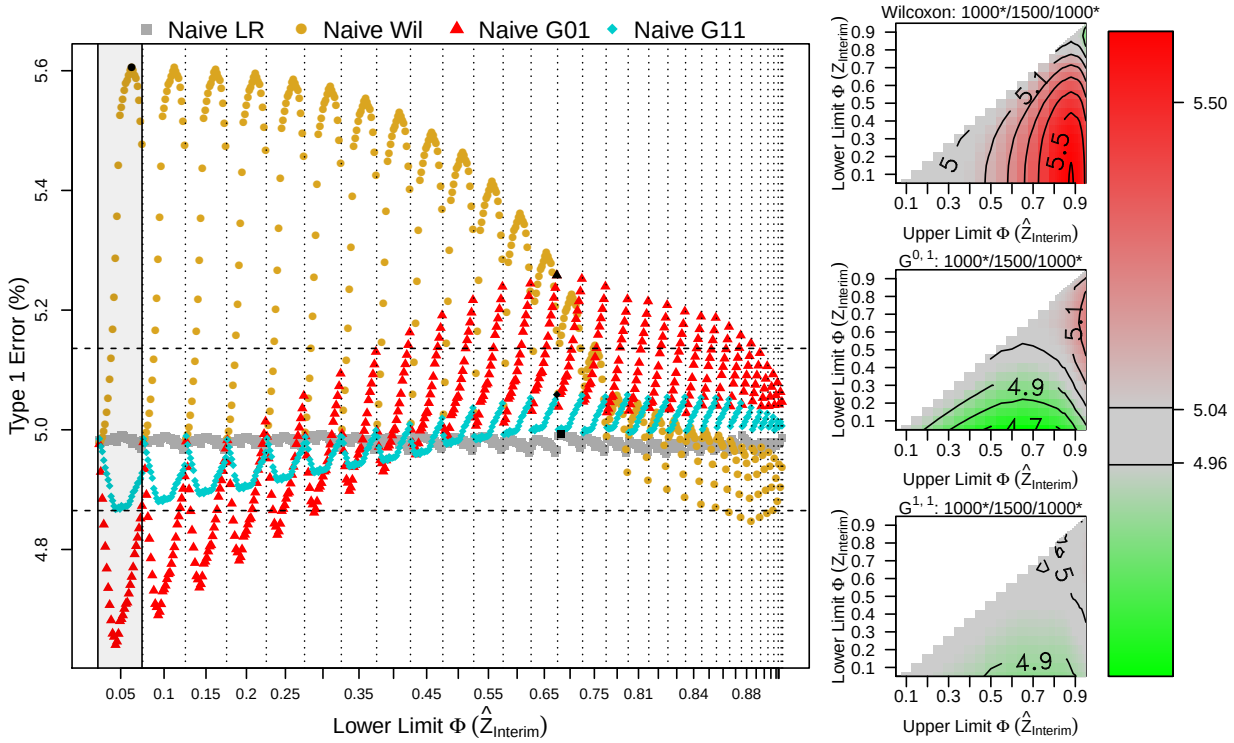

Figure E.11: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 1500 in the promising zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

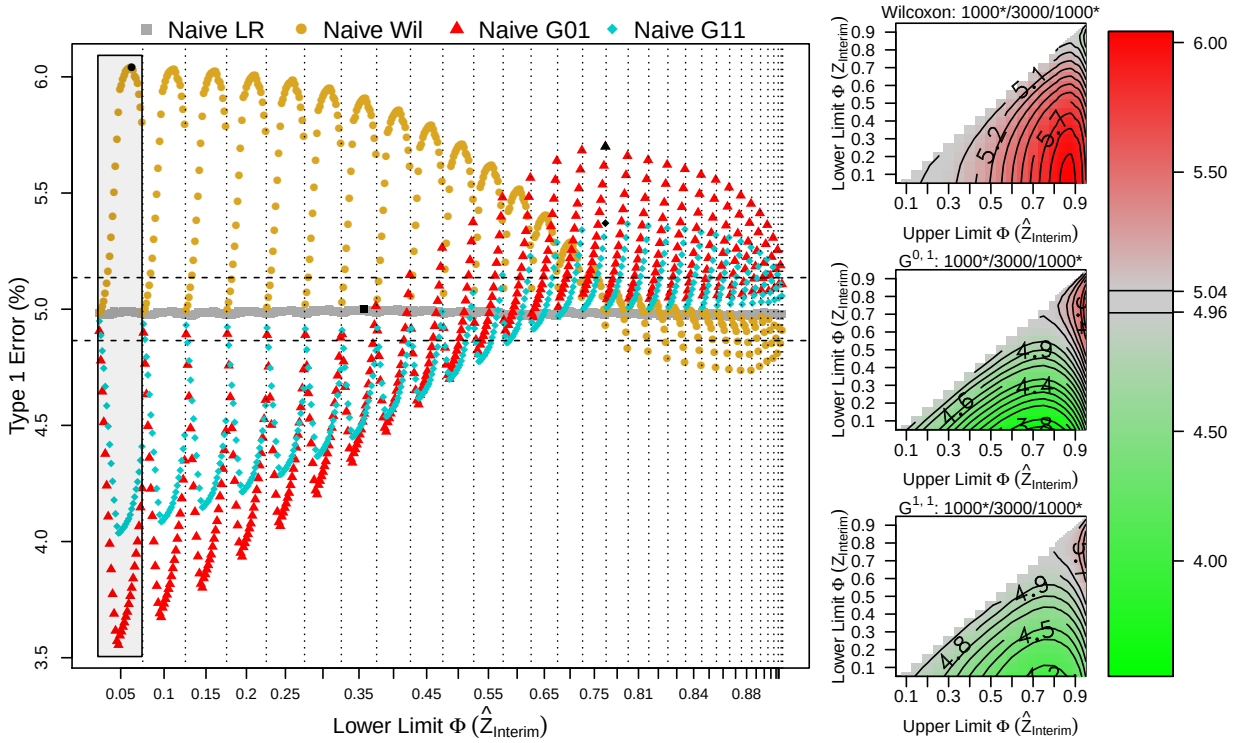

Figure E.12: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 3000 in the promising zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

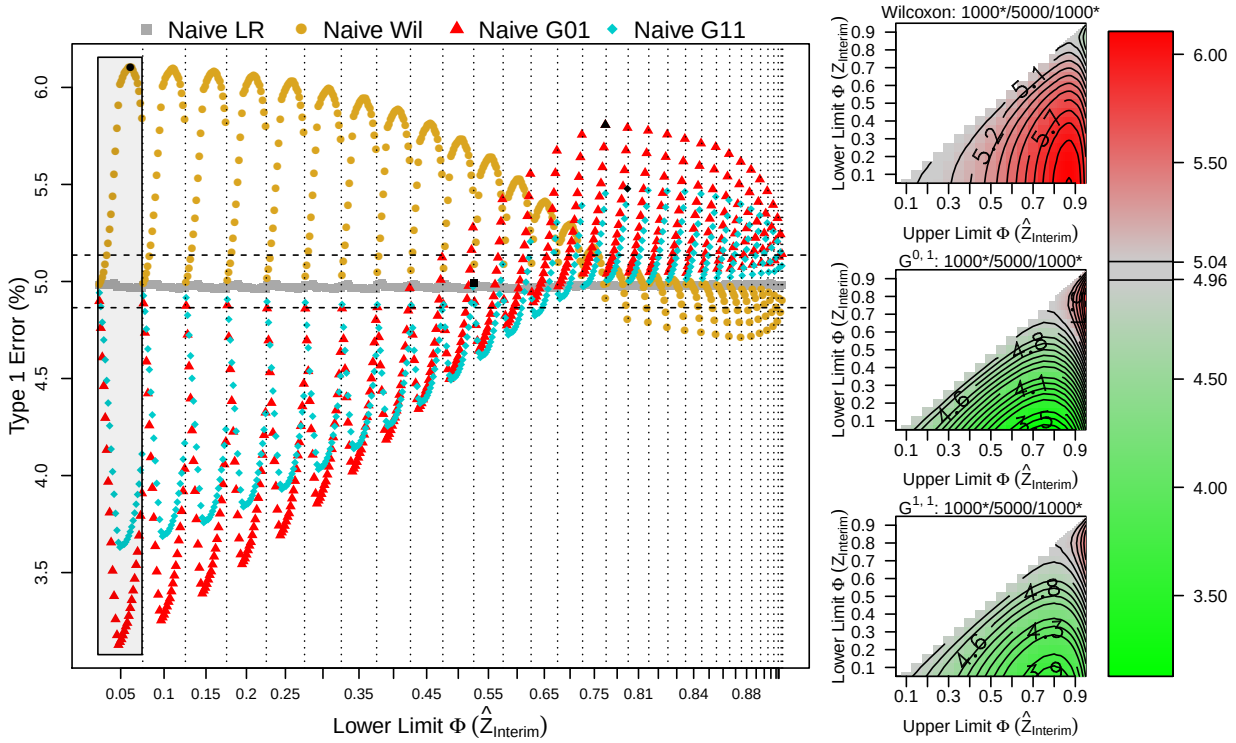

Figure E.13: Degree of inflation of overall Type 1 error using the adaptive rule to only increase accrual size to 5000 in the promising zone under uniform accrual when an adaptation is conducted at an interim analysis 1/3 of the total event size.

Table E.9: Table of prespecified and flexible scenarios under various incorrect specification of maximum statistical information. If no adaptation is done, then our final critical value remains the same as designed since $\mathcal{V}_{\text{Final}} = \mathcal{V}_{\text{Prespecified}}$.

	$\mathcal{V}_{\text{Interim}}$	$\mathcal{V}_{\text{Prespecified}}$	$\mathcal{V}_{\text{Final}}$
AA1	Estimated	$\widehat{\mathcal{V}}_{\text{Maximum}}$	Estimated
AA2	Re-estimated	$\widehat{\mathcal{V}}_{\text{Maximum}}$	Estimated
		Fully prespecified Procedures	
		LR, Wil, $G^{0,1}$, $G^{1,1}$	
B1	Estimated	191.25,110,20,6.5	Estimated
B2	Re-estimated	191.25,110,20,6.5	Estimated
B3	Estimated/Re-estimated**	191.25,110,20,6.5	Estimated
D1	Estimated	191.25,82.5,37.25,7.75*	Estimated
D2	Re-estimated	191.25,82.5,37.25,7.75*	Estimated
D3	Estimated/Re-estimated**	191.25,82.5,37.25,7.75*	Estimated
F1	Estimated	Mean/Estimated	Estimated
F2	Re-estimated	Mean/Estimated	Estimated
F3	Estimated/Re-estimated**	Mean/Estimated	Estimated
		Semi prespecified Procedures	
C1	Estimated	191.25,110,20,**6.5**/Estimated^	Estimated
C2	Re-estimated	191.25,110,20,**6.5**/Estimated^	Estimated
C3	Estimated/Re-estimated**	191.25,110,20,**6.5**/Estimated^	Estimated
E1	Estimated	191.25,82.5,37.25,7.75*/Estimated^	Estimated
E2	Re-estimated	191.25,82.5,37.25,7.75*/Estimated^	Estimated
E3	Estimated/Re-estimated**	191.25,82.5,37.25,7.75*/Estimated^	Estimated
G1	Estimated	Mean/Estimated	Estimated
G2	Re-estimated	Mean/Estimated	Estimated
G3	Estimated/Re-estimated**	Mean/Estimated	Estimated

$\mathcal{V}_{\text{Prespecified}}$: This would correspond to the statistical information for continuing the course of the trial without performing a accrual size adaptation.

$\mathcal{V}_{\text{Final}}$: This would correspond to the statistical information for continuing the course of the trial after performing a accrual size adaptation.

* Maximum statistical information is specified based on the maximum information using simulation. In the logrank setting, this is specified using the theoretical ratio.

^: In this setting, if there is no adaptation performed, we observed the maximum statistical information, and thus use the observed maximum statistical information rather than the prespecified statistical information.

** In this scenario, we only re-estimate the interim information when our estimated information at final is less than information at interim.

E.4 Impact of Additional Accrual on Short Term Survival

Figure E.14 shows the plot of the average statistical information vs the proportion of events (relative to final event size of 765) when we make accrual size increments at either 1/3 or 2/3 of the final event size assuming a short term survival under the null hypothesis. For the logrank statistic, information growth is linear regardless of any accrual size adjustment. However, when making accrual size adjustment with the use of the weighted versions of the logrank statistics, the information growth behaves differently. Under the limiting case of immediate accrual, the proportionate information for the other test statistics of interest has kinks depending on when the interim analysis for this added adaptation was scheduled. The $G^{1,0}$ statistic down-weights earlier information relative to the new events from new accrual size that occurred later as seen when the black solid lines becomes more and more weighted towards the diagonal line. In the $G^{0,1}$ and $G^{1,1}$ scenario, increase in accrual size suggests that the proportionate information is weighted heavily towards the earlier part of the information growth. Under the extreme scenario of immediate accrual, when our interim analysis is conducted at 2/3 of the total number of events, the proportionate information is close to 1 at such interim analysis.

Using the $G^{1,0}$ statistic, the estimated average statistical information at interim analysis is increasing nonlinearly with respect to the proportion of event size after making accrual size adjustment. At either 1/3 or 2/3 of the way through the study, when we increases the accrual size, information growth tends to grow more linearly. For other accrual patterns, information growth appears to be more linear regardless of the degree to which the accrual size was changed during the interim analysis. However, we see that the maximum information at the prespecified final event size has high variability. Unless we can determine the underlying survival distribution, it is difficult to estimate the maximum information during the course of the trial.

For the $G^{0,1}$ statistic, there is more variability in the information growth. The information growth under the immediate pattern is the slowest in this family. Interim adaptations to the

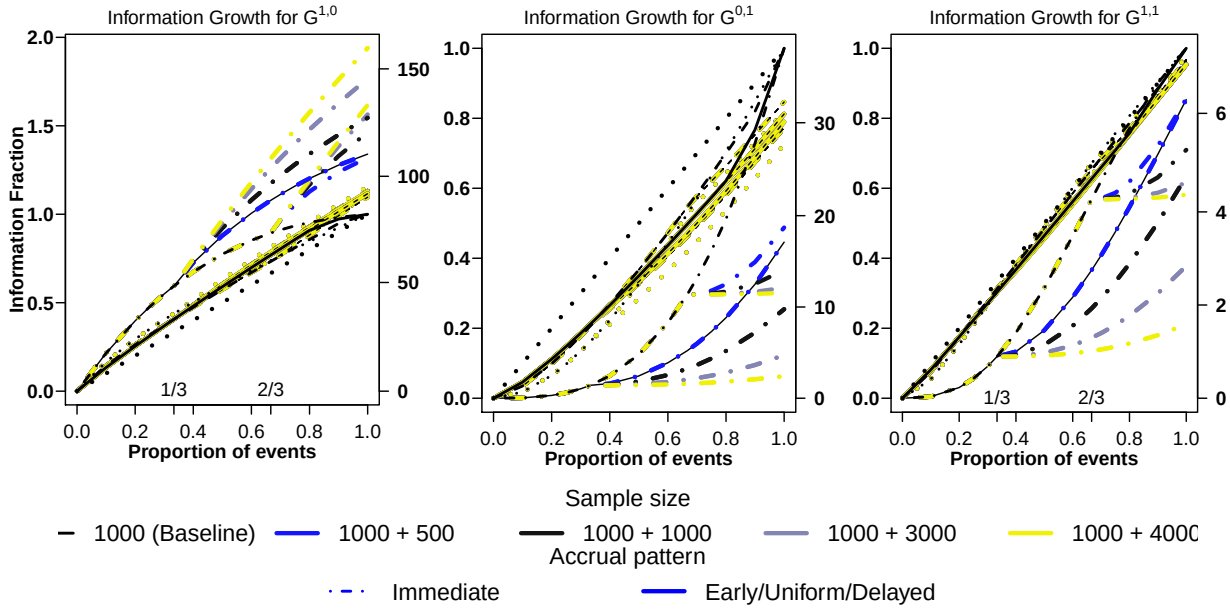

Figure E.14: Information growth for short term survival. Under this setting, the patterns of accrual have differential impact on the variability of the information growth for the various weighted logrank statisics.

accrual size results in the information growth to plateau or flattened as the weight function $1 - S(t^-)$ at this interim analysis increases dramatically. Staggered accrual patterns appear to balance out this drastic modification of our weight function (as a consequence of accrual size adaptation) and lead to a relative linear increase in information growth.

For the $G^{1,1}$ statistic, we see trends in information growth similar to the $G^{0,1}$ statistic. Information growth is again non-linear and discontinuous under the immediate accrual and increasing accrual size subsequently at interim analysis. Under the staggered accrual pattern, our information growth appears to be linear in trend. Using the weight function $(1 - \hat{S}(t-))\hat{S}(t-)$, these weights changed the rate of information growth over time. When we accrue patients immediately, the weight function grows at a rate of $1/n^*$ rather than $1/n$, where $n^* >> n$, thus the information growth increases in a relatively slow manner as more patients are accrued during interim analysis.

Under varying accrual patterns (early, delayed and uniform), there is a clear difference in the estimated maximum information for both long term and short term survival. The rate of information growth is modified further by changes to the accrual distribution when our underlying survival distribution, and total number of events remain unchanged. Under short term survival, different adaptations to the accrual patterns do not appear to have a direct impact on information growth. However, we note that the rate of information growth tends to be close to linear under different accrual patterns. In particular, there is an increased variability in the information growth. This indicates that we may not be able to consistently determine the final statistical information for the weighted logrank statistic as precisely as possible unless we know both the true censoring distribution as well as the underlying survival distribution.

E.5 Implications of Censoring on the Precision of the Variance Estimate at Interim Analyses

Consider the group sequential design with a total of J analyses (recall section 2.2), and our target parameter to be the difference in treatment means, and σ^2 to be the variance of the treatment. At the j^{th} interim analysis, our estimate of the statistical information can be denoted as $I_j(\sigma^2, \mu) = \frac{n_j}{\widehat{\sigma}_j^2} = V_j$, where σ is often estimated based on the data. The information fraction is then computed using $\Pi_j = \frac{n_j}{n_J} \frac{\widehat{\sigma}_J^2}{\widehat{\sigma}_j^2}$ where $\widehat{\sigma}_J^2$ can be the original statistical information based on design assumptions.

Sometimes, monitoring procedures may assume that $\widehat{\sigma}_J^2 = \widehat{\sigma}_j^2$, and the information fraction at interim analyses can be simplified to $\Pi_j = \frac{n_j}{n_J}$ for $j = 1, \cdots, J$. Procedures for estimation of σ^2 may vary across interim analyses as a consequence of the precision of the test statistic as we accumulate more statistical information. In order to recalibrate monitoring boundaries with the use of error spending approaches, or constrained boundary procedures, one can choose to update the variability of the test statistic based on current data, and revise the boundaries accordingly with this revised update in the estimate of statistical information [Burington and Emerson, 2003].

In the time to event setting, the use of the logrank statistic naïvely translate to the ratio of events at the interim analyses relative to the total planned number of events. The statistical information of the weighted logrank statistics no longer has such simple relationship with the number of events as seen earlier since this involves the number at risk, as well as the estimate of the overall survival weights $\widehat{S}(t-)$. An adaptive modification to the accrual of the subjects changes the censoring distribution, further affecting the precision of V_j at interim and the final statistical information V_J. We describe some isses when this imprecision of weights earlier on, as a consequence of a smaller accrual size, can affect our estimation of information growth at interim analyses, and possibly at the final analysis.

We consider the long term survival with decreasing hazard based on the Weibull distribution with shape parameter 0.5 and mean time to be 120.1 to illustrate the above problem under the null hypothesis. Under the extreme scenario of immediate accrual of 1000 subjects at baseline, and increasing accrual again at interim analysis from 1000 to a final accrual size of 5000, we observed an apparent reversal of statistical information for the $G^{0,1}$ and $G^{1,1}$ statistic. Roughly 20% and 5% of the estimated statistical information at the final analysis for $G^{0,1}$ and $G^{1,1}$ are less than the estimated interim statistical information. We explained the apparent reversal of statistical information below.

Following interim analysis, after accumulating 510 events, the remaining 255 events were contributed mostly from patients who entered after the interim analysis. In this extreme scenario, the calendar time to completion following this extreme increase in accrual size is almost instantaneous. However, all 255 events contribute to improving the estimate of the overall survival at the earlier time point. When the interim estimate of these survival curves are less optimistic (blue dotted lines), i.e., the estimated overall survival was much lower than the true survival (red line) at the earlier time-point, the remaining 255 events coming from the 2nd stage patients re-weights the overall survival curve by pulling the (blue dotted) survival curve closer to the true survival curves as indicated now by the black solid line. Our estimate of $\widehat{S}(t_{\text{Int}})$ is less precise compared to the estimate of $\widehat{S}(t_{\text{Final}})$ since our total number at risk is different. This improvement in the precision of the estimates of the total number at

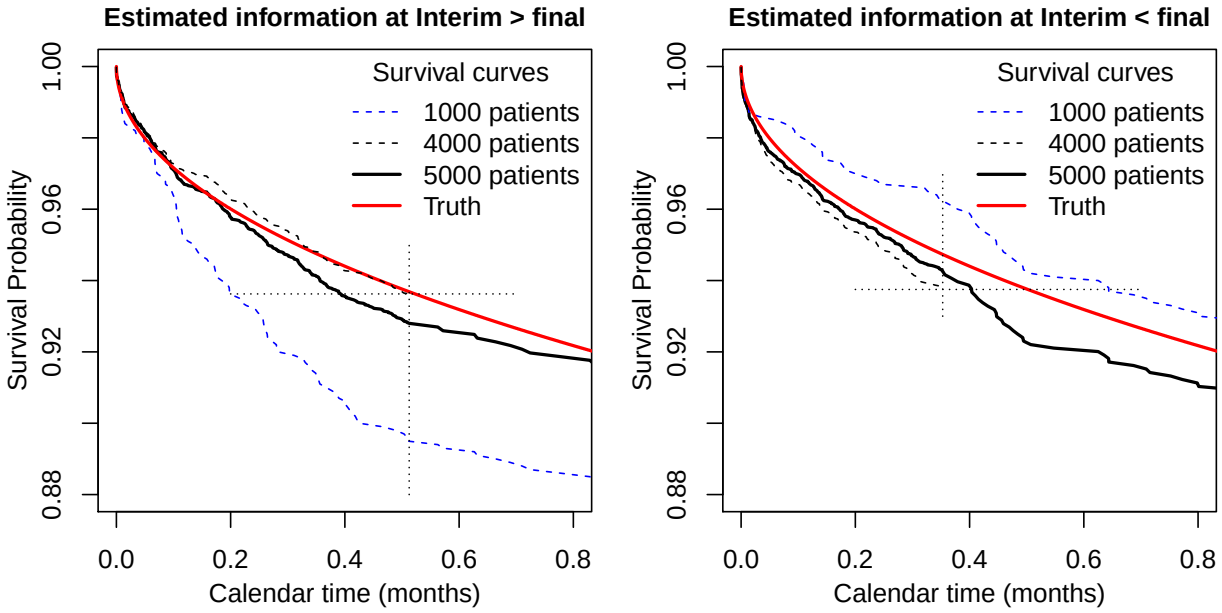

Figure E.15: Estimated survival curves for the different accrual size conducted at 2/3 of the final event size, and the overall estimated survival curves at the end of the trial.

risk provides the apparent contradiction of non-monotonic or reversal of information growth in the $G^{0,1}$ and $G^{1,1}$ setting. On the other hand, when the current estimate of the survival curves prior to interim analysis is overly optimistic, i.e., the true survival experience is worse than anticipated, such reversal of statistical information is not observed.

The above issues can be mitigated by applying the weights for these weighted logrank statistics based on the "revised" total number at risk at the final analysis and restricting these weights to the timing of the interim analysis. Such can overcome the imprecision as a consequence of random high bias of the survival distribution at interim analysis and facilitate a proper update of the interim statistical information. Alternatively, we can also prespecify a weighting scheme based on some pre-specified survival distribution to replace $w(t)$ when estimating the variance of $G^{\rho,\gamma}$. This observation is consistent with Gillen and Emerson [2007], indicating that the precision of the statistical information can be greatly affected by the drastic changes to our censoring distribution. However, the use of such pre-

specified survival weights is less appealing since they are not based on the current clinical data. Such apparent reversal of final statistical information can present challenges in the adaptive setting when applying CHW to control our overall Type 1 error.

Appendix F

Additional Results for Chapter 7

F.1 Asymptotic Properties of the Composite Statistics

O'Brien [1984] examined the general setting of applying a global test statistic to combining endpoints to test for treatment differences in similar direction. The use of the generalized (weighted) least squares statistics in O'Brien [1984] has the appealing property of the best linear unbiased estimates property under the Gauss Markov Theorem. His procedure was contrasted to the Hotelling's T^2 statistics and the rank sum procedure that were both directed at addressing the hypothesis of whether one or more treatments are different. The latter tests lack specific direction on which treatment were favorable. Subsequently, Pocock et al. [1987] extended O'Brien [1984]'s method of combining means to additional scenarios such as combining both binary or survival endpoints or any combination of test statistics that have asymptotically normal distributions. The linear composite and quadratic tests based on Logan et al. follow similar principles as O'Brien [1984].

In realistic settings, it is unlikely that all m endpoints are independent. For purpose of characterizing the properties of the composite statistics, we can envision that for the same endpoint of interest, the different test statistics applied separately on the disjoint time intervals can be combined linearly as motivated by Logan et al.. Hence, the data/events prior to the pre-specified crossing time and after this pre-specified crossing time can be considered time-wise independent as argued by Logan et al. [2008].

The adaptive strategy proposed in the literature considered switching the endpoints from one stage of the data collection to another stage (such as progression free survival to overall

survival). By adaptively switching between endpoints that are correlated, the resulting test statistic can be correlated across different stages of the data collection, inducing a correlation between test statistics. When such data from different stages are combined without acknowledging the underlying correlation, this can inflate the overall Type 1 error [Bauer and Posch, 2004, Jenkins et al., 2011, Irle and Schäfer, 2012, Magirr et al., 2016]. One of the appealing property of Logan et al. [2008]'s proposal is that data can be combined independently by first pre-specifying the time of crossing, τ, and then using the Nelson-Aalen on data prior to time of crossing and the weighted log-rank test after crossing. By this partitioning of the data, the resulting test statistics are independent.

We describe the asymptotic behavior of Logan's composite statistics by evaluating the alternative hypothesis conditional on this pre-specified crossing. Using the asymptotic properties for the individual components, this allow us to characterize the power of the composite statistics and better understand the relative behavior of the composite statistics in this bivariate parameter space. The asymptotic behavior of the composite statistics is limited since in realistic settings, as with many clinical trials, there are administrative, accrual and logistical constraints such that we do not have complete follow-up of all subjects in the study. Simulations are used to examine how these alternatives (in particular non proportional hazards alternatives discussed in section 7.4) may vary when subjected to censoring, as naturally induced with the use of interim analyses, and/or different accrual patterns.

F.1.1 Formulation under Local Alternatives

Let $\delta_1, \cdots, \delta_m$ be any constants such that $\delta = \delta_1^2 + \cdots + \delta_m^2$ [Shorack, 2010]. Then, our non-central $\chi^2_m(\delta)$ distribution with m degrees of freedom can be expressed as some function of the standard normal using some alternative δ by writing

$$\sum_{k=1}^{m}(Z_k + \delta_k)^2 \cong (Z_1 + \sqrt{\delta})^2 + \sum_{k=2}^{m} Z_k^2 \cong \chi^2_m(\delta)$$

with mean and variance

$$\mathbf{E}\left[\chi^2_m(\delta)\right] = m + \delta$$
$$\mathbf{Var}\left[\chi^2_m(\delta)\right] = 2m + 4\delta$$

In the setting of uncensored continuous outcome, let X_{0ik} and X_{1ik} be potential, independent observations measured from treatment group 0 and treatment group 1 respectively where i denote the i^{th} observation taking values $1, \cdots, n$, and k to denote the k^{th} endpoint for $k = 1, \cdots, K$. Also, let X_{0ik} and X_{1ik} have mean ω_{0k} and ω_{1k} respectively with common variance $V_k/2$. Assume that our target parameter of interest for the k^{th} endpoint, $\theta_k = \omega_{1k} - \omega_{0k}$, is the difference in the mean of the responses comparing subjects randomized to group 1 relative to subjects randomized to group 0. Then, an unbiased estimator of the treatment effect θ_k is $\widehat{\theta}_k = n^{-1}\sum_{i=1}^{n}(X_{1ik} - X_{0ik})$ with the associated (known) variance V_k/n can be computed for each endpoint of interest. Similar to O'Brien [1984], we can consider a linear combination of the K independent endpoints, and define the estimator of interest as $\mu = \sum_{k=1}^{K} w_k\theta_k$.

Appealing to asymptotics, each estimator $\widehat{\theta}_k$ is asymptotically $\mathcal{N}(\theta_k, V_k/n)$, for $k = 1, \cdots, K$. Thus, our (weighted) linear combination statistic has an asymptotic normal distribution as follows

$$\sum_{k=1}^{K} w_k\widehat{\theta}_k \;\dot{\sim}\; \mathbf{N}\left(\sum_{k=1}^{K} w_k\theta_k \;,\; n^{-1}\sum_{k=1}^{K} w_k^2 V_k\right)$$

with the $\sum_{k=1}^{K} w_k = 1$.

A suitable hypothesis of interest would be the test of $\mathbb{H}_0 : \sum_{k=1}^{K} w_k\theta_k = 0$ vs $\mathbb{H}_1 : \sum_{k=1}^{K} w_k\theta_k \neq 0$. We can define the test statistic $Z = \sum_{k=1}^{K} w_k\hat{\theta}_k/\sqrt{V_T}$. Under the null, Z has mean 0 and variance $V_T = \sum_{k=1}^{K} w_k^2 V_k/n$ which is standard normal.

Each of the θ_k is independent of another, the mean and variance can be estimated independently. Thus, if we assume further that our outcome in the continuous setting was coming

from a normal distribution, then V_k as estimated under $\mathbb{H}_0$ in either the Score version of the test statistic, Wald, or Likelihood Ratio test would be equivalent to V_T. However, when the endpoints are correlated such that either a mean variance relationship exists or some correlation is induced across different endpoints (such as binary or survival endpoints or repeated measurements), then our estimated variance evaluated under the null hypothesis can differ from the estimated variance evaluated based on the MLE. The use of asymptotic normality mitigates part of the difficulty in the survival setting by allowing us to quantify approximately how the alternatives of the composite statistics may behave under the ideal scenario of immediate accrual when we have complete follow up of all subjects in the population at each analysis time.

We apply the above extension to describe the asymptotic alternatives using the composite statistics as described by Logan et al. [2008] in the context of time-to-event endpoint. Let τ_0 be the time of crossing. Thus, if we set $\tau_0 = 0$, the composite test reduces to the usual log rank test statistics. Alternatively, we can simply use the Nelson-Aalen test statistic solely at the end of the trial if we are interested in the difference in survival probability at a prespecified time point.

Under $\mathbb{H}_0$, the composite test statistics have asymptotic distributions of the following

$$\begin{aligned}
\frac{Z_{NA}(\tau_0, t) + Z_{LR}(\tau_0, t)}{\sqrt{2}} &\doteq \mathcal{N}(0, 1) \\
\left[Z_{NA}(\tau_0, t)\right]^2 + \left[Z_{LR}(\tau_0, t)\right]^2 &\doteq \chi^2_2 \\
w_1 Z_{NA}(\tau_0, t) + w_2 Z_{LR}(\tau_0, t) &\doteq \mathcal{N}(0, w_1^2 + w_2^2)
\end{aligned}$$

The last equation represents the linear composite statistic that is weighted by w_1, w_2 respectively for each test statistic computed before and after the time of crossing.

Under local alternatives, the individual components of the composite statistics has asymp-

totic distribution of the form

$$
\begin{aligned}
Z_{NA}(\tau_0, t) &\stackrel{\cdot}{\sim} \mathcal{N}(\delta_1, 1) \\
Z_{LR}(\tau_0, t) &\stackrel{\cdot}{\sim} \mathcal{N}(\delta_2, 1) \\
Z_{NA} \stackrel{\cdot}{\sim} \mathcal{N}(\theta_{NA}/\sqrt{V_{NA}}, 1) &\cong Z_{NA}^2 \stackrel{\cdot}{\sim} \chi_1^2(\delta_{NA}^2 = \theta_{NA}^2/V_{NA}) \\
Z_{LR} \stackrel{\cdot}{\sim} \mathcal{N}(\theta_{LR}/\sqrt{V_{LR}}, 1) &\cong Z_{LR}^2 \stackrel{\cdot}{\sim} \chi_1^2(\delta_{LR}^2 = \theta_{LR}^2/V_{LR})
\end{aligned}
$$

Thus, the composite statistics have asymptotic distributions of the form

$$
\begin{aligned}
Z_{\text{OLS}}^{Opt} = \frac{w_1 Z_{NA}(\tau_0, t) + w_2 Z_{LR}(\tau_0, t)}{\sqrt{w_1^2 + w_2^2}} &\stackrel{\cdot}{\sim} \mathcal{N}\left(\frac{w_1\delta_1 + w_2\delta_2}{\sqrt{w_1^2 + w_2^2}}, 1\right) \\
Z_{\text{OLS}} = \frac{Z_{NA}(\tau_0, t) + Z_{LR}(\tau_0, t)}{\sqrt{2}} &\stackrel{\cdot}{\sim} \mathcal{N}\left(\frac{\delta_1 + \delta_2}{\sqrt{2}}, 1\right) \\
Z_{\text{Quad}} = \left[Z_{NA}(\tau_0, t)\right]^2 + \left[Z_{LR}(\tau_0, t)\right]^2 &\stackrel{\cdot}{\sim} \sum_{i=\{NA,LR\}} (Z_i + \delta_i)^2 \cong \chi_2^2(\delta = \delta_1^2 + \delta_2^2)
\end{aligned}
$$

Therefore, the quadratic versions for Z_{OLS}^{Opt}, and Z_{OLS} can be represented asymptotically as

$$
\begin{aligned}
(Z_{\text{OLS}}^{Opt})^2 = \left(\frac{w_1 Z_{NA}(\tau_0, t) + w_2 Z_{LR}(\tau_0, t)}{\sqrt{w_1^2 + w_2^2}}\right)^2 &\cong \chi_1^2\left(\frac{(w_1\delta_1 + w_2\delta_2)^2}{w_1^2 + w_2^2}\right) \\
Z_{\text{OLS}}^2 = \left(\frac{Z_{NA}(\tau_0, t) + Z_{LR}(\tau_0, t)}{\sqrt{2}}\right)^2 &\cong \chi_1^2\left(\frac{(\delta_1 + \delta_2)^2}{2}\right)
\end{aligned}
$$

We shall refer to these alternatives $\delta_1, \delta_2, \delta, \delta_{NA}$, and δ_{LR} as standardized alternatives for Nelson-Aalen restricted to time of crossing using complete follow up until time t, truncated logrank statistic between time of crossing and t, quadratic test, Nelson Aalen test at time t, and the overall logrank test at time t, since they are a function of the parameter of interest as well as the precision of the test statistic.

We can characterize the probability of rejecting the null hypothesis using the composite statistics by considering the bivariate parameter space spanned by the standardized alternatives based on each test statistic. We note that such characterization assumes rejecting the

composite null hypothesis when using the composite statistics. The use of individual components of the composite statistics rejects the null hypothesis of no difference in survival.

Regions that represent the same such probability of rejecting the null for the quadratic statistic can be described via the equation $\delta_1^2 + \delta_2^2 = r_q$, where r_q is the square of the radius of the contour such that we have $\Pr(\chi_2^2(\delta_1^2 + \delta_2^2) > \chi_{2,\alpha}^2(0)) = q$. For the linear composite statistics, the equation $(\delta_1 + \delta_2)/\sqrt{2} = r_l$ would consist of the set of lines for which the alternatives have the same probability of rejecting the null hypothesis $\Pr(Z_{OLS} > Z_\alpha | \sqrt{\delta} = (\delta_1 + \delta_2)/\sqrt{2})$ when using the appropriate critical value under $\mathbb{H}_0$. For example, at 5% level under the null, we can obtain the critical value, $\chi_2^2(0) = 5.991465$ for the quadratic statistic. For the linear composite statistic, the two-sided test under the null for the $\alpha = 0.05$ would define a critical value of $Z_\alpha = \pm 1.959964$.

We can solve the pair of simultaneous equation to obtain the combination of alternatives such that the linear and quadratic statistics have equivalent probability of rejecting the null hypothesis. By substituting $\delta_1 = r_l\sqrt{2} - \delta_2$ into the pair of equations, we have $\delta_2^2 = r_q - (r_l\sqrt{2} - \delta_2)^2$, giving $\delta_2 = \frac{\sqrt{2}}{2} r_l \pm \sqrt{\frac{r_q - r_l^2}{2}}$.

F.1.2 Asymptotic Probability of Rejecting $\mathbb{H}_0$ for the Composite Statistics under Local Alternatives

Using the asymptotic results, we generated regions in the two dimensional space and characterize regions when the composite statistics may have higher probability of rejecting the null hypothesis over other test statistics for the same value of the standardized alternative. Without loss of generality, we shall refer to δ as our alternatives for simplicity. This can allow us to characterize the asymptotic probability of rejecting $\mathbb{H}_0$ based on the bivariate space defined previously.

We compare Logan's statistics with the procedure of using a single test statistic to address the primary question of whether the treatment is superior as compared to placebo. The choice of a single test statistic such as either the log-rank test, or the Nelson Aalen statistics addresses different scientific questions. This standardized alternative chosen from picking

either the log-rank or Nelson Aalen statistic, thus describes different power profiles as compared to the same value of this standardized alternative computed based on the composite tests.

Under $\mathbb{H}_0$, at level α=0.025, for the linear composite test statistics, we set our critical value based on the asymptotic standard normal distribution, $N(0,1)$. We reject $\mathbb{H}_0$ if $Z > \Phi_{0.975} = 1.9599$. For the quadratic test, we set our critical value based on the χ^2_2 distribution, using a two-sided level $\alpha' = 2\alpha = 0.05$. We thus reject the null hypothesis if $Z_{Quad} > 5.991465$. In addition, we assume positive values of δ to be consistent with concluding the experimental treatment is superior relative to the placebo arm.

F.1.2.1 Probability of rejecting $\mathbb{H}_0$ for composite statistics using different α's

We evaluated the probability of rejecting $\mathbb{H}_0$ based on the bivariate space for the linear composite, quadratic, and the individual components of the composite statistics computed based on either 1-sided $\alpha = 0.025$ or 0.005 (Figure F.1). The corresponding α' for the quadratic statistics are 0.05 or 0.01 respectively.

In the left figure, at level $\alpha = 0.025$, the gold region, as bounded by the grey lines, represents 90% probability of rejecting the null hypothesis based on the quadratic test. The baby blue lines represents $> 90\%$ probability of rejecting the null hypothesis based on the linear composite statistics while the dotted lines represents 90% probability of rejecting the null hypothesis based on either Z_{NA} or Z_{LR}. Note that despite this characterization, we acknowledge that each test statistic is addressing a different null hypothesis.

In this bivariate space, to achieve at least 90% power based on the alternative $\delta_1 = 3.24$, the linear composite statistic would require an alternative $\delta_2 \geq 1.34$. However, for the quadratic statistic, when we hold $\delta_1 = 3.24$ fixed, δ_2 has to be at least greater than 1.465 in order to ensure the same 90% probability of rejecting $\mathbb{H}_0$. The use of the single test statistic, Z_1, or Z_2, has at least 90% power to detect an alternative of either $\delta_1 = 3.24$ or $\delta_2 = 3.24$ as denoted by the dotted lines. The right figure characterizes how the alternative changes as we hold fixed the probability of rejecting the null hypothesis and vary our α to 0.005.

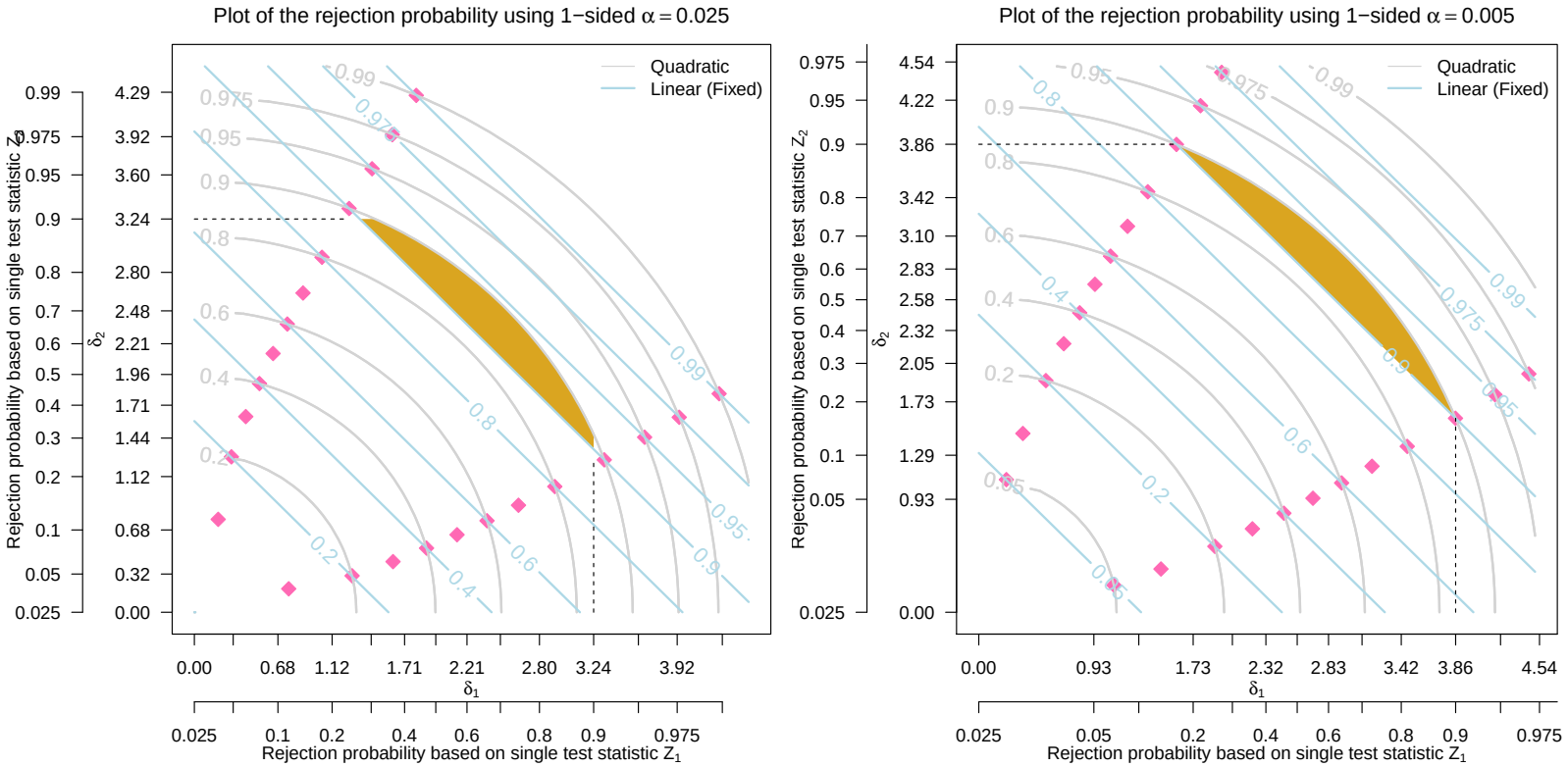

Figure F.1: Contour plots of the probability of rejecting $\mathbb{H}_0$ based on standardized alternative when assuming a 1-sided $\alpha = 0.025$ or 0.005. The critical value of the χ^2_2 distribution is based on 2α since no direction is provided for the p-value. The x-axis (y-axis) comprised of labels corresponding to the alternative δ_1 (δ_2), the probability of rejecting the null hypothesis obtained from the use of only Z_{NA} (or Z_{LR}) statistic. Changes are observed in the contour regions of standardized alternatives having similar probability of rejecting the null hypothesis as we modify our α level.

The results shown in quadrant I have interpretations consistent with the direction of the standardized alternatives. When our standardized alternatives are in quadrant I or III, we can interpret this probability of rejecting the null hypothesis as the power of the test statistics since we know the ideal treatment. However, in quadrant II/IV, we interpret the contour lines for the quadratic test or other tests as the probability of rejecting $\mathbb{H}_0$ without imposing strictly which alternative to be correct. In such situations, we may not be able to quantify the better alternative unless we seen the survival curves or have some prior knowledge on what is the decision rule leading us to select the better treatment.

F.1.2.2 Comparison of the probability of rejecting $\mathbb{H}_0$ with a single test statistic

By symmetry of the bivariate parameter space, we describe the rejection probability for the composite statistics in Quadrant I and IV. Quadrant I can be seen to describe the behavior of the composite statistics when the alternatives for both the Nelson-Aalen test at time of crossing, and the weighted log-rank test statistics after time of crossing are consistently picking the experimental arm. We define positive values of the standardized alternatives to be consistent with the experimental (Exp) treatment arm being superior relative to the placebo (Ctrl) arm whereas the negative values of the standardized alternatives are consistent with the placebo arm being superior over the experimental treatment arm. Thus, when we presume superiority of experimental treatment over the placebo such that $\log[\Lambda_{Ctrl}(t)/\Lambda_{Exp}(t)] > 0$, this translates to the Nelson Aalen hypothesis of testing if $S_{Exp}(t) > S_{Ctrl}(t)$ or equivalently $\log \Lambda_{Exp}(t) < \log \Lambda_{Ctrl}(t)$. Similarly, for the truncated log rank statistics, we are thus interested in the alternative $\sum_{t \geq \tau_0} [\Lambda_{Ctrl}(t) - \Lambda_{Exp}(t)] > 0$.

The asymptotic rejection probability for the linear composite statistics can be described using the bivariate parameter space that is symmetric about the axis $\delta_1 = \delta_2$ (Figure F.1). For the quadratic statistic, the asymptotic probability of rejecting $\mathbb{H}_0$ can be parametrized in the form of a circle with rings describing similar probability of rejecting the null hypothesis. In Figure F.2 and F.3, a second x-axis can be used to characterize the standardized alternatives that is mapped from assuming the use of a test statistic picked from either Z_{NA}, or Z_{LR}.

Consider the contour lines that represent 90% probability of rejecting the null in quadrant I (See Figure F.2). The shaded gray region characterizes the set of alternatives, i.e., δ_1 and δ_2, for which the use of the single test statistic gives a higher probability of rejecting the null as opposed to using the linear or quadratic statistic. The baby blue region indicates that the linear composite statistic has higher probability of rejecting the null when the two alternatives have magnitudes in the same direction over the single or quadratic statistics. Combining them in an equally weighted manner provides a synergistic effect to increase such probability of rejecting the null hypothesis over the use of either the single test statistic or the quadratic statistic. Compared to the linear or single test statistic, the quadratic test statistic rejects less frequently.

When one of the alternatives is weaker (as indicated by the gray region), for example, if $\delta_1 = 3.24$ and $\delta_2 = 0.32$, or $\delta_2 = 3.24$ and $\delta_1 = 0.32$, then both the linear and quadratic composite statistics have relatively low probability of rejecting the null hypothesis compared to having picked the right test statistic to detect this joint difference. In this case, the probability of rejecting the null is higher for the quadratic statistic over the linear statistic in this shaded region. However, this probability of rejecting the null based on the quadratic statistic is still lower than having picked the right single test statistic in the first place.

In Quadrant IV (as in Figure F.3), the combination of bivariate alternatives can be described as antagonistic. This has a direct impact on the probability of rejecting $\mathbb{H}_0$ particularly for the linear composite statistics whereby the stronger alternative effect is weakened when weighted equally with a standardized alternative in the opposite direction. In the context of the composite statistics, we have high probability in detecting a difference in survival such that the treatment is superior to the placebo. However, after the time of crossing, this estimate of the alternative based on the truncated log rank test statistic provides evidence in favor of the placebo being superior to the treatment. This switching of the preferential treatment relates most closely to the understanding the behavior of the individual components of the composite statistics in this partially ordered hazards space. Here, we see that the truncated logrank statistic may not weigh the difference in hazards appropriately since

the history of prior survival is ignored. This consequence is illustrated using simulations in finite sample settings in section 7.4 and characterized further in Appendix F.3.2.

In Quadrant IV, the single test statistic is dominant over a smaller region when the treatment alternative in the opposing direction is relatively small. Since the quadratic statistic ignores the direction of alternatives, there is high probability of rejecting $\mathbb{H}_0$ even when our alternatives are antagonistic. This gives a bigger shaded region that provides the appearance of high "statistical power" to claim a crossing as in Logan et al. [2008].

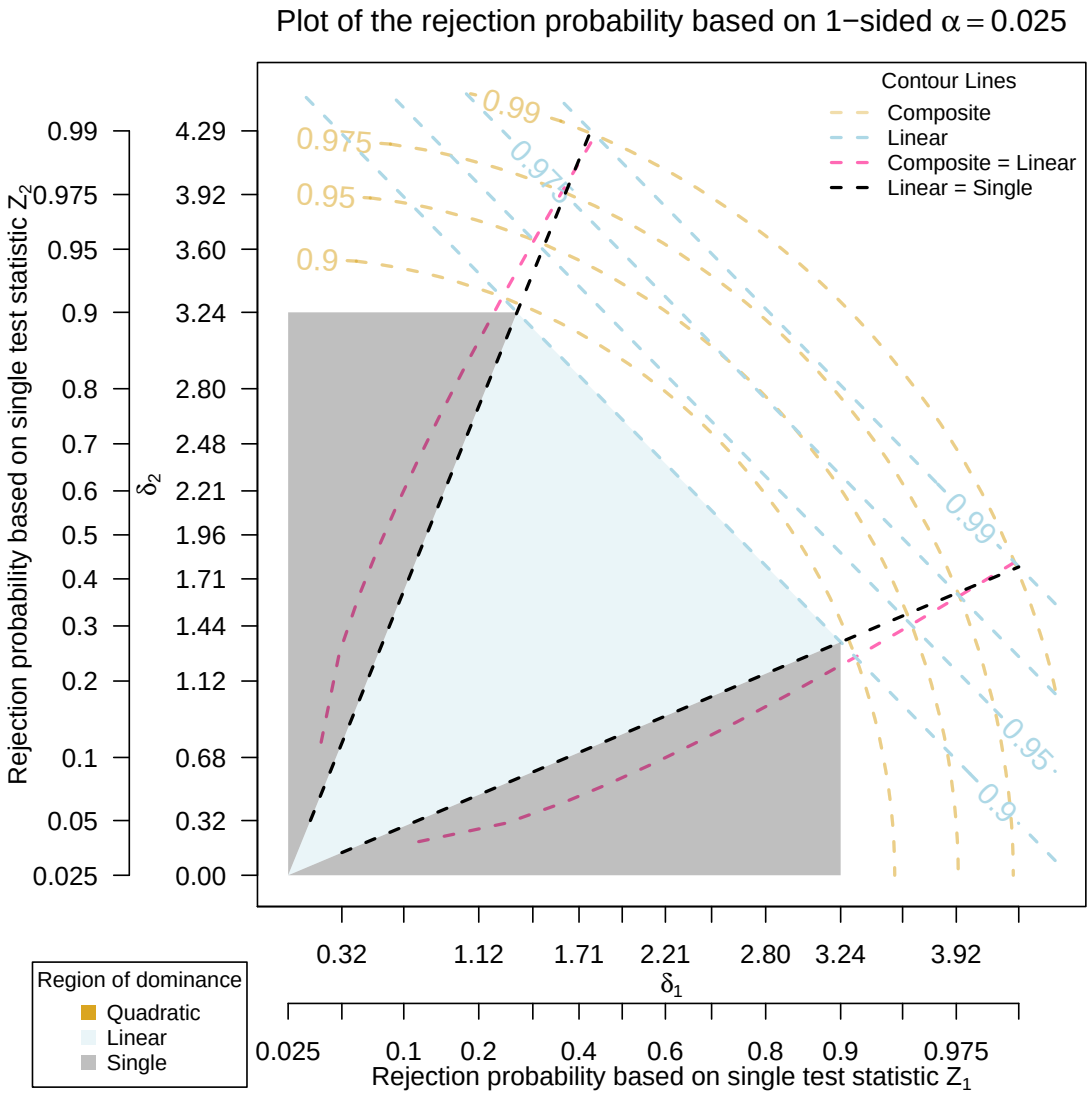

Figure F.2: Contour plot of the probability of rejecting $\mathbb{H}_0$ based on the alternatives in Quadrant I when assuming a 1-sided $\alpha = 0.025$ (This is similar to Quadrant III). The critical value of the χ^2_2 distribution is based on 2α since no direction is provided for the p-value. Using probability of rejecting $\mathbb{H}_0$ at $\leq 90\%$ as benchmark, the shaded light-blue region corresponds to the parameter space whereby the linear composite statistic dominates over the other two statistics. The light gray region corresponds to the parameter space for which the rejection probability of the single statistic (when chosen correctly) dominates. There is no region for which the quadratic statistic has higher probability of rejecting $\mathbb{H}_0$ when compared to the single test statistic or the linear composite statistic.

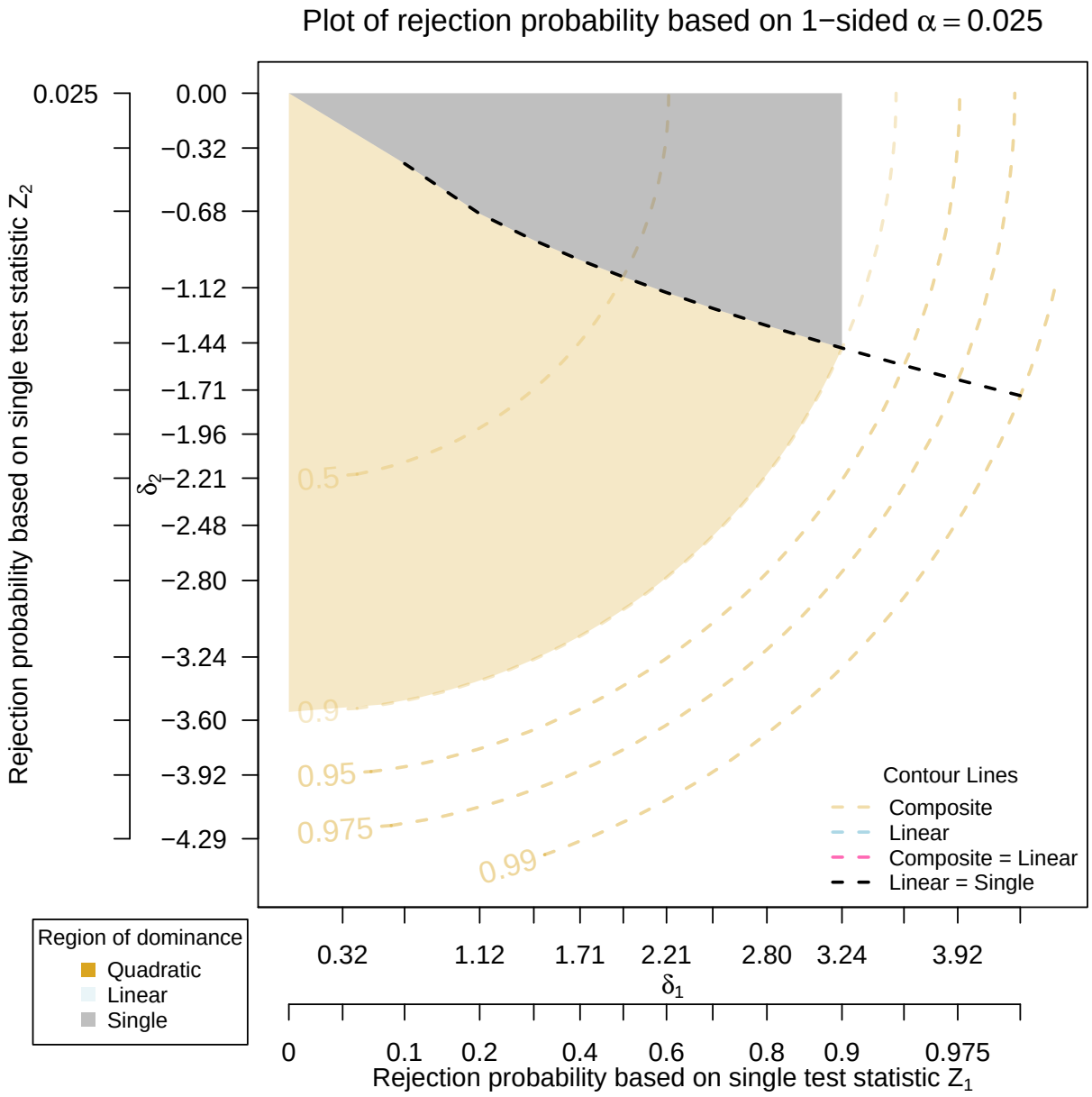

Figure F.3: Contour plot of the probability of rejecting $\mathbb{H}_0$ based on alternatives in Quadrant IV when assuming a 1-sided $\alpha = 0.025$ (This is similar to Quadrant II). The quadratic statistic has a bigger region, which highlights a higher probability of rejecting $\mathbb{H}_0$, relative to the use of the linear composite statistics or the single test statistic. Note that the quadratic statistic ignores the directions of the standardized alternatives.

F.1.2.3 Comparison of the Probability of Rejecting $\mathbb{H}_0$ with the Bonferroni Corrected Test

An alternative (and less common) strategy in the clinical setting involved testing multiple endpoints. This procedure involves choosing the minimum of the two p-values computed from two statistical tests, each testing the superiority of the treatment over the placebo. A Bonferroni multiplicity correction, applied for each test at level $\alpha/2$, is then applied to determine whether we reject the null hypothesis of interest. In our context, we may interpret this as using both the logrank and Nelson Aalen test statistic, and starting follow-up from time of first randomization until the end of the trial, and applying the Bonferroni adjustment procedure.

We present the numerical results as previous, and further include the comparison with the Bonferroni corrected test statistic in Figure F.4 and F.5. A third x-axis is added to describe the cumulative probability of rejecting $\mathbb{H}_0$ (adjusted for multiplicity) for the alternative based on the x-axis. In contrast, the second x-axis describes the cumulative probability of rejecting $\mathbb{H}_0$ based on the single test statistic similar to the previous section. By comparison between the second and third x-axis, the use of the single test statistic has a higher probability of rejecting $\mathbb{H}_0$ as compared to the Bonferroni corrected test. The gray region serves as a comparison from the previous section relative to using the Bonferroni corrected test. With a multiplicity adjustment using the Bonferroni correction, the alternative providing at least 90% probability of rejecting $\mathbb{H}_0$ would need to be be much further away than 3.24.

The use of the Bonferroni corrected test has higher probability of rejecting $\mathbb{H}_0$ relative to applying the composite statistics when one of the alternative is indicative of a strong treatment effect, while the other alternative, in the same direction, is relatively weaker. The advantage of the Bonferroni corrected test is minimal since when one of the alternatives becomes weaker, the quadratic statistic dominates in terms of higher probability of rejecting the null. In other words, either the Bonferroni corrected, or quadratic test has sufficiently high probability of rejecting $\mathbb{H}_0$ so long as one of the alternatives has sufficiently high prob-

ability to reject the null hypothesis. Picking the Bonferroni corrected procedure thus does not present an advantage when both of the alternatives have moderate effect.

The linear composite test procedure presents an advantage over the Bonferroni corrected or quadratic test when alternatives are synergistic, and are moderately effective as shown in the blue region. Within the blue region, the linear combination test characterized a bigger bivariate space where there is higher probability of rejecting $\mathbb{H}_0$ over the quadratic or Bonferroni correct test.

In the previous section, clinicians are offered the choice of only a single test statistic to select the better treatment. Picking a single, correct test statistic does not result in regions whereby the quadratic test can dominate. Here, the Bonferroni corrected procedure allows clinicians to choose between two potentially important endpoint by correcting for multiplicity. The multiplicity correction allows a fairer comparison with the composite statistics even though we may be quantifying different alternatives with the test statistics.

In Quadrant IV, the quadratic test now dominates majority of the bivariate space (Figure F.5). In contrast to Figure F.4, the alternative δ_1 has to be more extreme in order for the Bonferroni corrected test to gain advantage over the quadratic test. Similarly, the quadratic test has high probability of rejecting $\mathbb{H}_0$ even when δ_1 is weak so long as δ_2 is highly negative.

In Quadrant IV, the Nelson-Aalen test has sufficiently high probability of rejecting the null hypothesis in favor of the experimental treatment arm even when the alternative for the truncated log-rank test is contrary to the conclusions by the Nelson-Aalen test at time of crossing. The Bonferroni corrected test picks the experimental arm over the placebo arm with high probability of concluding the placebo arm as superior over the experimental arm. Note that in this space, for any pair of δ's along the line of $\delta_1 = -\delta_2$, the linear composite test has a level of 0.025 probability of rejecting $\mathbb{H}_0$. The quadratic test has the "advantage" of summing the square of these alternatives to produce sufficiently higher probability of rejecting $\mathbb{H}_0$ over the linear composite test. This resultant procedure is consistent with the claim by Logan et al. [2008] where the quadratic test has high "power" to detect a crossing and lack interpretability on the preferred treatment.

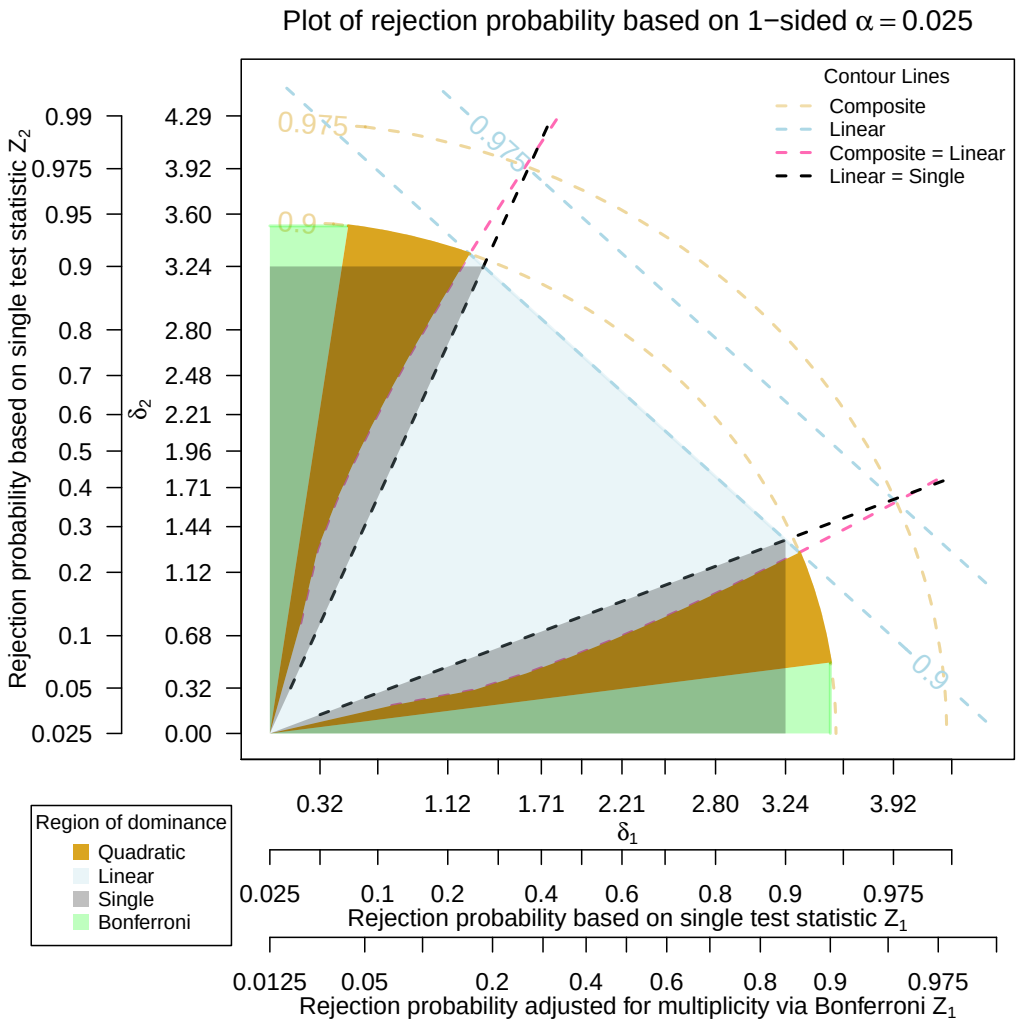

Figure F.4: Contour plot of the probability of rejecting $\mathbb{H}_0$ in Quadrant I for the composite statistics and the Bonferroni correction for multiple testing based on the standardized alternatives in Quadrant I when assuming a 1-sided $\alpha = 0.025$ (This is similar to Quadrant III). The critical value of the χ^2_2 distribution is based on 2α since no direction is provided for the p-value. With rejection probability of $\leq 90\%$, the shaded light-blue region corresponds to the parameter space for which the rejection probability of the linear composite statistic dominates. Incorporating the Bonferroni test statistic, the gold region corresponds to the parameter space for which the quadratic test has higher rejection probability over linear or Bonferroni corrected test. The green region represents the region for which the Bonferroni corrected test has higher rejection probability over either linear or quadratic test. The light gray region corresponds to the parameter space for which the rejection probability of the single statistic (chosen correctly) dominates.

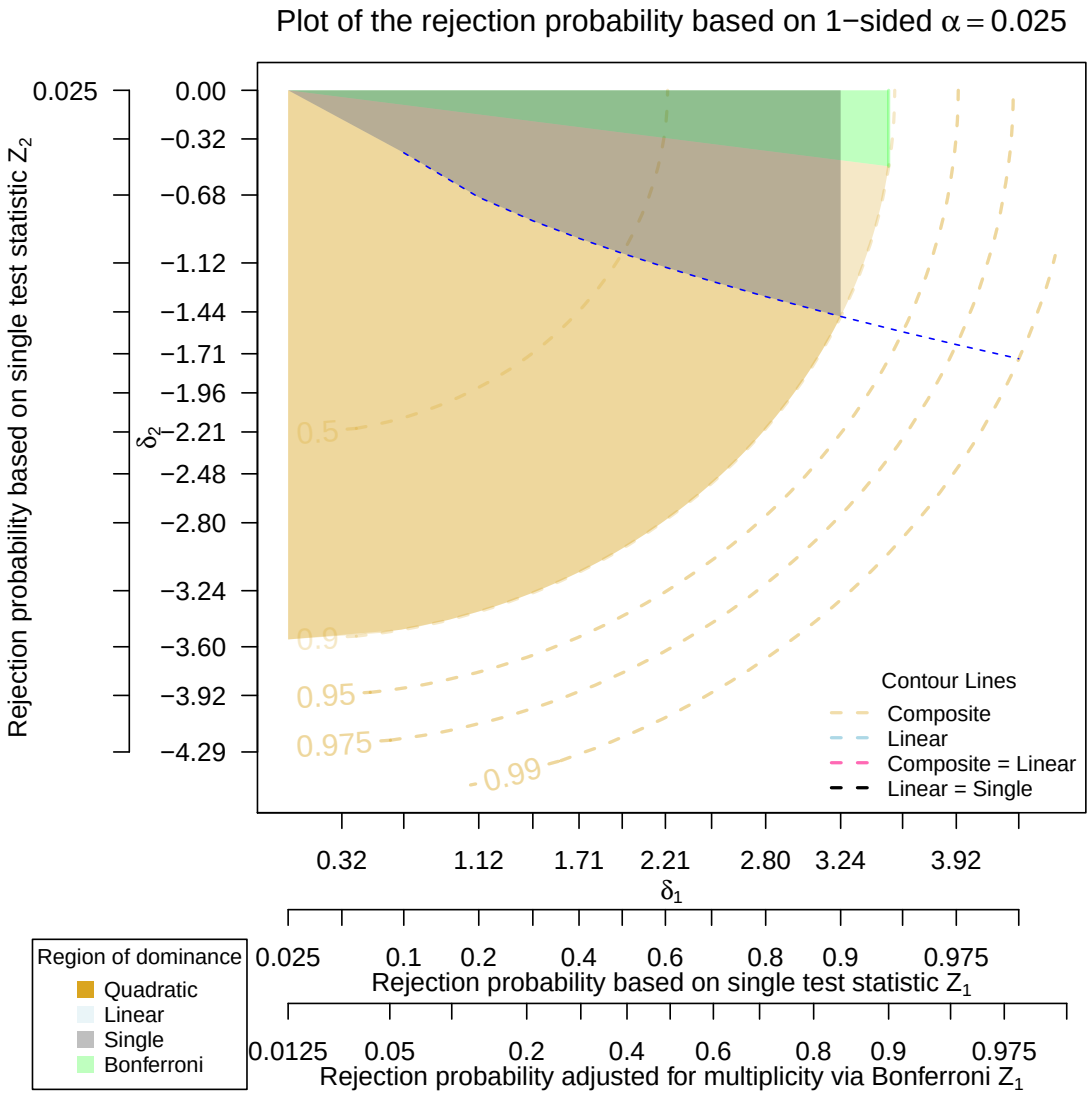

Figure F.5: Contour plots of the probability of rejecting $\mathbb{H}_0$ in Quadrant IV for the composite statistics and the Bonferroni correction for multiple testing based on the standardized alternatives in Quadrant IV when assuming a 1-sided $\alpha = 0.025$ (This is similar to Quadrant II). The critical value of the χ^2_2 distribution is based on 2α since no direction is provided for the p-value. With rejection probability $\leq 90\%$, the shaded light-blue region corresponds to the parameter space for which the rejection probability of the linear composite statistic dominates. In this quadrant (respectively in quadrant II), the quadratic test has the appearance of being more "powerful" over the other test statistics purely as a consequence of picking the magnitude of the alternatives and ignoring the direction of the treatment effect.

F.2 Additional Results for Section 7.3.1

F.2.1 Additional Simulation Results for Stochastic Ordered, Crossing Hazards Survival Curves

The results in section 7.3.2 (Table F.1) are seen to be affected by patterns of accrual. In comparison, with censoring, a higher proportion of curves (27.1% relative to 17.2%) has spurious crossings when analyzed at time 2. One direct consequence of accrual is that, by time 5, on average, a lower number of events is observed. Although this proportion of survival curves crossing by time 5 is similar, there is a higher proportion of survival curves that crosses at earlier calendar time (for example, time 2).

The probability of rejecting the null hypothesis for other common test statistics under the censored setting (Table F.2) was generally similar to the results with immediate accrual (Table 7.3). The results for the composite statistics were also similar (99%) with high probability of rejecting the null hypothesis incorrectly in favor of A in the censored setting.

Table F.1: Summary statistics at various calendar time where survival curves are stochastically ordered without true crossings over the first five years with uniform accrual over a 3 year period based on 10,000 simulation. Treatment group B is the preferred treatment in terms of survival probability at all times relative to group A.

		Stochastically ordered survival curves[@] (Crossing Hazards)				
		$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Accrual over 3 years	Total number of Events	231 (14)	558 (20)	898 (22)	1009 (22)	1026 (22)
	Events (B vs A)	86 vs 145	241 vs 317	409 vs 489	494 vs 515	511 vs 515
	Number of events $\geq \tau_0$	0 (0)	0 (0)	3 (2)	10 (3)	16 (4)
	Events (B vs A)	0 vs 0	0 vs 0	3 vs 0	10 vs 0	16 vs 0
	$\text{HR}_{\text{Ref: B}}$ ($\pm$logSD)	2.09 (0.135)	1.65 (0.085)	1.49 (0.067)	1.31 (0.064)	1.26 (0.063)
	$\text{RMS}_A(t)^{\ddagger}$	0.44 (0.019)	0.93 (0.033)	1.41 (0.043)	1.90 (0.058)	2.38 (0.074)
	$\text{RMS}_B(t)^{\ddagger}$	0.59 (0.015)	1.15 (0.030)	1.65 (0.040)	2.14 (0.052)	2.62 (0.066)
	$\widehat{S}_A(t)$	0.4858 (0.03)	0.4849 (0.02)	0.4847 (0.02)	0.4848 (0.02)	0.4848 (0.02)
	$\widehat{S}_B(t)$	0.5888 (0.06)	0.5057 (0.03)	0.4890 (0.02)	0.4856 (0.02)	0.4850 (0.02)
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)^{\dagger}$	5.6	27.1	42.6	48.0	49.0

†: Percentage of times a crossing is observed.

‡: The restricted mean statistic is truncated to 3 months just prior to the calendar time t.

Table F.2: Table of the statistically significant results for level $\alpha = 2.5\%$ based on 10,000 simulations under the setting where survival curves are stochastically ordered without true crossings over the first five years under uniform accrual of subjects over 3 years. The columns, A and B, indicate the total number of times the test statistic concludes the trial in favor of the treatment group A or B.

				Crossing at time 2				Crossing by time 5			
		Overall		Crossing $\widehat{S}_A(2) > \widehat{S}_B(2)$ n= 2712		No Crossing $\widehat{S}_A(2) < \widehat{S}_B(2)$ n= 7288		Crossing $\widehat{S}_A(5) > \widehat{S}_B(5)$ n= 4904		No Crossing $\widehat{S}_A(5) < \widehat{S}_B(5)$ n= 5096	
Statistic	Sig	A	B	A	B	A	B	A	B	A	B
$Z_{\text{LR}}(5)$	9529	0	9529	0	2346	0	7183	0	4433	0	5096
$Z_{\text{NA}}(5)$	506	264	242	215	3	49	239	264	0	0	242
$Z_{\text{RMS}}(5)$	6815	0	6815	0	1041	0	5774	0	1720	0	5095
$Z_{\text{NA}}(\tau_0)$	1183	30	1153	30	0	0	1153	30	132	0	1021
$Z_{\text{LR}}(\tau_0, 5)$	9998	9998	0	2711	0	7287	0	4904	0	5094	0
$Z_{\text{OLS}}(5)$	5929	5929	0	2290	0	3639	0	4825	0	1104	0
$Z_{\text{Quad}}(5)$	9954	9954	9954	2693	2693	7261	7261	4884	4884	5070	5070

F.2.2 Simulation Results for Crossing Survival Curves

Table F.3 shows the summary statistics at each calendar time under the fixed sample setting when we have crossing survival, and crossing hazards in Figure 7.1. With patients being accrued uniformly, the common summary statistics, such as $S(t)$, restricted mean up to time t, tend to have more variation even though on average, they are consistent with summary statistics obtained in the immediate accrual setting. The proportion of crossing at time 2 is higher when we do not have immediate accrual.

Table F.3: Summary statistics based on 10,000 simulations for the crossing survival curves with either immediate accrual or uniform accrual of subjects over the first three years.

		Crossing survival curves				
		$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Immediate Accrual	Total number of Events*	827 (22)	1009 (22)	1099 (22)	1155 (22)	1194 (22)
	Events (B vs A)	331 vs 497	494 vs 515	584 vs 516	639 vs 516	678 vs 516
	Number of events $\geq \tau_0$*	0 (0)	0 (0)	90 (9)	146 (11)	185 (12)
	Events (B vs A)	0 vs 0	0 vs 0	89 vs 1	145 vs 1	184 vs 1
	$\text{HR}_{\text{Ref: B}}$ ($\pm$ logSD)	1.86 (0.071)	1.28 (0.064)	1.06 (0.062)	0.95 (0.062)	0.88 (0.061)
	$\text{RMS}_A(t)^{\ddagger}$	0.51 (0.009)	1.00 (0.024)	1.49 (0.040)	1.97 (0.056)	2.46 (0.072)
	$\text{RMS}_B(t)^{\ddagger}$	0.64 (0.007)	1.26 (0.020)	1.74 (0.033)	2.15 (0.046)	2.50 (0.059)
	$\widehat{S}_A(t)$	0.5032 (0.02)	0.4850 (0.02)	0.4841 (0.02)	0.4839 (0.02)	0.4837 (0.02)
	$\widehat{S}_B(t)$	0.6693 (0.01)	0.5058 (0.02)	0.4164 (0.02)	0.3612 (0.02)	0.3220 (0.01)
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)^{\dagger}$	0.0	17.4	99.8	100	100
Accrual over 3 years	Total number of Events*	184 (13)	494 (19)	847 (22)	1040 (23)	1121 (22)
	Events (B vs A)	62 vs 122	202 vs 292	383 vs 464	526 vs 514	605 vs 516
	Number of events $\geq \tau_0$*	0 (0)	0 (0)	16 (4)	56 (7)	112 (10)
	Events (B vs A)	0 vs 0	0 vs 0	16 vs 0	56 vs 1	111 vs 1
	$\text{HR}_{\text{Ref: B}}$ ($\pm$ logSD)	2.37 (0.156)	1.76 (0.091)	1.48 (0.069)	1.19 (0.063)	1.02 (0.062)
	$\text{RMS}_A(t)^{\ddagger}$	0.51 (0.017)	1.00 (0.032)	1.49 (0.042)	1.97 (0.056)	2.46 (0.072)
	$\text{RMS}_B(t)^{\ddagger}$	0.64 (0.013)	1.26 (0.028)	1.74 (0.038)	2.15 (0.048)	2.50 (0.060)
	$\widehat{S}_A(t)$	0.5041 (0.04)	0.4853 (0.02)	0.4842 (0.02)	0.4839 (0.02)	0.4837 (0.02)
	$\widehat{S}_B(t)$	0.6696 (0.06)	0.5068 (0.05)	0.4171 (0.04)	0.3617 (0.03)	0.3222 (0.03)
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)^{\dagger}$	2.0	29.4	97.4	100	100

†: Percentage of times a crossing is observed.

‡: The restricted mean statistic is truncated to 3 months just prior to the calendar time t.

Table F.4: Table of the statistically significant results for level $\alpha = 2.5\%$ based on 10,000 simulations under the setting of crossing survival curves scenario with immediate accrual. The columns, A and B, indicate the total number of times the test statistic concludes the trial in favor of the treatment arm.

				Condition on time 2				Condition on time 5			
		Overall		Crossing $\widehat{S}_A(2) > \widehat{S}_B(2)$ n= 1737		No Crossing $\widehat{S}_A(2) < \widehat{S}_B(2)$ n= 8273		Crossing $\widehat{S}_A(5) > \widehat{S}_B(5)$ n= 10,000		No Crossing $\widehat{S}_A(5) < \widehat{S}_B(5)$ n= 0	
Statistic	Sig	A	B	A	B	A	B	A	B	A	B
$Z_{\text{LR}}(5)$	5574	5574	0	1732	0	3842	0	5574	0	0	0
$Z_{\text{NA}}(5)$	10000	10000	0	1737	0	8263	0	10000	0	0	0
$Z_{\text{RMS}}(5)$	708	79	629	79	0	0	629	79	629	0	0
$Z_{\text{NA}}(\tau_0)$	1586	20	1566	20	0	0	1566	20	1566	0	0
$Z_{\text{LR}}(\tau_0, 5)$	10000	10000	0	1737	0	8263	0	10000	0	0	0
$Z_{\text{OLS}}(5)$	10000	10000	0	1737	0	8263	0	10000	0	0	0
$Z_{\text{Quad}}(5)$	10000	10000	10000	1737	1737	8263	8263	10000	10000	0	0

Table F.5: Table of the statistically significant results for level $\alpha = 2.5\%$ based on 10,000 simulations under the setting of crossing survival curves scenario with uniform accrual over 3 years. The columns, A and B, indicate the total number of times the test statistic concludes the trial in favor of the treatment arm.

				Condition on time 2				Condition on time 5			
		Overall		Crossing $\widehat{S}_A(2) > \widehat{S}_B(2)$ n= 2941		No Crossing $\widehat{S}_A(2) < \widehat{S}_B(2)$ n= 7059		Crossing $\widehat{S}_A(5) > \widehat{S}_B(5)$ n= 10000		No Crossing $\widehat{S}_A(5) < \widehat{S}_B(5)$ n= 0	
Statistic	Sig	A	B	A	B	A	B	A	B	A	B
$Z_{\text{LR}}(5)$	619	157	462	113	33	44	429	157	462	0	0
$Z_{\text{NA}}(5)$	9965	9965	0	2936	0	7029	0	9965	0	0	0
$Z_{\text{RMS}}(5)$	708	80	628	61	43	19	585	80	628	0	0
$Z_{\text{NA}}(\tau_0)$	1272	14	1258	14	0	0	1258	14	1258	0	0
$Z_{\text{LR}}(\tau_0, 5)$	10000	10000	0	2941	0	7059	0	10000	0	0	0
$Z_{\text{OLS}}(5)$	10000	10000	0	2941	0	7059	0	10000	0	0	0
$Z_{\text{Quad}}(5)$	10000	10000	10000	2941	2941	7059	7059	10000	10000	0	0

F.3 Additional Results for Section 7.4

F.3.1 Simulation Setup for Mixtures of Weibull Distributions

In this section, we simulated our survival curves based on mixtures of Weibull distributions described as follows. Denote $M \sim \mathcal{B}\text{ernoulli}(\pi)$ to be the random variable that characterizes the mixture of Weibull survival distribution after being randomized treatment to either treatment A or B. After being randomized to treatment $k = \{A, B\}$, the survival time for a patient has some probability π of coming from the Weibull distribution corresponding to $M = 1$ with shape parameter, a^1, and "rate" parameter, λ_k^1 and probability $1 - \pi$ coming from another Weibull distribution corresponding to $M = 0$ with shape parameter, a^0, and "rate" parameter λ_k^0 as below. Thus, our survival time for the i^{th} subject randomized to the k treatment group can be simulated from the mixtures of Weibull distribution described as follows:

$$f_{ik}(t) = \begin{cases} a_k^1 \lambda_w^1 (\lambda_w^1 t)^{a^1 - 1} \exp(-(\lambda_k^1 t)^{a^1}), & M = 1 \text{ with probability } \pi \\ a_k^0 \lambda_w^0 (\lambda_w^0 t)^{a^0 - 1} \exp(-(\lambda_k^0 t)^{a^0}), & M = 0 \text{ with probability } 1 - \pi \end{cases}$$

The expected survival at time t, $S_A(t)$, for *any* patient in the control group is $\pi \exp(-(\lambda_A^1 t)^{a_A^1}) + (1 - \pi) \exp(-(\lambda_A^0 t)^{a_A^0})$. Similarly, the expected survival at time t, $S_B(t)$, for *any* patient in the treatment group is $\pi \exp(-(\lambda_B^1 t)^{a_B^1}) + (1 - \pi) \exp(-(\lambda_B^0 t)^{a_B^0})$. This general formulation can allow us to simulate additional, more flexible survival functionals. We describe additional results corresponding to scenario 4 - 7 under stochastically ordered, crossing hazards survival curves, and scenario 4 and 5 for crossing survival, crossing hazards survival curves that are simulated from the mixture of Weibulls setup. Results for crossing survival, crossing hazards survival curves for scenario 1 and 3 are also provided that are simulated from mixtures of Exponentials described in section 7.3.1.

F.3.2 Stochastically Ordered, Crossing Hazards Survival Curves

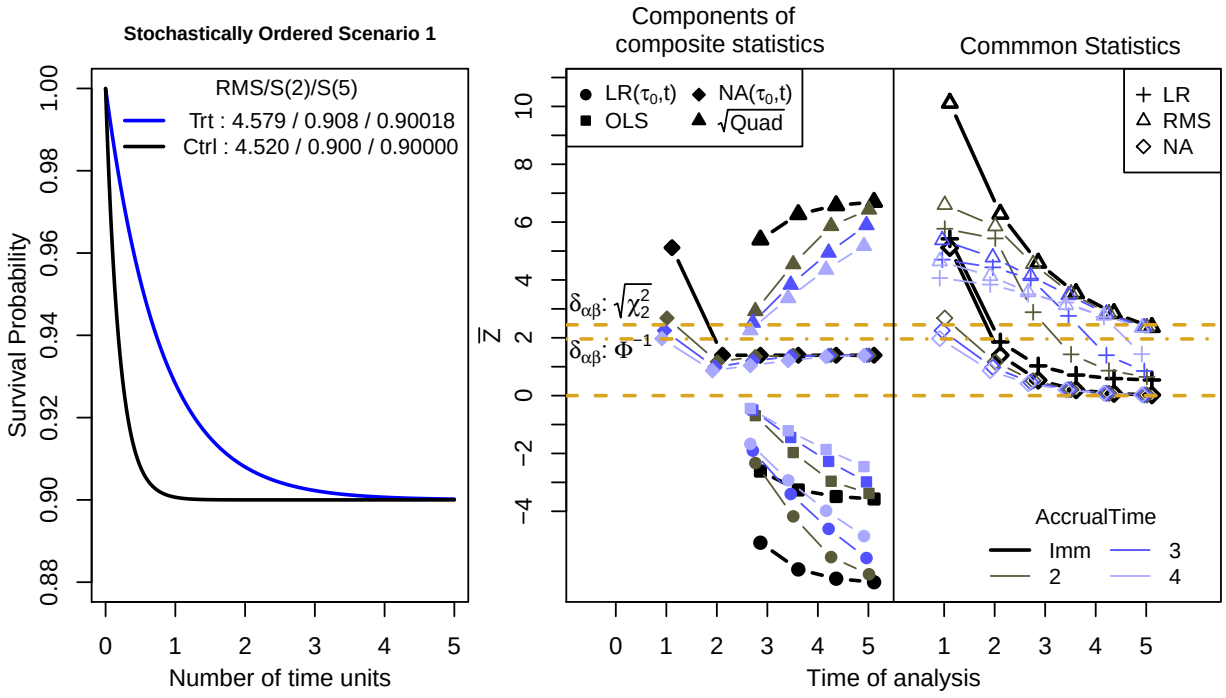

Figure F.6: Survival curves and plot of standardized alternatives for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 1) under various accrual patterns and different interim analyses. The commonly used test statistics have high probability of declaring a treatment as superior for overall LR and RMS. The combination of alternatives for $\text{NA}(\tau_0, t)$ and $\text{LR}(\tau_0, t)$ resides in quadrant IV and describes conflicting conclusions prior to crossing and after crossing. This observation is as speculated and described in Table 7.1. The net effect for the linear composite statistics concludes that the placebo arm is better than the treatment arm even though the treatment arm is better than placebo across all analyses time from 0 to 5.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
$\text{LR}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
$\text{NA}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.6: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 1) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	923.74	1036.82	1063.19	1073.45	1077.41	1078.96	
	2.00	333.33	829.87	1002.51	1050.39	1068.49	1075.50	
	3.00	222.25	553.13	816.25	1003.08	1051.23	1068.85	
	4.00	166.79	414.85	612.13	812.65	995.02	1049.83	
Events (τ_0, t)	Imm			26.37	36.63	40.59	42.14	
	2.00			5.74	17.85	31.67	38.68	
	3.00			3.82	11.94	21.65	32.03	
	4.00			2.86	8.92	16.24	24.02	
Z_{LR}	Imm	85.61	96.09	98.54	99.49	99.86	100.00	269.65
	2.00	30.99	77.16	93.21	97.66	99.35	100.00	268.79
	3.00	20.79	51.75	76.36	93.84	98.35	100.00	267.13
	4.00	15.88	39.51	58.30	77.41	94.78	100.00	262.37
$Z_{\text{LR}}(\tau_0, t)$	Imm			62.57	86.92	96.31	100.00	10.54
	2.00			14.83	46.13	81.88	100.00	9.67
	3.00			11.93	37.26	67.60	100.00	8.01
	4.00			11.90	37.14	67.62	100.00	6.00
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	25446.36
	2.00		62.45	93.58	99.47	100.00	100.00	25446.36
	3.00		42.41	73.54	93.77	99.10	100.00	25446.36
	4.00		32.65	55.95	76.05	94.07	100.00	25103.18

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

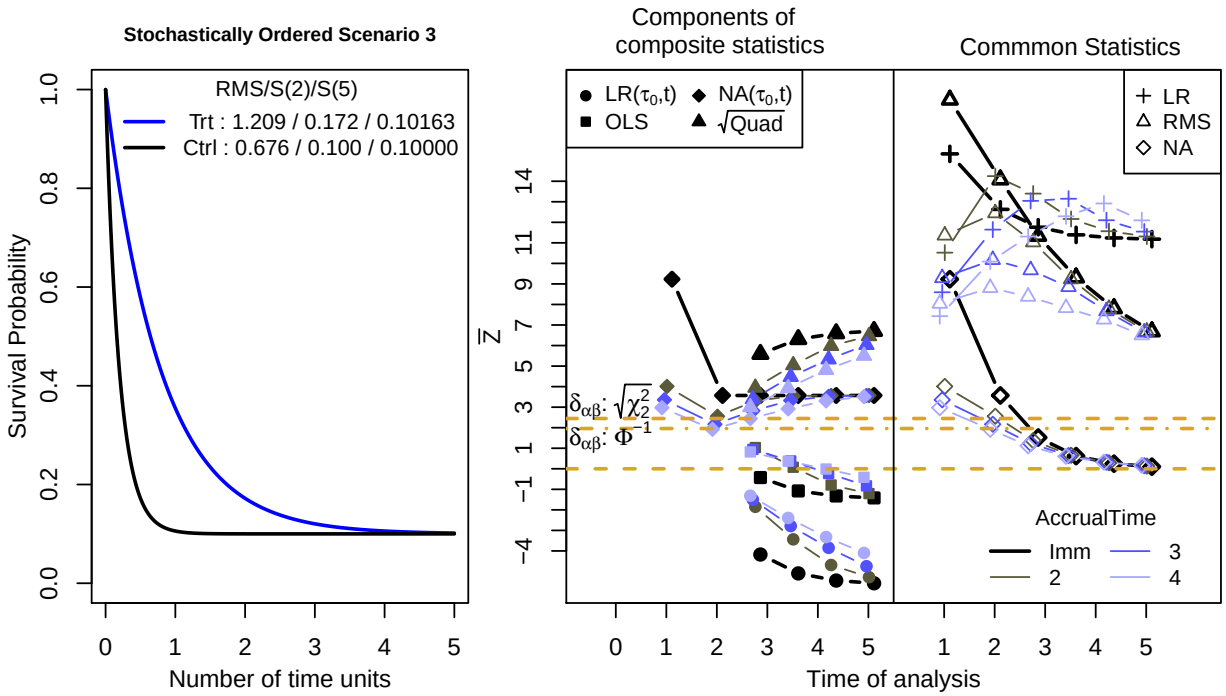

Figure F.7: Survival curves and plot of standardized alternatives for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 3) under various accrual patterns and different interim analyses. The combination of alternatives for $NA(\tau_0, t)$ and $LR(\tau_0, t)$ resides in quadrant IV and describe conflicting conclusions prior to crossing and after crossing. This observation is as speculated and described in Table 7.1. The net effect for the linear composite statistics conclude placebo is better than the treatment despite treatment being better at all times from 0 to 5 relative to placebo.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.7: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 3) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	923.64	1036.78	1063.31	1073.63	1077.64	1079.18	
	2.00	333.45	829.71	1002.54	1050.44	1068.64	1075.71	
	3.00	222.40	553.33	816.18	1003.12	1051.25	1068.99	
	4.00	166.81	415.00	612.16	812.63	995.11	1049.87	
Events (τ_0, t)	Imm			26.52	36.85	40.86	42.40	
	2.00			5.77	17.96	31.86	38.92	
	3.00			3.83	11.97	21.75	32.20	
	4.00			2.86	8.96	16.35	24.15	
Z_{LR}	Imm	85.92	95.62	98.32	99.41	99.84	100.00	233.30
	2.00	32.48	78.46	93.01	97.38	99.25	100.00	232.43
	3.00	21.79	52.64	77.20	93.98	98.21	100.00	230.78
	4.00	16.63	40.18	58.95	78.09	95.29	100.00	226.47
$Z_{\text{LR}}(\tau_0, t)$	Imm			61.70	86.55	96.27	100.00	10.22
	2.00			14.09	45.27	81.32	100.00	9.35
	3.00			11.13	36.33	67.02	100.00	7.70
	4.00			10.91	36.07	67.07	100.00	5.75
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	43.53
	2.00		45.36	85.19	98.29	100.00	100.00	43.53
	3.00		31.72	61.53	86.49	97.43	100.00	43.53
	4.00		25.57	47.89	68.84	88.62	100.00	42.02

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

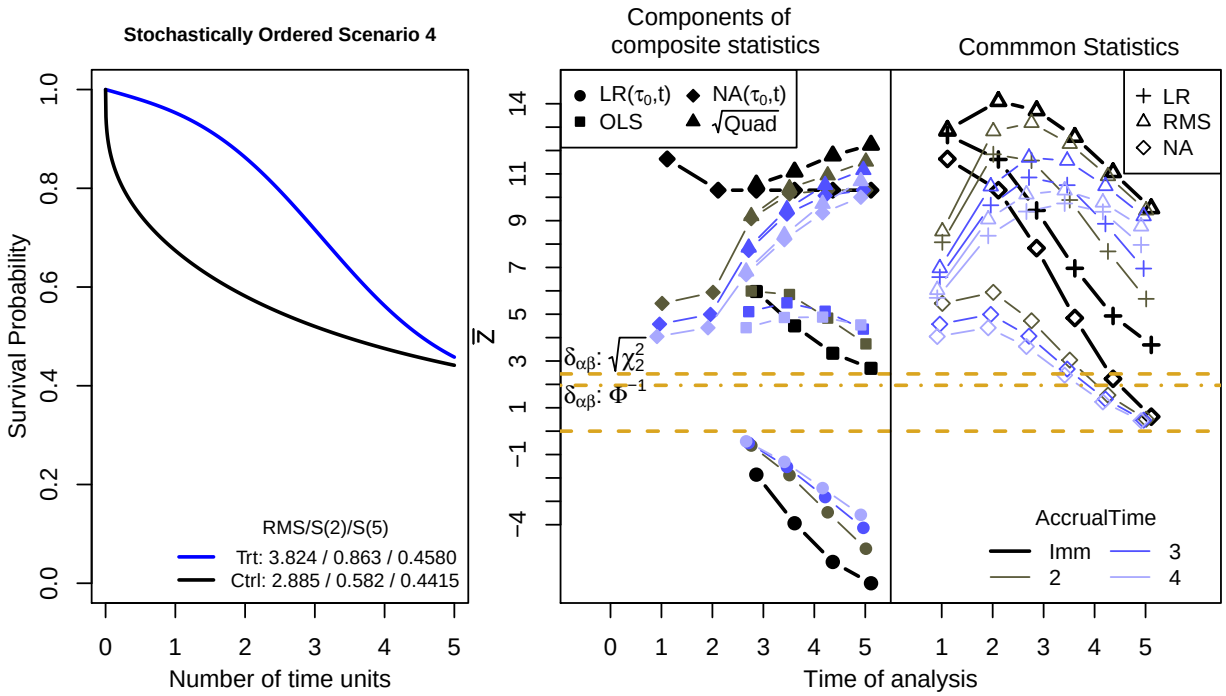

Figure F.8: Survival curves and plot of standardized alternatives for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 4) under various accrual patterns and different interim analyses. The combination of alternatives for the composite statistics resides in quadrant IV. However, we created a sufficiently large difference at the time of crossing such that the net effect of the linear composite statistics conclude the treatment is better than the placebo. The commonly used statistics are consistent in concluding some treatment effect relative to placebo.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.8: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 4) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	224.16	333.69	426.06	520.88	602.22	660.26	
	2.00	78.89	217.95	307.43	396.24	487.52	571.13	
	3.00	52.60	145.26	240.04	338.56	426.99	512.40	
	4.00	39.49	108.94	180.00	268.89	368.68	453.79	
Events (τ_0, t)	Imm			92.37	187.18	268.53	326.57	
	2.00			16.98	69.63	153.83	237.43	
	3.00			11.34	46.37	103.79	178.71	
	4.00			8.48	34.74	77.83	133.98	
Z_{LR}	Imm	34.24	50.55	64.32	78.62	91.06	100.00	161.83
	2.00	14.00	38.50	54.03	69.38	85.28	100.00	139.66
	3.00	10.40	28.59	47.10	66.22	83.35	100.00	125.27
	4.00	8.80	24.18	39.83	59.38	81.32	100.00	111.02
$Z_{\text{LR}}(\tau_0, t)$	Imm			27.84	56.77	81.92	100.00	80.02
	2.00			7.03	29.02	64.46	100.00	57.85
	3.00			6.23	25.68	57.82	100.00	43.46
	4.00			6.20	25.64	57.81	100.00	32.55
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	685.12
	2.00		35.92	78.02	96.98	100.00	100.00	685.12
	3.00		25.81	56.63	81.39	95.60	100.00	685.12
	4.00		21.71	45.18	67.14	87.13	100.00	646.04

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

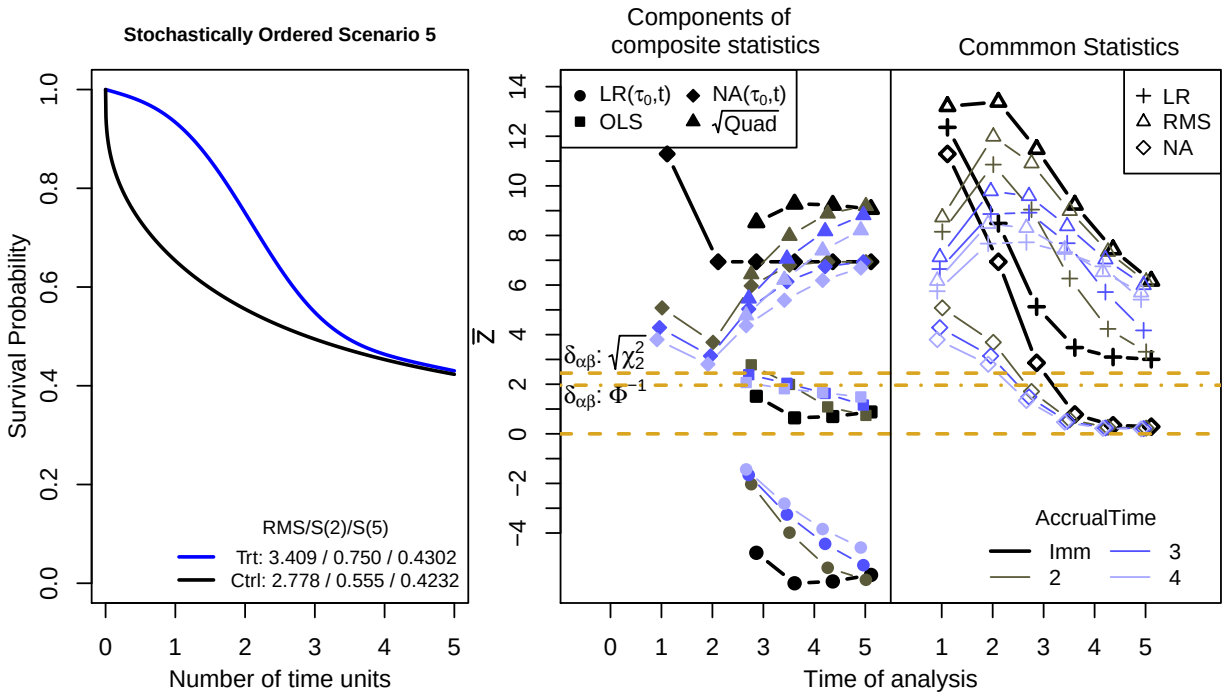

Figure F.9: Survival curves and plot of standardized alternatives for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 5) under various accrual patterns and different interim analyses. In this scenario, the combination of alternatives for the composite statistics are still in quadrant IV. However, we created a sufficiently large difference at time of crossing such that the net effect of the linear composite statistics conclude the treatment is better than the placebo. The commonly used statistics are consistent in concluding some treatment effect relative to placebo.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.9: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 5) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	247.94	416.97	542.13	620.65	660.73	687.83	
	2.00	83.90	248.47	374.10	493.42	586.53	643.80	
	3.00	55.95	165.56	286.16	411.89	515.19	596.31	
	4.00	42.00	124.15	214.56	324.54	439.01	529.52	
Events (τ_0, t)	Imm			125.16	203.68	243.76	270.86	
	2.00			24.47	87.80	169.55	226.82	
	3.00			16.29	58.50	114.95	179.33	
	4.00			12.20	43.88	86.25	134.49	
Z_{LR}	Imm	36.13	60.33	78.54	90.09	96.00	100.00	169.33
	2.00	13.12	38.64	57.97	76.43	90.98	100.00	158.31
	3.00	9.45	27.80	47.97	69.01	86.31	100.00	146.49
	4.00	7.98	23.47	40.49	61.26	82.90	100.00	130.04
$Z_{\text{LR}}(\tau_0, t)$	Imm			45.91	75.03	89.93	100.00	67.18
	2.00			10.64	38.49	74.58	100.00	56.16
	3.00			8.94	32.43	63.97	100.00	44.34
	4.00			8.90	32.40	63.98	100.00	33.22
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	530.31
	2.00		29.20	73.96	96.13	100.00	100.00	530.31
	3.00		21.15	52.66	78.23	94.46	100.00	530.31
	4.00		18.11	42.59	64.90	85.47	100.00	492.97

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

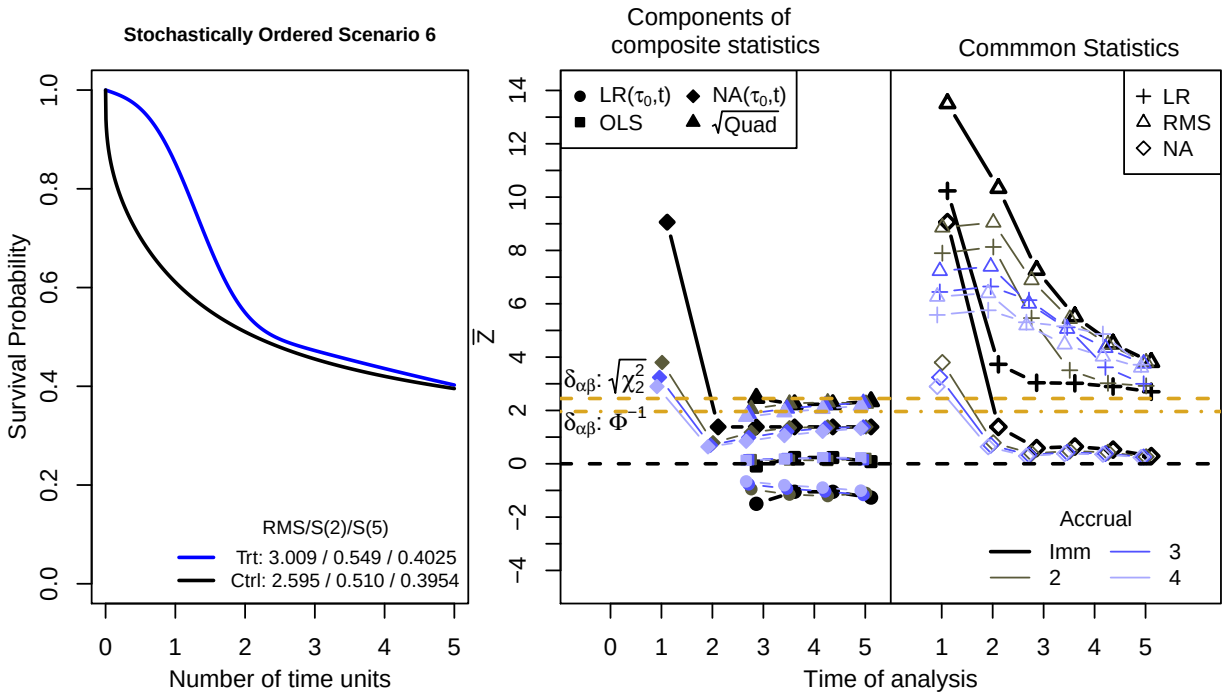

Figure F.10: Survival curves and plot of standardized alternatives for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 6) under various accrual patterns and different interim analyses. We have an advantage in survival prior to time τ_0 but this difference wears off such that there is negligible difference by time of crossing with negligible long term difference in survival past time τ_0.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.10: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 6) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	322.53	564.38	630.51	665.98	695.46	721.14	
	2.00	99.84	327.25	489.76	598.93	652.40	684.85	
	3.00	66.61	218.03	368.95	507.97	605.12	660.45	
	4.00	49.99	163.46	276.62	398.40	519.54	609.01	
Events (τ_0, t)	Imm			66.13	101.60	131.08	156.75	
	2.00			14.47	46.24	88.02	120.47	
	3.00			9.66	30.84	59.98	96.06	
	4.00			7.23	23.09	44.98	72.01	
Z_{LR}	Imm	44.54	78.02	87.29	92.27	96.40	100.00	178.04
	2.00	14.58	47.64	71.29	87.30	95.20	100.00	168.96
	3.00	10.08	32.90	55.74	76.77	91.53	100.00	162.86
	4.00	8.20	26.75	45.33	65.35	85.27	100.00	150.05
$Z_{\text{LR}}(\tau_0, t)$	Imm			42.19	64.82	83.62	100.00	39.13
	2.00			11.94	38.33	73.04	100.00	30.06
	3.00			9.97	32.01	62.38	100.00	23.95
	4.00			9.92	31.95	62.38	100.00	17.94
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	337.74
	2.00		27.38	72.15	96.10	100.00	100.00	337.74
	3.00		19.61	50.47	76.33	94.14	100.00	337.74
	4.00		16.78	41.11	63.26	84.03	100.00	311.81

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

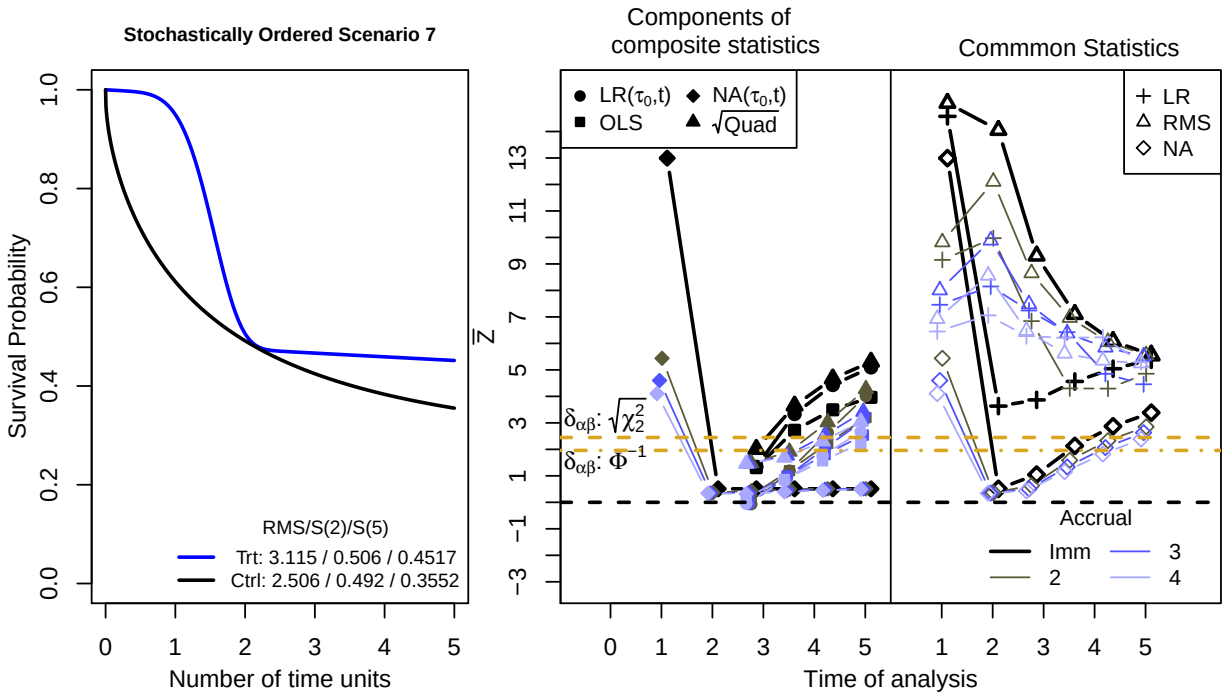

Figure F.11: Survival curves and plot of standardized alternatives for various test statistics when survival curves are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 7) under various accrual patterns and different interim analyses. We now created the scenario where there is negligible difference by time 2 but a difference exists at time 5 which still corresponds to a treatment benefit. Commonly used statistics works sufficiently well to identify the perferred treatment without the complexity of composite statistics providing conflicting conclusions before and after time of crossing.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.11: Average information growth for survival curves that are stochastically ordered without true crossings over the first five years (Stochastically Ordered Scenario 7) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	263.51	601.71	656.00	681.46	700.84	716.22	
	2.00	82.58	298.58	484.56	621.28	671.60	693.53	
	3.00	55.03	199.12	358.27	506.57	618.71	676.51	
	4.00	41.25	149.36	268.72	394.15	518.93	615.40	
Events (τ_0, t)	Imm			54.30	79.75	99.13	114.52	
	2.00			13.06	38.40	69.90	91.82	
	3.00			8.71	25.62	48.07	74.80	
	4.00			6.52	19.21	36.08	56.14	
Z_{LR}	Imm	36.81	83.77	91.49	95.10	97.84	100.00	175.60
	2.00	12.00	42.95	69.65	89.41	96.79	100.00	169.98
	3.00	8.19	29.34	52.86	74.76	91.34	100.00	165.73
	4.00	6.75	24.19	43.60	64.02	84.33	100.00	150.62
$Z_{\text{LR}}(\tau_0, t)$	Imm			47.58	69.82	86.68	100.00	28.50
	2.00			14.18	41.82	76.18	100.00	22.88
	3.00			11.57	34.20	64.24	100.00	18.64
	4.00			11.50	34.14	64.23	100.00	13.97
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	298.92
	2.00		21.02	65.96	94.37	100.00	100.00	298.92
	3.00		15.10	45.44	71.64	91.85	100.00	298.92
	4.00		13.45	38.00	60.62	82.09	100.00	268.54

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

In the results for scenario 1 and 3, if the long term treatment effect is the objective of the study, i.e., we are primarily interested in the benefit of the treatment at some pre-defined calendar time, then choosing the Nelson-Aalen test that evaluates the difference in survival probability at year 5 would have directly address the question of interest. If the relative benefit of this long term effect has to be weighted by the degree of early differences, then picking the log rank test, or the restricted mean statistic may in fact be more appropriate

over Logan's composite statistics to address the scientific question of which is the preferred treatment. The linear composite statistics, in this case, is weighted so heavily by the truncated log-rank test that the aggregate direction based on the equal weighting scheme will favor the placebo as the preferred treatment rather than the experimental treatment arm.

In scenario 4, 5, and 6 (Figure F.8, F.9, and F.10 respectively), we constructed survival curves with negligible difference in survival probability by time 5, and $S(5)$ to be approximately between 40% to 50%. However, we varied the timing for which the survival probability is similar to the placebo arm with sufficiently large early survival differences. For example, in Figure F.10 (Scenario 6), the survival curves are constructed such that experimental treatment survival curve is stochastically better than the placebo arm between 0 and 5. A significant difference is observed early on by time 1. However, this difference in probability of survival disappears by time 2.

In Figure F.11, scenario 7, we constructed a difference in survival again by year 5. In particular, this is the only stochastic ordered, non proportional scenario we investigated such that the alternatives of the truncated log rank test is estimated in the positive (and correct) direction. The commonly used statistics have relatively similar estimates when early differences exists, with a positive upswing in the estimates of the alternatives for the overall logrank statistic, and the Nelson-Aalen test statistics at calendar time of analyses. This leads to selecting the experimental treatment arm as the preferred treatment over the placebo arm by time 5. The composite statistics correctly pick up this positive upswing for scenario 7. This is however weighted downwards by the alternatives by the Nelson-Aalen test at time of crossing.

F.3.3 Crossing Hazards, Crossing Survival Curves

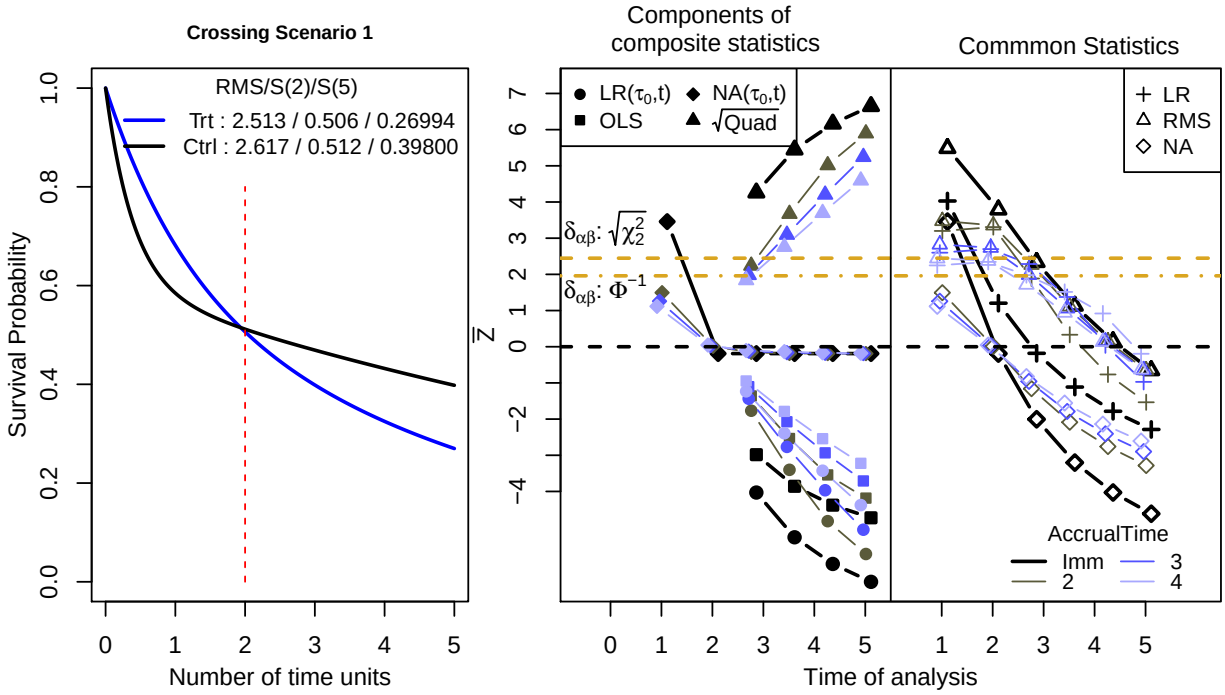

Figure F.12: Survival curves and plot of standardized alternatives for various test statistics when we have crossing survival curves (Crossing scenario 1) under various accrual patterns and different interim analyses. Simulated crossing hazards, crossing survival curves where we added variations to whether the curves crosses just before 2, and after 2. The combination of composite alternatives changes from Quadrant III (-,-), Quadrant III/IV(0, -), Quadrant IV (+,-) with the net result providing conclusion of preferring the placebo over treatment. The commonly used statistics is seen to have a changing alternative that switches from positive to negative.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.12: Average information growth for crossing survival curves (Crossing Scenario 1) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (τ_0, t)	Imm			70.36	125.51	171.11	210.06	
	2.00			13.84	50.92	105.10	154.48	
	3.00			9.24	33.96	71.16	118.94	
	4.00			6.91	25.47	53.41	89.13	
Events (t)	Imm	440.04	589.16	659.52	714.67	760.27	799.22	
	2.00	136.20	397.87	547.54	632.64	694.26	743.64	
	3.00	90.76	265.40	421.77	565.44	648.57	708.10	
	4.00	68.05	198.92	316.41	445.27	577.68	661.71	
Z_{LR}	Imm	55.17	73.89	82.73	89.62	95.25	100.00	198.51
	2.00	18.31	53.53	73.68	85.15	93.42	100.00	184.95
	3.00	12.79	37.46	59.57	79.88	91.62	100.00	176.15
	4.00	10.24	30.02	47.80	67.30	87.32	100.00	164.60
$Z_{\text{LR}}(\tau_0, t)$	Imm			33.86	60.23	81.80	100.00	51.82
	2.00			8.96	33.08	68.22	100.00	38.26
	3.00			7.74	28.59	59.90	100.00	29.46
	4.00			7.70	28.58	59.97	100.00	22.06
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	311.81
	2.00		35.38	78.90	97.31	100.00	100.00	311.81
	3.00		25.12	56.21	81.66	95.98	100.00	311.81
	4.00		20.91	44.62	66.28	86.52	100.00	295.13

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

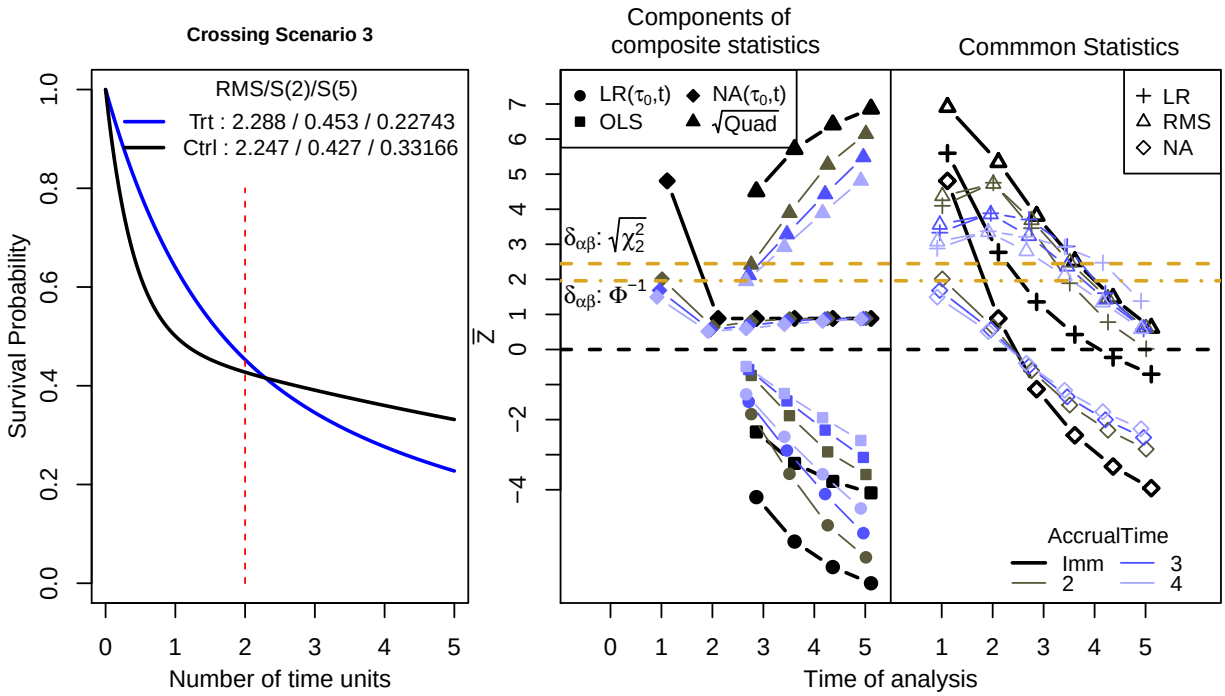

Figure F.13: Survival curves and plot of standardized alternatives for various test statistics when we have crossing survival curves (Crossing scenario 3) under various accrual patterns and different interim analyses. Simulated crossing hazards, crossing survival curves where we added variations to whether the curves crosses just before 2, and after 2. The combination of composite alternatives changes from Quadrant III (-,-), Quadrant III/IV(0, -), Quadrant IV (+,-) with the net result providing conclusion of preferring the placebo over treatment. The commonly used statistics is seen to have a changing alternative that switches from positive to negative.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.13: Average information growth for crossing survival curves (Crossing Scenario 3) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	515.48	671.71	739.57	790.20	830.72	864.48	
	2.00	161.51	463.47	627.54	713.04	771.15	815.76	
	3.00	107.61	309.12	485.90	643.50	727.15	783.17	
	4.00	80.74	231.74	364.49	508.02	652.82	738.25	
Events (τ_0, t)	Imm			67.86	118.49	159.02	192.78	
	2.00			13.47	48.82	99.45	144.05	
	3.00			8.99	32.58	67.40	111.46	
	4.00			6.73	24.43	50.55	83.58	
Z_{LR}	Imm	59.66	77.74	85.65	91.52	96.17	100.00	214.10
	2.00	19.80	56.80	76.88	87.40	94.55	100.00	202.18
	3.00	13.72	39.43	62.02	82.14	92.84	100.00	194.09
	4.00	10.90	31.34	49.33	68.80	88.44	100.00	182.90
$Z_{\text{LR}}(\tau_0, t)$	Imm			35.54	61.91	82.82	100.00	47.67
	2.00			9.33	33.96	69.19	100.00	35.75
	3.00			8.01	29.22	60.51	100.00	27.66
	4.00			7.96	29.18	60.49	100.00	20.71
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	236.15
	2.00		35.72	79.12	97.37	100.00	100.00	236.15
	3.00		25.30	56.30	81.77	96.06	100.00	236.15
	4.00		20.98	44.64	66.24	86.48	100.00	223.72

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

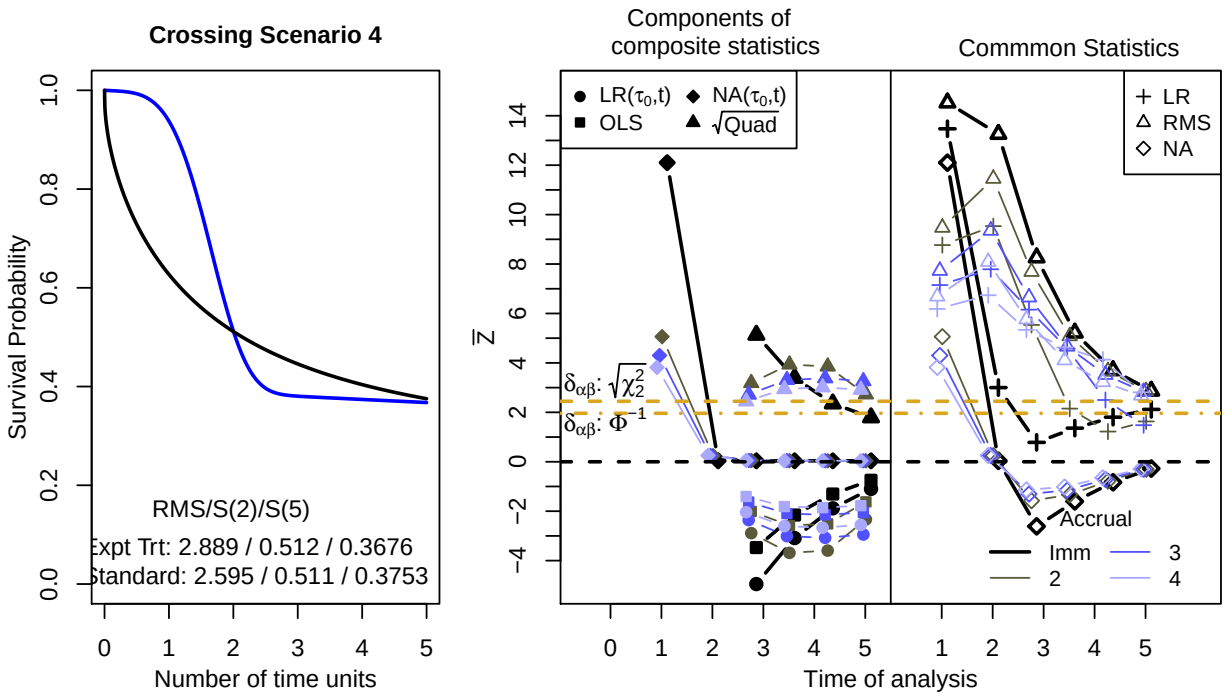

Figure F.14: Survival curves and plot of standardized alternatives for various test statistics when we have crossing survival curves (Crossing scenario 4) under various accrual patterns and different interim analyses. The commonly used statistics capture the large magnitude in survival difference earlier on during the trial.The difference in survival is exaggerated earlier on but treatment benefit wears off.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.14: Average information growth for crossing survival curves (Crossing Scenario 4) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	261.39	586.05	694.52	720.69	739.70	754.73	
	2.00	80.71	288.56	483.18	637.02	708.47	732.56	
	3.00	53.86	192.42	356.36	514.98	637.95	710.74	
	4.00	40.41	144.32	267.28	400.05	532.23	636.87	
Events (τ_0, t)	Imm			108.47	134.64	153.65	168.69	
	2.00			26.17	72.21	122.43	146.51	
	3.00			17.49	48.10	84.29	124.69	
	4.00			13.12	36.04	63.23	93.45	
Z_{LR}	Imm	34.58	77.40	91.93	95.43	97.98	100.00	185.71
	2.00	11.07	39.26	65.74	86.80	96.67	100.00	180.17
	3.00	7.61	26.97	50.02	72.32	89.65	100.00	174.74
	4.00	6.37	22.57	41.87	62.76	83.55	100.00	156.42
$Z_{\text{LR}}(\tau_0, t)$	Imm			64.30	79.78	91.06	100.00	41.98
	2.00			17.80	49.24	83.55	100.00	36.43
	3.00			13.93	38.49	67.54	100.00	31.00
	4.00			13.92	38.45	67.59	100.00	23.21
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	315.07
	2.00		19.57	64.94	93.68	100.00	100.00	315.07
	3.00		14.29	44.77	71.10	91.34	100.00	315.07
	4.00		12.85	37.62	60.43	82.05	100.00	281.88

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

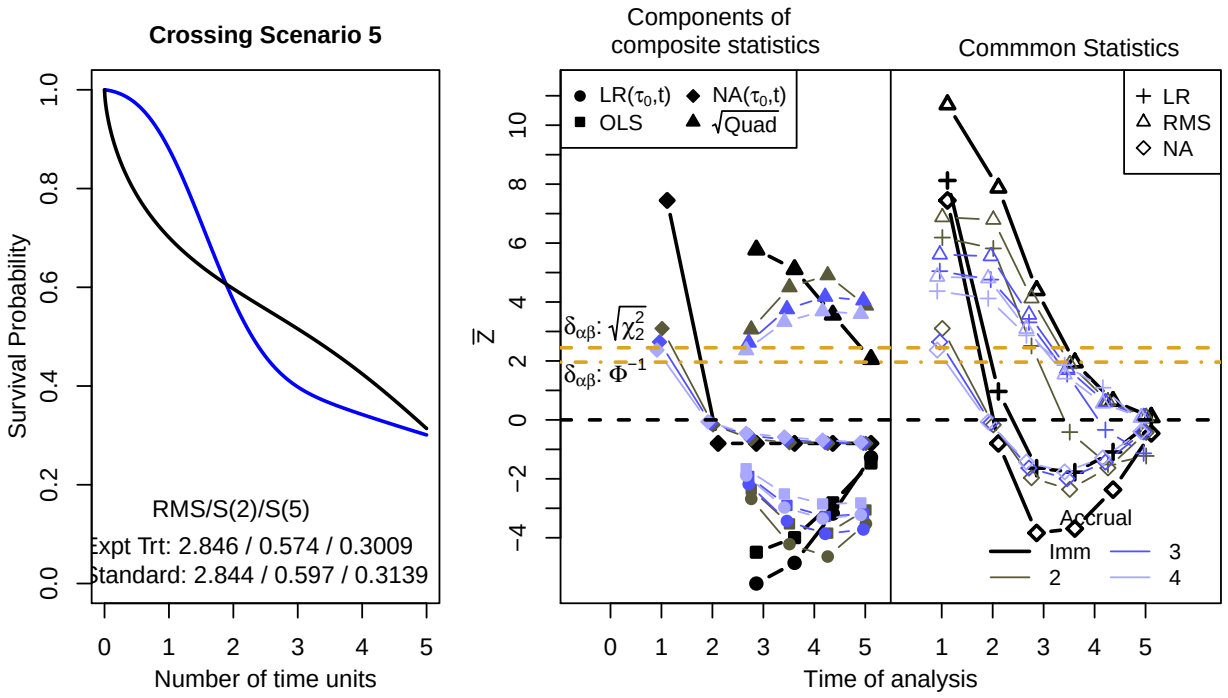

Figure F.15: Survival curves and plot of standardized alternatives for various test statistics when we have crossing survival curves (Crossing scenario 5) under various accrual patterns and different interim analyses. Differential preference in survival is again seen in this scenario where the treatment effect essentially wears off by time 5 to reflect no meaningful difference. On average, the hazard ratio based on the log rank test statistic is 1. However, the crossing survival curves at time 2 lead to averaging out this treatment effect via the use of the overall logrank statistic.
LR: Overall logrank statistic conducted at time t.
NA: Nelson-Aalen test statistic at time t.
RMS: Restricted mean statistics conducted at time $t - 0.25$.
LR(τ_0, t): Truncated logrank statistic up to time t starting at time τ_0.
NA(τ_0, t): Nelson-Aalen test conducted at time t restricted to time τ_0.
OLS: Linear composite statistics.
Quad: Quadratic test statistics.
$\sqrt{\chi^2_{2,\alpha}}$: line corresponding to the square root of the critical value based on the χ^2_2 at $\alpha = 0.05$.
$\Phi^{-1}(z_{1-\alpha/2})$: line corresponding to the critical value based on the standard normal.

Table F.15: Average information growth for crossing survival curves (Crossing Scenario 5) for the various test statistics under patterns of accrual and different interim analyses.

Statistic	Accrual	$t = 1$	$t = 2$	$t = 2.75$	$t = 3.5$	$t = 4.25$	$t = 5$	$\widehat{\mathcal{I}}_5$
Events (t)	Imm	251.15	497.78	624.25	696.98	761.81	831.04	
	2.00	71.26	258.93	428.32	570.86	669.26	740.80	
	3.00	47.52	172.57	314.52	465.44	591.47	688.91	
	4.00	35.63	129.39	235.87	360.14	493.18	610.51	
Events (τ_0, t)	Imm			126.46	199.20	264.03	333.25	
	2.00			26.11	87.95	171.47	243.01	
	3.00			17.42	58.64	116.46	191.13	
	4.00			13.06	44.00	87.26	143.36	
Z_{LR}	Imm	30.22	59.98	75.22	83.89	91.65	100.00	206.08
	2.00	9.63	34.95	57.85	77.13	90.40	100.00	183.65
	3.00	6.89	25.02	45.65	67.57	85.86	100.00	170.83
	4.00	5.82	21.15	38.61	58.99	80.79	100.00	151.34
$Z_{\text{LR}}(\tau_0, t)$	Imm			38.09	59.75	79.13	100.00	82.48
	2.00			10.76	36.26	70.63	100.00	60.05
	3.00			9.09	30.69	60.92	100.00	47.23
	4.00			9.07	30.67	60.85	100.00	35.39
$Z_{\text{NA}}(\tau_0, t)$	Imm		100.00	100.00	100.00	100.00	100.00	424.01
	2.00		23.16	68.19	94.81	100.00	100.00	424.01
	3.00		16.89	47.16	73.43	92.68	100.00	424.01
	4.00		14.97	39.06	61.52	82.78	100.00	385.15

Z_{LR}: Overall logrank statistic conducted at time t.
$Z_{\text{NA}}(\tau_0, t)$: Nelson-Aalen test conducted at time t restricted to time τ_0.
$Z_{\text{LR}}(\tau_0, t)$: Truncated logrank statistic up to time t starting at time τ_0.
Imm refers to the setting of immediate accrual.

F.4 Additional Results for Section 7.5

F.4.1 Summary Statistics for Other Scenarios

Table F.16: Summary statistics based on 10,000 simulations under the stochastically ordered, crossing hazards survivals scenario. Descriptives are presented in the format mean (standard deviation). (Calibrated to blue curves)

		Null					Alternative				
		$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Scenario A	No of events by t	156 (12)	458 (21)	801 (27)	1001 (30)	1058 (31)	222 (15)	553 (23)	905 (29)	1040 (31)	1069 (31)
	Events (B vs A)	78 vs 78	229 vs 229	401 vs 401	501 vs 500	529 vs 529	78 vs 144	229 vs 324	401 vs 504	501 vs 540	529 vs 540
	No of events in (2,t)	0 (0)	0 (0)	12 (4)	37 (6)	64 (8)	0 (0)	0 (0)	6 (3)	18 (4)	32 (6)
	Events (B vs A)	0 vs 0	0 vs 0	6 vs 6	18 vs 18	32 vs 32	0 vs 0	0 vs 0	6 vs 0	18 vs 0	32 vs 0
	$HR_{\text{Ref: B}}$	1.00 (0.161)	1.00 (0.094)	1.00 (0.072)	1.00 (0.064)	1.00 (0.063)	1.92 (0.141)	1.46 (0.087)	1.30 (0.068)	1.11 (0.063)	1.05 (0.062)
	$\text{RMS}_A(t)$‡	0.7235	1.646	2.552	3.454	4.354	0.6944	1.595	2.495	3.395	4.295
	$\text{RMS}_B(t)$‡	0.7235	1.645	2.552	3.453	4.354	0.7235	1.645	2.552	3.453	4.354
	$\widehat{S}_A(t)$	0.9284	0.9081	0.9023	0.9007	0.9002	0.9008	0.9001	0.9	0.9	0.9
	$\widehat{S}_B(t)$	0.9285	0.908	0.9022	0.9006	0.9002	0.9285	0.908	0.9022	0.9006	0.9002
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)$†	49.9	50.3	51.3	50.2	50.2	3.4	18.4	36.2	45.9	48.7
Scenario C	No of events by t	156 (12)	458 (17)	801 (16)	1001 (13)	1058 (11)	222 (13)	553 (17)	905 (14)	1040 (12)	1069 (11)
	Events (B vs A)	78 vs 78	229 vs 229	401 vs 401	501 vs 501	529 vs 529	78 vs 145	229 vs 324	401 vs 504	501 vs 540	529 vs 540
	No of events in (2,t)	0 (0)	0 (0)	13 (3)	37 (6)	64 (8)	0 (0)	0 (0)	6 (3)	18 (4)	32 (6)
	Events (B vs A)	0 vs 0	0 vs 0	6 vs 6	18 vs 18	32 vs 32	0 vs 0	0 vs 0	6 vs 0	18 vs 0	32 vs 0
	$HR_{\text{Ref: B}}$	1.00 (0.160)	1.00 (0.093)	1.00 (0.071)	1.00 (0.063)	1.00 (0.061)	3.24 (0.144)	2.70 (0.093)	2.44 (0.073)	2.16 (0.069)	2.03 (0.069)
	$\text{RMS}_A(t)^{\ddagger}$	0.5112	0.8099	0.9657	1.081	1.186	0.2491	0.3529	0.4528	0.5525	0.6523
	$\text{RMS}_B(t)^{\ddagger}$	0.5116	0.81	0.9665	1.082	1.187	0.5116	0.81	0.9665	1.082	1.187
	$\widehat{S}_A(t)$	0.3594	0.1743	0.121	0.1057	0.1015	0.1062	0.09966	0.09981	0.09977	0.09978
	$\widehat{S}_B(t)$	0.3584	0.1743	0.1212	0.1061	0.1018	0.3584	0.1743	0.1212	0.1061	0.1018
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)$†	50.1	50.2	49.6	48.9	49.3	1.5	6.1	21.9	36.6	45.3

†: Percentage of times a crossing is observed.

‡: The restricted mean statistic is truncated to 3 months just prior to the calendar time t.

Table F.17: Summary statistics based on 10,000 simulations under the crossing survival scenario. Descriptives are presented in the format mean (standard deviation). (Calibrated to blue curves except for F where a different parameterization was done to ensure the survival probability for both curves at $t = 5$ is 10%)

		Null					Alternative				
		$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Scenario D	No of events by t	102 (10)	358 (19)	717 (26)	1048 (31)	1283 (34)	159 (12)	458 (21)	817 (28)	1055 (31)	1181 (33)
	Events (B vs A)	51 vs 51	179 vs 179	359 vs 359	524 vs 524	642 vs 642	51 vs 108	179 vs 280	359 vs 459	524 vs 531	642 vs 539
	No of events in (2,t)	0 (0)	0 (0)	45 (7)	165 (13)	342 (18)	0 (0)	0 (0)	24 (5)	86 (9)	177 (13)
	Events (B vs A)	0 vs 0	0 vs 0	23 vs 23	82 vs 82	171 vs 171	0 vs 0	0 vs 0	23 vs 1	82 vs 3	171 vs 6
	$HR_{\text{Ref: B}}$	1.00 (0.200)	1.00 (0.107)	1.00 (0.076)	1.00 (0.063)	1.00 (0.057)	2.17 (0.171)	1.61 (0.097)	1.32 (0.072)	1.04 (0.063)	0.86 (0.060)
	$\text{RMS}_A(t)^{\ddagger}$	0.7335	1.672	2.578	3.462	4.33	0.7116	1.619	2.52	3.42	4.32
	$\text{RMS}_B(t)^{\ddagger}$	0.7335	1.672	2.578	3.462	4.329	0.7335	1.672	2.578	3.462	4.329
	$\widehat{S}_A(t)$	0.947	0.9129	0.8891	0.8711	0.8568	0.911	0.9012	0.9001	0.9	0.9
	$\widehat{S}_B(t)$	0.9471	0.9131	0.8888	0.8711	0.8568	0.9471	0.9131	0.8888	0.8711	0.8568
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)^{\dagger}$	49.2	49.5	50.5	50.6	50.6	1.9	14.3	87.7	100.0	100.0
Scenario F	No of events by t	138 (11)	421 (16)	755 (17)	969 (14)	1044 (12)	187 (12)	519 (17)	884 (15)	1073 (11)	1119 (9)
	Events (B vs A)	69 vs 69	211 vs 211	378 vs 377	485 vs 485	522 vs 522	79 vs 108	239 vs 280	425 vs 459	541 vs 531	580 vs 539
	No of events in (2,t)	0 (0)	0 (0)	17 (4)	52 (7)	93 (9)	0 (0)	0 (0)	10 (3)	30 (5)	53 (7)
	Events (B vs A)	0 vs 0	0 vs 0	8 vs 8	26 vs 26	46 vs 46	0 vs 0	0 vs 0	9 vs 1	26 vs 3	47 vs 6
	$HR_{\text{Ref: B}}$	1.00 (0.171)	1.00 (0.097)	1.00 (0.073)	1.00 (0.065)	1.00 (0.062)	1.64 (0.148)	1.50 (0.089)	1.39 (0.069)	1.28 (0.062)	1.19 (0.060)
	$\text{RMS}_A(t)^{\ddagger}$	0.5411	0.8918	1.077	1.205	1.311	0.404	0.5727	0.6793	0.7797	0.8793
	$\text{RMS}_B(t)^{\ddagger}$	0.5408	0.8912	1.077	1.205	1.312	0.5082	0.7725	0.8629	0.8933	0.9036
	$\widehat{S}_A(t)$	0.4159	0.2098	0.1366	0.1098	0.1002	0.2029	0.1112	0.1006	0.09953	0.09936
	$\widehat{S}_B(t)$	0.4151	0.2095	0.1365	0.1104	0.1003	0.3359	0.1166	0.04262	0.01648	0.006771
	% of $\widehat{S}_A(t) > \widehat{S}_B(t)^{\dagger}$	50.5	50.5	50.3	49.3	50.1	8.9	44.2	97.7	100.0	100.0

†: Percentage of times a crossing is observed.

‡: The restricted mean statistic is truncated to 3 months just prior to the calendar time t.

F.4.2 Example of Constrained Boundaries Approach

We present a worked example of the constrained boundaries based on the logrank test statistic to illustrate the slight differences in the two procedures described in section 7.5.1.2. We assume a four look, two sided level $\alpha = 0.05$, symmetric GSD using an OBF boundary that has 97.5% power to detect a standardized alternative of 7.929. In this GSD, the maximal statistical information is presumed to be 264.2263, with annual interim analysis starting from year 2, and ending at year 5. The sequence of average statistical information is 114.2607, 200.1134, 250.1231, and 264.2263, which corresponds to the sequence of information fraction Π_j to be 0.432, 0.757, 0.947, and 1.

Table F.18: The example illustrates the use of the constrained boundaries algorithm where we revised our monitoring boundary (highlighted in yellow) at each interim analysis based on the observed statistical information estimated from the logrank test statistic. Horizontally, the second and third row of the table correspond to the true observed statistical information at each interim analysis. The diagonal reflects the revised monitoring boundary for each consecutive interim analysis as we observed the estimated statistical information relative to the planned amount of statistical information. (b) This approach only adjusts the final critical value at the end of the trial while holding fixed the earlier boundaries.

					(a)				
	j			Orig.	1	2	3	4	(b)
		$\mathcal{I}_j$			121.94	208.46	261.97	275.22	
			Π_j		0.46	0.79	0.99	1.04	
Upp. boundary Z statistic	1	114.26	0.43	3.134	3.096	3.096	3.096	3.096	3.1335
	2	200.11	0.76	2.368	2.368	2.330	2.330	2.330	2.3678
	3	250.12	0.95	2.118	2.118	2.118	2.030	2.030	2.1179
	4	264.23	1.00	2.061	2.061	2.061	2.061	2.203	2.055
Error Spending Scale	1	114.26	0.43	0.035	0.039	0.039	0.039	0.039	0.035
	2	200.11	0.76	0.371	0.371	0.410	0.410	0.410	0.370
	3	250.12	0.95	0.797	0.797	0.799	0.958	0.958	0.799
	4	264.23	1.00	1.000	1.000	1.000	1.000	1.000	1.000

At the first interim analysis, our observed statistical information at year 2 is 121.94 instead of 114.2607, with $\widehat{\Pi}_1 = 0.46$ based on our pre-specified maximal statistical information. By constraining on the future boundaries, our revised boundaries on the Z scale at the first

analysis time is 3.0961. The proportion of error spent is 0.039 instead of 0.035, translating to having used up a cumulative $\alpha = 0.001$.

At the second interim analysis, the statistical information at year 3 is 208.46 with $\widehat{\Pi}_2 = 0.79$ using 264.2263 as our maximum statistical information. Holding fixed the revised boundaries from the past analyses, as well as the future boundaries, our revised oboundary at the second analysis is now 2.3303. The proportion of error spent is now 0.41α. In other words, we have spend a cumulative error of $0.001 + 0.009275 \approx 0.01025$.

At the third interim analysis, our observed statistical information is 261.97, with $\widehat{\Pi}_3 = 0.99$ when assuming 264.2263 as our maximum statistical information. Our revised boundary for the third analysis is 2.0304. The proportion of error spent is now 0.958 with a cumulative error of 0.02396 spent. At this analysis, it is possible to terminate the trial since more than 95% of our error is used up. If we chose to stop at the third interim analysis, this leads to a revised critical value of 2.0075 instead. Had we choose to continue and stop at the maximum planned calendar time of 5, then our observed statistical information is more than what was pre-specified (275.22). This leads to our revised, final, critical value to be 2.203, which is more extreme than our original final critical value.

We can apply the alternative strategy (b) by keeping our boundaries fixed (defined on the Z statistic scale) at the first three analyses, and only adjust the final boundary value to account for the observed sequence of information growth over the course of the trial. Design assumptions are revised at the end of the trial by using the observed sequence of statistical information with respect to the true maximum statistical information based on the final analysis (Column (b) in Table F.18). Such application will give rise to a slightly different final critical value relative to approach (a).

F.4.3 Prespecified Boundaries Based on Equally Spaced Information Growth

The monitoring boundaries as constructed based on either the Z or E scale (or other scales) can be used to reflect the degree of conservatism/anti-conservatism at interim analyses. We considered the Pocock monitoring boundaries with a total of three equally spaced analyses

Table F.19: Various monitoring boundaries presented based on the assumption of either the calendar time of analyses correspond to equally spaced information time or equally spent α on the error spending scale. All monitoring boundaries are calibrated with respect to OBF and constrained to maintain the same maximal statistical information.

			Equal Information Π						Unified family E			
			Three Analyses			Four Analyses			Three Analyses		Four Analyses	
	Time	Π_j	OBF	POC	HP	OBF	POC	HP	OBF	POC	OBF	POC
	Power		0.975	0.959	0.976	0.975	0.956	0.976	0.975	0.965	0.975	0.964
Upper Z	$t=2$	1/4	-	-	-	4.049	2.361	3.090	-	-	3.460	2.498
	$t=3$	1/3	3.471	2.289	3.090	-	-	-	3.193	2.394	-	-
		1/2	-	-	-	2.863	2.361	3.090	-	-	2.811	2.407
	$t=4$	2/3	2.454	2.289	3.090	-	-	-	2.493	2.294	-	-
		3/4	-	-	-	2.337	2.361	3.090	-	-	2.378	2.321
	$t=5$	1	2.004	2.289	1.970	2.024	2.361	1.976	2.001	2.200	2.020	2.245
Error Spent	$t=2$	1/4	-	-	-	0.001	0.364	0.040	-	-	0.011	0.250
	$t=3$	1/3	0.010	0.441	0.040	-	-	-	0.028	0.333	-	-
		1/2	-	-	-	0.084	0.631	0.073	-	-	0.105	0.500
	$t=4$	2/3	0.286	0.759	0.073	-	-	-	0.268	0.667	-	-
		3/4	-	-	-	0.418	0.835	0.101	-	-	0.393	0.750
	$t=5$	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

HP: Haybittle-Peto

for illustration. We first presume three equally spaced analysis on the calendar time (with equal statistical information). On the information based scale, our Z statistic corresponds to ± 2.289 at all calendar time (highlighted in yellow) in Table F.19. On the E scale (a 1-1 mapping from the Z scale to the cumulative amount of error spent), our cumulative proportion of error spent at each analysis translates to 44.1%, 75.9%, and 100%.

Clinical trialists may alternatively choose to plan the same study using the Lan-DeMets error spending function using the Pocock monitoring boundary. Thus, on the error spending scale, we presumably spend an equal amount of error at each calendar time, therefore our cumulative error is thus 33.3%, 66.7%, and 100% (highlighted in light blue) in Table F.19. However, conversion of the error spending scale to the Z scale results in slightly different monitoring rule (2.394, 2.294, and 2.200 at the first, second, and third analysis) as compared to the above monitoring boundaries presuming equally spaced statistical information defined on the Z scale based on the calendar time. This difference in monitoring rule translates to requiring a more extreme test statistic at the first interim analysis relative to the "fixed" Z

statistic, as would be obtained based on an equal, information-based Pocock design. Under the null hypothesis, and additional assumptions, either scale should ideally maintain control of the overall Type 1 error under the strong null setting.

The OBF monitoring boundary, as defined on the Z statistic scale, tends to be more conservative when assuming an the equally spaced, information scale as compared to defining it on the error spending scale (E) in `RCTdesign`. On the information scale, when the information fraction $\Pi < 0.5$, the OBF monitoring boundary (defined on the Z statistic scale) tends to be more extreme. Therefore, the (cumulative) error spent at earlier interim analysis is extremely small relative to the OBF boundary derived on the error spending scale. For an OBF design with a total of four equally spaced analyses, the cumulative proportion of error is 0.418 when we are 3/4 of the way through the trial (highlighted in pink in Table F.19). Compared to the boundaries defined on the unified family of E scale, the cumulative error is only 0.393, with a more extreme Z statistic relative to the OBF design planned directly using the information scale.

The Haybittle-Peto (HP) boundary mimics some form of Pocock design, by the use of an extremely conservative boundary even at later interim analysis. The extreme boundary at later interim analysis makes it harder to stop the trial even when accumulated data may have demonstrated convincing evidence of efficacy/futility. As such, the critical value at the final analysis for the HP design tends to be similar to conducting a fixed sample design with the same maximum statistical information.

F.4.4 Results for Recalibrated Boundaries Based on Naïve Information Growth

Under the strong null hypothesis, the revised final boundaries tend to be more extreme than the original critical boundaries if the overall Type 1 error without calibration is anti-conservative. Likewise, the revised boundaries tend to be less extreme than the original critical boundaries at the final analysis when the overall Type 1 error is conservative.

The test statistics, namely, Z_{LR}, $Z_{\text{NA}}(t)$, and Z_{RMS}, tend to be less extreme than the final boundaries of the original monitoring rule under most scenarios except for D. This

observation is consistent when the overall Type 1 error for the pre-specified rule for the corresponding test statistic is rejecting less often than the nominal α. In particular, $Z_{\text{NA}}(t)$ tends to reject more often for Scenario D under the Pocock monitoring rule designed using the unified family specified either on the statistical information or error spending scale (Table F.20 & F.21).

Thus, the recalibrated final boundaries for $Z_{\text{NA}}(t)$ must be more extreme than, for example, 2.2895, when presuming an equally spaced, information monitoring boundary. We note that while $Z_{\text{NA}}(t)$ does not have an independent increment structure when performing the timing analysis at different monitoring time, and changing the analysis, under the strong null hypothesis, we observe a higher probability of rejecting the null hypothesis. When using the naïve linear combination statistic (that does not have independent increments), and presuming an equally spaced information growth, the final boundary of the GSD requires mild correction to obtain a fixed level $\alpha = 0.05$.

For the quadratic test statistic, we perform the analysis by computing the p-value using the χ^2_2 distribution at the interim analyses, and then back transforming this computed p-value, using the inverse normal distribution $\Phi^{-1}(p)$, to the "Z"-statistic scale. We then compare this "Z-statistic" obtained with our monitoring boundaries to evaluate whether we have crossed the monitoring boundaries. This naïve approach tends to lead to an inflated nominal Type 1 error, which can be avoided by recalibrating our final critical value to ensure a fixed Type 1 error rate of α.

Table F.22 and F.23 show the recalibrated results under the weak null hypothesis for scenarios A-F based on naïve information growth.

Table F.20: Original and recalibrated boundaries to ensure a fixed overall Type 1 error for various test statistics under the different monitoring rules for survival curves that are stochastically ordered without true crossings over the first five years based on naïve information growth.

		Three Analyses					Four Analyses				
		OBF	OBF^E	POC	POC^E	HP	OBF	OBF^E	POC	POC^E	HP
Original	$\Pi = 1/4$						4.0486	3.4599	2.3613	2.4979	3.0902
	$\Pi = 1/3$	3.4711	3.1929	2.2895	2.3941	3.0902					
	$\Pi = 1/2$						2.8628	2.8113	2.3613	2.4073	3.0902
	$\Pi = 2/3$	2.4544	2.4935	2.2895	2.2937	3.0902					
	$\Pi = 3/4$						2.3375	2.3782	2.3613	2.3209	3.0902
	$\Pi = 1$	2.0040	2.0009	2.2895	2.2002	1.9704	2.0243	2.0196	2.3613	2.2451	1.9759
Scenario A	Z_{LR}	1.9939	1.9906	2.0937	2.0440	1.9921	1.9981	1.9976	2.2539	2.1486	1.9979
	Z_{NA}	1.9931	1.9908	2.2387	2.1296	1.9863	2.0197	2.0196	2.3447	2.2266	1.9945
	Z_{RMS}	1.9874	1.9874	2.0782	2.0463	1.9876	1.9913	1.9930	2.2119	2.1361	1.9935
	$Z_{\text{NA}}(2,t)$	2.0123	2.0123	2.1176	2.0615	2.0129	2.0170	2.0204	2.2118	2.1347	2.0187
	$Z_{\text{OLS Fixed}}$	2.0040	2.0092	2.3550	2.2414	1.9855					
	$Z^{\text{N}}_{\text{OLS}}$	2.0069	2.0069	2.1206	2.0689	2.0072					
	$Z^{\text{S}}_{\text{OLS}}$	2.0024	2.0009	2.2989	2.1940	1.9871					
	Z_{Quad}	2.0625	2.0625	2.4872	2.3633	2.0050					
Scenario B	Z_{LR}	1.9600	1.9605	2.0248	1.9982	1.9615	1.9680	1.9695	2.1468	2.0831	1.9660
	Z_{NA}	1.9607	1.9562	2.1271	2.0944	1.9414	1.9936	1.9817	2.2685	2.1824	1.9457
	Z_{RMS}	1.9666	1.9677	2.0428	2.0031	1.9691	1.9723	1.9727	2.1434	2.0869	1.9723
	$Z_{\text{NA}}(2,t)$	1.9593	1.9590	2.0868	2.0445	1.9593	1.9819	1.9620	2.1757	2.1290	1.9597
	$Z_{\text{OLS Fixed}}$	1.9811	1.9774	2.2028	2.1205	1.9560					
	$Z^{\text{N}}_{\text{OLS}}$	1.9555	1.9557	2.0760	2.0325	1.9558					
	$Z^{\text{S}}_{\text{OLS}}$	1.9698	1.9698	2.2066	2.1400	1.9609					
	Z_{Quad}	2.0340	2.0369	2.5609	2.3807	1.9694					
Scenario C	Z_{LR}	1.9736	1.9725	2.0546	2.0249	1.9725	1.9769	1.9817	2.1947	2.1192	1.9822
	Z_{NA}	1.9523	1.9511	2.0742	2.0303	1.9278	1.9791	1.9702	2.1297	2.0954	1.9305
	Z_{RMS}	1.9605	1.9586	2.1280	2.0757	1.9586	1.9873	1.9926	2.2703	2.1819	1.9778
	$Z_{\text{NA}}(2,t)$	1.9678	1.9655	2.0897	2.0536	1.9632	1.9854	1.9779	2.1366	2.1012	1.9692
	$Z_{\text{OLS Fixed}}$	2.0194	2.0105	2.2846	2.1913	1.9909					
	$Z^{\text{N}}_{\text{OLS}}$	2.0050	2.0031	2.1829	2.1122	1.9789					
	$Z^{\text{S}}_{\text{OLS}}$	2.0164	2.0120	2.2456	2.1788	1.9857					
	Z_{Quad}	2.0841	2.0774	2.6446	2.4002	2.0063					

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z^{\text{S}}_{\text{OLS}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.

Table F.21: Original and recalibrated boundaries to ensure a fixed overall Type 1 error for various test statistics under the different monitoring rules for crossing survival curves based on naïve information growth.

		Three Analyses					Four Analyses				
		OBF	OBF^E	POC	POC^E	HP	OBF	OBF^E	POC	POC^E	HP
Original	$\Pi = 1/4$						4.0486	3.4599	2.3613	2.4979	3.0902
	$\Pi = 1/3$	3.4711	3.1929	2.2895	2.3941	3.0902					
	$\Pi = 1/2$						2.8628	2.8113	2.3613	2.4073	3.0902
	$\Pi = 2/3$	2.4544	2.4935	2.2895	2.2937	3.0902					
	$\Pi = 3/4$						2.3375	2.3782	2.3613	2.3209	3.0902
	$\Pi = 1$	2.0040	2.0009	2.2895	2.2002	1.9704	2.0243	2.0196	2.3613	2.2451	1.9759
Scenario D	Z_{LR}	2.0109	2.0088	2.2581	2.1722	1.9931	2.0485	2.0426	2.5573	2.3487	2.0069
	Z_{NA}	2.0274	2.0228	2.3220	2.2603	1.9771	2.0822	2.0785	2.5967	2.3456	1.9890
	Z_{RMS}	1.9951	1.9922	2.1472	2.0987	1.9860	2.0084	2.0062	2.3300	2.2438	1.9954
	$Z_{\text{NA}}(2,t)$	1.9989	1.9989	2.1221	2.0698	1.9936	2.0137	2.0068	2.2328	2.1382	1.9989
	$Z_{\text{OLS Fixed}}$	2.0470	2.0451	2.5325	2.3393	2.0170					
	$Z^{\text{N}}_{\text{OLS}}$	1.9930	1.9915	2.1224	2.0682	1.9899					
	$Z^{\text{S}}_{\text{OLS}}$	2.0508	2.0438	2.4555	2.3155	2.0068					
	Z_{Quad}	2.0656	2.0621	2.5189	2.3707	2.0041					
Scenario E	Z_{LR}	1.9831	1.9794	2.1015	2.0654	1.9718	2.0023	1.9998	2.2973	2.1744	1.9863
	Z_{NA}	1.9477	1.9430	2.1301	2.0791	1.9069	1.9947	1.9776	2.2000	2.1362	1.9160
	Z_{RMS}	1.9710	1.9692	2.0932	2.0499	1.9611	1.9898	1.9926	2.3017	2.1710	1.9810
	$Z_{\text{NA}}(2,t)$	1.9561	1.9561	2.1272	2.0444	1.9548	1.9581	1.9578	2.1543	2.0909	1.9561
	$Z_{\text{OLS Fixed}}$	1.9848	1.9787	2.2615	2.1903	1.9591					
	$Z^{\text{N}}_{\text{OLS}}$	1.9688	1.9677	2.1529	2.0885	1.9631					
	$Z^{\text{S}}_{\text{OLS}}$	1.9790	1.9784	2.2295	2.1556	1.9566					
	Z_{Quad}	2.0465	2.0515	2.3748	2.2966	1.9773					
Scenario F	Z_{LR}	1.9959	1.9957	2.1021	2.0613	1.9957	2.0046	2.0046	2.2194	2.1379	2.0046
	Z_{NA}	1.8870	1.8827	2.0324	1.9877	1.8649	1.9141	1.9082	2.0229	2.0049	1.8657
	Z_{RMS}	2.0058	2.0041	2.1660	2.1016	1.9969	2.0243	2.0225	2.3163	2.2143	2.0027
	$Z_{\text{NA}}(2,t)$	1.9671	1.9645	2.1079	2.0527	1.9630	1.9913	1.9913	2.1324	2.0902	1.9631
	$Z_{\text{OLS Fixed}}$	2.0221	2.0232	2.2924	2.2105	2.0006					
	$Z^{\text{N}}_{\text{OLS}}$	2.0343	2.0289	2.1916	2.1495	2.0146					
	$Z^{\text{S}}_{\text{OLS}}$	2.0247	2.0221	2.2663	2.1723	1.9968					
	Z_{Quad}	2.0228	2.0189	2.6013	2.4163	1.9869					

$^{\text{E}}$ represents the error spending version of the boundary shape describing specific class of monitoring rules. In this case, the OBF monitoring boundary specified on the Z statistic scale gives a boundary shape function different from that specified on the error spending scale.
$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z^{\text{S}}_{\text{OLS}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.

Table F.22: Probability of rejecting the weak null hypothesis for survival curves that are stochastically ordered without true crossings over the first five years based on naïve, equally spaced information growth with boundaries calibrated to reject the strong null hypothesis at a fixed 5% error rate for the different test statistics and monitoring rules.

		Alt Hypothesis				
		OBF Overall/Std/Trt	OBF (Error) Overall/Std/Trt	Pocock Overall/Std/Trt	Pocock(Error) Overall/Std/Trt	Haybittle-Peto Overall/Std/Trt
Scenario A	Z_{LR}	68.17/0.26/67.91	77.35/0.26/77.09	94.79/0.20/94.59	93.70/0.24/93.46	80.29/0.26/80.03
	Z_{NA}	4.87/2.22/2.65	4.91/2.23/2.68	5.62/1.42/4.20	5.46/1.70/3.76	5.00/2.27/2.73
	Z_{RMS}	75.41/0.00/75.41	79.46/0.00/79.46	94.80/0.00/94.80	93.71/0.00/93.71	81.05/0.00/81.05
	$Z_{\text{NA}}(2,t)$	27.44/0.04/27.40	27.45/0.04/27.41	25.93/0.06/25.87	26.72/0.06/26.66	27.40/0.04/27.36
	$Z_{\text{OLS Fixed}}$	88.71/88.71/0.00	88.60/88.60/0.00	78.38/78.36/0.02	82.09/82.09/0.00	89.17/89.17/0.00
	$Z^{\text{N}}_{\text{OLS}}$	24.37/0.05/24.32	24.39/0.05/24.34	23.54/0.06/23.48	23.82/0.07/23.75	24.40/0.05/24.35
	$Z^{\text{S}}_{\text{OLS}}$	88.84/88.78/0.06	88.97/88.83/0.14	81.74/80.06/1.68	84.58/83.40/1.18	89.32/89.13/0.19
	Z_{Quad}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	87.35/0.25/87.10	91.44/0.24/91.20	98.60/0.05/98.55	98.16/0.12/98.04	93.12/0.23/92.89
	$Z_{\text{NA}}[J=4]$	5.11/2.14/2.97	5.74/2.14/3.60	12.60/1.08/11.52	11.19/1.36/9.83	6.70/2.21/4.49
	$Z_{\text{RMS}}[J=4]$	89.65/0.00/89.65	94.03/0.00/94.03	99.30/0.00/99.30	99.00/0.00/99.00	95.77/0.00/95.77
	$Z_{\text{NA}}(2,t)[J=4]$	27.36/0.04/27.32	27.40/0.04/27.36	26.26/0.05/26.21	26.71/0.05/26.66	27.55/0.04/27.51
Scenario B	Z_{LR}	99.97/0.00/99.97	99.99/0.00/99.99	100/0.00/100	100/0.00/100	100/0.00/100
	Z_{NA}	5.00/1.94/3.06	5.09/1.96/3.13	6.77/1.44/5.33	6.24/1.51/4.73	5.22/2.09/3.13
	Z_{RMS}	96.49/0.00/96.49	97.59/0.00/97.59	99.82/0.00/99.82	99.72/0.00/99.72	98.07/0.00/98.07
	$Z_{\text{NA}}(2,t)$	45.47/0.00/45.47	45.46/0.00/45.46	43.30/0.00/43.30	43.84/0.00/43.84	45.48/0.00/45.48
	$Z_{\text{OLS Fixed}}$	75.87/75.87/0.00	75.97/75.97/0.00	66.93/66.89/0.04	70.25/70.23/0.02	76.82/76.82/0.00
	$Z^{\text{N}}_{\text{OLS}}$	22.81/0.07/22.74	22.81/0.07/22.74	27.77/0.05/27.72	26.48/0.05/26.43	22.67/0.07/22.60
	$Z^{\text{S}}_{\text{OLS}}$	76.49/76.21/0.28	76.67/76.20/0.47	69.94/66.61/3.33	72.15/69.46/2.69	77.15/76.59/0.56
	Z_{Quad}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}[J=4]$	5.60/1.83/3.77	7.30/1.86/5.44	18.99/1.08/17.91	16.62/1.25/15.37	9.57/2.06/7.51
	$Z_{\text{RMS}}[J=4]$	99.54/0.00/99.54	99.84/0.00/99.84	100/0.00/100	100/0.00/100	99.96/0.00/99.96
	$Z_{\text{NA}}(2,t)[J=4]$	45.62/0.00/45.62	45.79/0.00/45.79	43.25/0.00/43.25	43.56/0.00/43.56	46.02/0.00/46.02
Scenario C	Z_{LR}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	Z_{NA}	7.00/1.51/5.49	7.63/1.51/6.12	15.64/1.06/14.58	14.38/1.22/13.16	7.63/1.68/5.95
	Z_{RMS}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)$	95.40/0.00/95.40	95.43/0.00/95.43	94.64/0.00/94.64	94.91/0.00/94.91	95.47/0.00/95.47
	$Z_{\text{OLS Fixed}}$	8.88/8.78/0.10	8.97/8.87/0.10	7.79/5.28/2.51	8.33/6.50/1.83	9.27/9.17/0.10
	$Z^{\text{N}}_{\text{OLS}}$	43.11/0.01/43.10	43.71/0.01/43.70	65.06/0.00/65.06	62.54/0.01/62.53	38.16/0.01/38.15
	$Z^{\text{S}}_{\text{OLS}}$	14.21/8.82/5.39	17.26/8.86/8.40	37.70/5.72/31.98	35.08/6.66/28.42	18.74/9.25/9.49
	Z_{Quad}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}[J=4]$	13.55/1.41/12.14	21.59/1.42/20.17	50.51/0.72/49.79	46.85/0.90/45.95	28.00/1.63/26.37
	$Z_{\text{RMS}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)[J=4]$	95.31/0.00/95.31	95.39/0.00/95.39	94.83/0.00/94.83	94.98/0.00/94.98	95.49/0.00/95.49

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z^{\text{S}}_{\text{OLS}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.

Table F.23: Probability of rejecting the weak null hypothesis for crossing survival curves based on naïve, equally spaced information growth with boundaries calibrated to reject the strong null hypothesis at a fixed 5% error rate for the different test statistics and monitoring rules.

		Alt Hypothesis				
		OBF Overall/Std/Trt	OBF (Error) Overall/Std/Trt	Pocock Overall/Std/Trt	Pocock(Error) Overall/Std/Trt	Haybittle-Peto Overall/Std/Trt
Scenario D	Z_{LR}	99.65/33.32/66.33	99.92/24.16/75.76	100/5.83/94.17	100/7.07/92.93	99.97/21.07/78.90
	Z_{NA}	99.89/99.89/0.00	99.89/99.89/0.00	99.80/99.75/0.05	99.82/99.79/0.03	99.87/99.87/0.00
	Z_{RMS}	80.38/0.93/79.45	87.30/0.93/86.37	97.98/0.50/97.48	97.51/0.60/96.91	89.29/0.92/88.37
	$Z_{\text{NA}}(2,t)$	52.59/0.00/52.59	52.59/0.00/52.59	51.02/0.00/51.02	51.73/0.00/51.73	52.82/0.00/52.82
	Z_{OLS}	100/100/0.00	100/100/0.00	100/100/0.00	100/100/0.00	100/100/0.00
	$Z^{\text{N}}_{\text{QLS}}$	34.61/0.02/34.59	34.58/0.02/34.56	38.54/0.01/38.53	37.86/0.01/37.85	34.30/0.02/34.28
	$Z^{\text{S}}_{\text{OLS}}$	100/99.98/0.02	100/99.92/0.08	100/98.95/1.05	100/99.17/0.83	100/99.88/0.12
	Z_{Quad}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	100/9.57/90.43	100/4.59/95.41	100/0.40/99.60	100/0.55/99.45	100/2.67/97.33
	$Z_{\text{NA}}[J=4]$	99.90/0.79/99.11	99.89/2.96/96.93	99.82/16.66/83.16	99.86/14.05/85.81	99.89/5.56/94.33
	$Z_{\text{RMS}}[J=4]$	96.84/0.61/96.23	98.88/0.32/98.56	99.93/0.03/99.90	99.90/0.03/99.87	99.40/0.21/99.19
	$Z_{\text{NA}}(2,t)[J=4]$	52.37/0.00/52.37	52.80/0.00/52.80	50.12/0.00/50.12	51.69/0.00/51.69	53.01/0.00/53.01
Scenario E	Z_{LR}	91.54/2.29/89.25	95.66/1.98/93.68	99.74/0.52/99.22	99.63/0.68/98.95	96.72/1.86/94.86
	Z_{NA}	99.88/99.88/0.00	99.89/99.89/0.00	99.84/99.83/0.01	99.86/99.85/0.01	99.84/99.84/0.00
	Z_{RMS}	72.90/0.88/72.02	81.04/0.88/80.16	96.55/0.55/96.00	95.64/0.65/94.99	83.64/0.90/82.74
	$Z_{\text{NA}}(2,t)$	30.69/0.06/30.63	30.67/0.06/30.61	28.81/0.05/28.76	29.59/0.05/29.54	30.94/0.06/30.88
	Z_{OLS}	100/100/0.00	100/100/0.00	100/100/0.00	100/100/0.00	100/100/0.00
	$Z^{\text{N}}_{\text{QLS}}$	17.45/16.07/1.38	17.67/16.08/1.59	19.52/12.12/7.40	19.77/13.54/6.23	17.50/16.18/1.32
	$Z^{\text{S}}_{\text{OLS}}$	100/99.99/0.01	100/99.99/0.01	100/99.63/0.37	100/99.71/0.29	100/99.98/0.02
	Z_{Quad}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	98.63/1.12/97.51	99.25/0.55/98.70	99.96/0.03/99.93	99.95/0.08/99.87	99.52/0.36/99.16
	$Z_{\text{NA}}[J=4]$	99.88/99.69/0.19	99.89/98.87/1.02	99.87/90.04/9.83	99.91/91.91/8.00	99.86/97.26/2.60
	$Z_{\text{RMS}}[J=4]$	94.47/0.64/93.83	97.72/0.35/97.37	99.87/0.03/99.84	99.79/0.06/99.73	98.95/0.23/98.72
	$Z_{\text{NA}}(2,t)[J=4]$	30.90/0.06/30.84	31.07/0.06/31.01	30.78/0.05/30.73	30.71/0.05/30.66	31.23/0.06/31.17
Scenario F	Z_{LR}	95.37/0.00/95.37	96.51/0.00/96.51	99.46/0.00/99.46	99.33/0.00/99.33	96.54/0.00/96.54
	Z_{NA}	99.99/99.99/0.00	99.99/99.99/0.00	99.98/99.98/0.00	99.99/99.99/0.00	100/100/0.00
	Z_{RMS}	55.45/0.82/54.63	64.37/0.82/63.55	89.54/0.49/89.05	87.41/0.60/86.81	67.87/0.83/67.04
	$Z_{\text{NA}}(2,t)$	4.99/1.93/3.06	5.06/1.95/3.11	4.87/1.87/3.00	5.03/1.94/3.09	5.04/1.97/3.07
	Z_{OLS}	99.96/99.96/0.00	99.96/99.96/0.00	99.88/99.88/0.00	99.89/99.89/0.00	99.96/99.96/0.00
	$Z^{\text{N}}_{\text{QLS}}$	96.63/96.63/0.00	96.64/96.64/0.00	95.37/95.31/0.06	95.74/95.70/0.04	96.74/96.74/0.00
	$Z^{\text{S}}_{\text{OLS}}$	99.96/99.96/0.00	99.96/99.96/0.00	99.89/99.88/0.01	99.91/99.91/0.00	99.96/99.96/0.00
	Z_{Quad}	100/0.00/100	100/0.00/100	100/0.03/99.97	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	98.09/0.00/98.09	98.49/0.00/98.49	99.73/0.00/99.73	99.67/0.00/99.67	97.89/0.00/97.89
	$Z_{\text{NA}}[J=4]$	99.99/99.99/0.00	99.99/99.99/0.00	99.99/98.33/1.66	99.99/98.84/1.15	100/99.90/0.10
	$Z_{\text{RMS}}[J=4]$	80.38/0.77/79.61	87.49/0.70/86.79	98.18/0.17/98.01	97.55/0.22/97.33	90.58/0.64/89.94
	$Z_{\text{NA}}(2,t)[J=4]$	4.97/1.93/3.04	4.93/1.90/3.03	5.67/1.65/4.02	5.50/1.76/3.74	5.08/1.97/3.11

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z^{\text{S}}_{\text{OLS}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.

F.4.5 Results for Recalibrated Boundaries Based on Average Information Growth

In this section, we present results after we recalibrated the final critical value of the monitoring rules based on the assumption of true average information growth for the various test statistics with independent increments. We assume the true average information growth under the null hypothesis for all the monitoring rules. Adjustments to our interim critical values of the monitoring rules are not performed even though the interim statistical information may differ from trial monitoring in this section. Instead, the critical values at the final analysis are adjusted to calibrate the overall Type 1 error rate under the strong null. Additionally, we considered recalibration of the final critical value based on the assumption of an equal amount of error spent at each interim analysis.

For test statistics that do not have independent increments, we assume the monitoring boundaries to be based on the information growth of the log rank test statistic. The information growth for the NA(t) is based on the average information growth based on $NA(2,t)$. The information growth for the non sequential standardized OLS statistic, $Z_{\text{OLS}}^{\text{Fixed}}$, and the quadratic combination statistic are assumed to follow the information growth based on the standardized OLS statistic.

Table F.26 and F.27 show the calibrated boundaries based on average information growth. Table F.24 and F.25 show the recalibrated results under the weak null hypothesis for scenarios A-F based on average information growth.

Table F.24: Probability of rejecting the weak null hypothesis for survival curves that are stochastically ordered without true crossings over the first five years based on average information growth with boundaries calibrated to reject the strong null hypothesis at a fixed 5% error rate for the different test statistics and monitoring rules.

		Alt Hypothesis		
		OBF Overall/Std/Trt	Pocock Overall/Std/Trt	Equal Error Overall/Std/Trt
Scenario A	Z_{LR}	93.89/0.16/93.73	95.90/0.06/95.84	93.11/0.24/92.87
	Z_{NA}	5.67/1.47/4.20	6.29/1.27/5.02	5.29/1.73/3.56
	Z_{RMS}	94.09/0.00/94.09	96.03/0.00/96.03	93.19/0.00/93.19
	$Z_{\text{NA}}(2,t)^*$	26.08/0.05/26.03	25.55/0.07/25.48	26.68/0.05/26.63
	$Z_{\text{OLS Fixed}}$	87.27/87.27/0.00	75.07/75.05/0.02	84.23/84.23/0.00
	$Z^{\text{N}}_{\text{OLS}}$	24.00/0.06/23.94	24.31/0.07/24.24	23.88/0.07/23.81
	$Z^{\text{S}}_{\text{OLS}}$	87.86/87.61/0.25	79.55/77.69/1.86	85.47/84.42/1.05
	$Z^{\text{S}}_{\text{Quad}}$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	96.25/0.13/96.12	98.74/0.01/98.73	98.00/0.13/97.87
	$Z_{\text{NA}}[J=4]$	6.89/1.43/5.46	13.67/0.92/12.75	11.45/1.34/10.11
	$Z_{\text{RMS}}[J=4]$	97.26/0.00/97.26	99.44/0.00/99.44	98.88/0.00/98.88
	$Z_{\text{NA}}(2,t)[J=4]^*$	26.09/0.05/26.04	25.32/0.05/25.27	26.70/0.04/26.66
Scenario B	Z_{LR}	100/0.00/100	100/0.00/100	100/0.00/100
	Z_{NA}	6.57/1.30/5.27	7.39/1.07/6.32	6.57/1.46/5.11
	Z_{RMS}	99.76/0.00/99.76	99.89/0.00/99.89	99.72/0.00/99.72
	$Z_{\text{NA}}(2,t)$	44.48/0.01/44.47	42.42/0.00/42.42	43.96/0.00/43.96
	$Z_{\text{OLS Fixed}}$	74.79/74.79/0.00	64.70/64.64/0.06	68.56/68.53/0.03
	$Z^{\text{N}}_{\text{OLS}}$	27.73/0.05/27.68	30.26/0.03/30.23	27.16/0.05/27.11
	$Z^{\text{S}}_{\text{OLS}}$	75.27/74.71/0.56	68.59/65.13/3.46	71.54/68.92/2.62
	$Z^{\text{S}}_{\text{Quad}}$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}[J=4]$	9.99/1.28/8.71	20.33/0.82/19.51	18.40/1.18/17.22
	$Z_{\text{RMS}}[J=4]$	99.95/0.00/99.95	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)[J=4]$	44.50/0.01/44.49	41.98/0.00/41.98	43.91/0.00/43.91
Scenario C	Z_{LR}	100/0.00/100	100/0.00/100	100/0.00/100
	Z_{NA}	15.04/0.93/14.11	17.87/0.79/17.08	18.04/0.76/17.28
	Z_{RMS}	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)$	94.93/0.00/94.93	94.29/0.00/94.29	94.73/0.00/94.73
	$Z_{\text{OLS Fixed}}$	8.63/8.50/0.13	7.66/5.07/2.59	8.25/6.40/1.85
	$Z^{\text{N}}_{\text{OLS}}$	54.18/0.01/54.17	66.64/0.00/66.64	64.08/0.00/64.08
	$Z^{\text{S}}_{\text{OLS}}$	16.15/8.53/7.62	37.96/5.65/32.31	35.91/6.34/29.57
	$Z^{\text{S}}_{\text{Quad}}$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}[J=4]$	31.32/0.89/30.43	52.60/0.55/52.05	55.08/0.44/54.64
	$Z_{\text{RMS}}[J=4]^*$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)[J=4]$	94.93/0.00/94.93	94.81/0.00/94.81	94.50/0.00/94.50

$Z_{\text{OLS Fixed}}$: OLS statistic based on information growth of $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.
$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z^{\text{S}}_{\text{OLS}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.
$Z^{\text{S}}_{\text{Quad}}$: Assumes the information growth of the standardized form of the linear composite statistic $Z^{\text{S}}_{\text{OLS}}$ where the variance estimator is 2 by time 5 as the information growth for the quadratic statistic.

Table F.25: Probability of rejecting the weak null hypothesis for crossing survival curves based on average information growth with boundaries calibrated to reject the strong null hypothesis at a fixed 5% error rate for the different test statistics and monitoring rules.

		Alt Hypothesis		
		OBF Overall/Std/Trt	Pocock Overall/Std/Trt	Equal Error Overall/Std/Trt
Scenario D	Z_{LR}	100/12.16/87.84	100/5.07/94.93	100/8.07/91.93
	Z_{NA}	99.88/99.87/0.01	99.79/99.73/0.06	99.84/99.81/0.03
	Z_{RMS}	94.60/0.79/93.81	98.19/0.37/97.82	97.02/0.63/96.39
	$Z_{\text{NA}}(2,t)$	51.10/0.00/51.10	49.43/0.00/49.43	51.55/0.00/51.55
	$Z_{\text{OLS Fixed}}$	100/100/0.00	100/100/0.00	100/100/0.00
	$Z^{\text{N}}_{\text{OLS}}$	39.22/0.01/39.21	40.71/0.01/40.70	38.31/0.01/38.30
	$Z^{\text{S}}_{\text{OLS}}$	100/99.94/0.06	100/98.92/1.08	100/99.25/0.75
	$Z^{\text{S}}_{\text{Quad}}$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	99.99/6.98/93.01	100/0.40/99.60	100/0.61/99.39
	$Z_{\text{NA}}[J=4]$	99.89/98.69/1.20	99.82/82.92/16.90	99.85/84.50/15.35
	$Z_{\text{RMS}}[J=4]$	97.92/0.45/97.47	99.92/0.02/99.90	99.89/0.03/99.86
	$Z_{\text{NA}}(2,t)[J=4]$	51.10/0.00/51.10	49.75/0.00/49.75	51.06/0.00/51.06
Scenario E	Z_{LR}	99.23/0.94/98.29	99.77/0.36/99.41	99.66/0.63/99.03
	Z_{NA}	99.87/99.86/0.01	99.82/99.79/0.03	99.85/99.84/0.01
	Z_{RMS}	93.44/0.69/92.75	97.09/0.38/96.71	95.59/0.60/94.99
	$Z_{\text{NA}}(2,t)$	29.86/0.06/29.80	27.47/0.05/27.42	29.13/0.05/29.08
	$Z_{\text{OLS Fixed}}$	100/100/0.00	100/100/0.00	100/100/0.00
	$Z^{\text{N}}_{\text{OLS}}$	18.48/14.43/4.05	18.54/10.25/8.29	19.64/13.24/6.40
	$Z^{\text{S}}_{\text{OLS}}$	100/99.99/0.01	100/99.62/0.38	100/99.70/0.30
	$Z^{\text{S}}_{\text{Quad}}$	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{LR}}[J=4]$	99.52/0.48/99.04	99.97/0.03/99.94	99.95/0.09/99.86
	$Z_{\text{NA}}[J=4]$	99.89/99.08/0.81	99.87/89.50/10.37	99.85/89.69/10.16
	$Z_{\text{RMS}}[J=4]$	97.93/0.35/97.58	99.87/0.02/99.85	99.78/0.06/99.72
	$Z_{\text{NA}}(2,t)[J=4]$	29.86/0.06/29.80	30.37/0.05/30.32	30.83/0.05/30.78
Scenario F	Z_{LR}	99.32/0.00/99.32	99.61/0.00/99.61	99.38/0.00/99.38
	Z_{NA}	99.98/99.98/0.00	99.94/99.94/0.00	99.92/99.92/0.00
	Z_{RMS}	86.72/0.44/86.28	91.31/0.19/91.12	87.83/0.59/87.24
	$Z_{\text{NA}}(2,t)$	4.92/1.92/3.00	5.02/1.94/3.08	4.97/1.92/3.05
	$Z_{\text{OLS Fixed}}$	99.96/99.96/0.00	99.88/99.88/0.00	99.90/99.90/0.00
	$Z^{\text{N}}_{\text{OLS}}$	96.35/96.35/0.00	95.02/94.95/0.07	95.69/95.64/0.05
	$Z^{\text{S}}_{\text{OLS}}$	99.96/99.96/0.00	99.88/99.87/0.01	99.91/99.91/0.00
	$Z^{\text{S}}_{\text{Quad}}$	100/0.00/100	100/0.03/99.97	100/0.00/100
	$Z_{\text{LR}}[J=4]$	99.38/0.00/99.38	99.76/0.00/99.76	99.70/0.00/99.70
	$Z_{\text{NA}}[J=4]$	99.98/99.93/0.05	99.95/98.07/1.88	99.85/97.33/2.52
	$Z_{\text{RMS}}[J=4]$	92.64/0.39/92.25	98.30/0.03/98.27	97.47/0.22/97.25
	$Z_{\text{NA}}(2,t)[J=4]$	4.92/1.92/3.00	5.67/1.65/4.02	5.95/1.47/4.48

$Z_{\text{OLS Fixed}}$: OLS statistic based on information growth of $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.
$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z^{\text{S}}_{\text{OLS}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.
$Z^{\text{S}}_{\text{Quad}}$: Assumes the information growth of the standardized form of the linear composite statistic $Z^{\text{S}}_{\text{OLS}}$ where the variance estimator is 2 by time 5 as the information growth for the quadratic statistic.

Table F.26: Original and recalibrated boundaries to ensure a fixed overall Type 1 error for various test statistics under the different monitoring rules for survival curves that are stochastically ordered without true crossings over the first five years when presuming true average information growth.

		OBF (True Avg Info)					POC		POC (Equal Error Spent)			
		$Z_{t=2}$	$Z_{t=3}$	$Z_{t=4}$	$Z_{t=5}$	$Z_{t=5}^{\text{Rev}}$	Z_t	Z_t^{Rev}	$Z_{t=2}^{\text{Rev}}$	$Z_{t=3}^{\text{Rev}}$	$Z_{t=4}^{\text{Rev}}$	$Z_{t=5}^{\text{Rev}}$
Scenario A	Z_{LR}	-	2.363	2.114	2.057	2.163	2.161	2.350	-	2.436	2.237	2.042
	Z_{NA}	-	2.388	2.093	2.053	2.381	2.162	3.071	-	2.446	2.290	2.110
	Z_{RMS}	-	2.363	2.114	2.057	2.108	2.161	2.262	-	2.435	2.227	2.036
	$Z_{\text{NA}}(2,t)$	-	2.388	2.093	2.053	2.164	2.162	2.477	-	2.444	2.259	2.056
	$Z_{\text{OLS Fixed}}$	-	3.729	2.529	1.996	1.994	2.301	2.311	-	2.452	2.341	2.171
	$Z_{\text{OLS}}^{\text{N}}$	-	2.389	2.094	2.053	2.169	2.163	2.561	-	2.442	2.252	2.058
	$Z_{\text{OLS}}^{\text{S}}$	-	2.965	2.313	2.024	2.169	2.256	2.561	-	2.442	2.252	2.058
	Z_{Quad}	-	3.729	2.529	1.996	2.041	2.301	2.453	-	2.447	2.358	2.276
	Z_{LR}	3.133	2.367	2.118	2.061	2.176	2.288	2.622	2.543	2.436	2.290	2.126
	Z_{NA}	3.626	2.389	2.094	2.053	2.404	2.308	3.071	2.452	2.476	2.372	2.211
	Z_{RMS}	3.133	2.367	2.118	2.061	2.118	2.288	2.569	2.540	2.425	2.282	2.116
	$Z_{\text{NA}}(2,t)$	3.626	2.389	2.094	2.053	2.168	2.308	2.359	2.452	2.438	2.325	2.141
Scenario B	Z_{LR}	-	2.363	2.114	2.057	2.057	2.161	2.176	-	2.387	2.206	2.017
	Z_{NA}	-	2.485	2.114	2.051	2.231	2.184	2.400	-	2.366	2.233	2.121
	Z_{RMS}	-	2.363	2.114	2.057	2.068	2.161	2.175	-	2.391	2.197	2.018
	$Z_{\text{NA}}(2,t)$	-	2.485	2.114	2.051	2.051	2.184	2.223	-	2.404	2.234	2.048
	$Z_{\text{OLS Fixed}}$	-	4.406	2.616	1.989	1.965	2.318	2.162	-	2.374	2.240	2.163
	$Z_{\text{OLS}}^{\text{N}}$	-	2.508	2.125	2.051	2.036	2.189	2.201	-	2.392	2.218	2.048
	$Z_{\text{OLS}}^{\text{S}}$	-	3.058	2.325	2.021	2.036	2.263	2.201	-	2.392	2.218	2.048
	Z_{Quad}	-	4.406	2.616	1.989	1.989	2.318	2.479	-	2.397	2.395	2.281
	Z_{LR}	3.133	2.368	2.118	2.061	2.057	2.289	2.273	2.473	2.386	2.258	2.115
	Z_{NA}	3.956	2.485	2.114	2.051	2.231	2.326	2.512	2.395	2.418	2.342	2.215
	Z_{RMS}	3.133	2.368	2.118	2.061	2.068	2.289	2.296	2.470	2.392	2.262	2.112
	$Z_{\text{NA}}(2,t)$	3.956	2.485	2.114	2.051	2.051	2.326	2.260	2.395	2.440	2.304	2.135
Scenario C	Z_{LR}	-	2.364	2.114	2.057	2.086	2.161	2.178	-	2.449	2.226	2.031
	Z_{NA}	-	2.716	2.179	2.042	2.152	2.223	2.200	-	2.141	2.194	2.198
	Z_{RMS}	-	2.364	2.114	2.057	2.159	2.161	2.341	-	2.412	2.237	2.088
	$Z_{\text{NA}}(2,t)$	-	2.716	2.179	2.042	2.050	2.223	2.165	-	2.358	2.242	2.083
	$Z_{\text{OLS Fixed}}$	-	4.530	2.633	1.987	1.995	2.321	2.230	-	2.381	2.304	2.203
	$Z_{\text{OLS}}^{\text{N}}$	-	2.905	2.250	2.032	2.054	2.247	2.207	-	2.342	2.270	2.159
	$Z_{\text{OLS}}^{\text{S}}$	-	3.277	2.368	2.014	2.054	2.277	2.207	-	2.342	2.270	2.159
	Z_{Quad}	-	4.530	2.633	1.987	2.043	2.321	2.479	-	2.369	2.435	2.277
	Z_{LR}	3.136	2.368	2.118	2.061	2.100	2.289	2.413	2.505	2.456	2.275	2.111
	Z_{NA}	4.642	2.716	2.179	2.042	2.148	2.358	2.197	2.210	2.213	2.291	2.274
	Z_{RMS}	3.136	2.368	2.118	2.061	2.182	2.289	2.853	2.493	2.416	2.321	2.176
	$Z_{\text{NA}}(2,t)$	4.642	2.716	2.179	2.042	2.050	2.358	2.137	2.210	2.440	2.339	2.170

$Z_{\text{OLS}}^{\text{N}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z_{\text{OLS}}^{\text{S}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.

Table F.27: Original and recalibrated boundaries to ensure a fixed overall Type 1 error for various test statistics under the different monitoring rules for crossing survival curves that are stochastically ordered without true crossings over the first five years when presuming true average information growth.

		OBF (True Avg Info)					POC		POC (Equal Error Spent)			
		$Z_{t=2}$	$Z_{t=3}$	$Z_{t=4}$	$Z_{t=5}$	$Z_{t=5}^{\text{Rev}}$	Z_t	Z_t^{Rev}	$Z_{t=2}^{\text{Rev}}$	$Z_{t=3}^{\text{Rev}}$	$Z_{t=4}^{\text{Rev}}$	$Z_{t=5}^{\text{Rev}}$
Scenario D	Z_{LR}	-	2.725	2.254	2.037	2.087	2.232	2.400	-	2.462	2.289	2.138
	Z_{NA}	-	2.552	2.134	2.049	2.159	2.197	2.503	-	2.381	2.269	2.277
	Z_{RMS}	-	2.725	2.254	2.037	2.031	2.232	2.246	-	2.463	2.273	2.085
	$Z_{\text{NA}}(2,t)$	-	2.552	2.134	2.049	2.120	2.197	2.290	-	2.410	2.242	2.081
	$Z_{\text{OLS Fixed}}$	-	5.073	2.809	1.977	2.006	2.332	2.355	-	2.467	2.378	2.236
	$Z_{\text{OLS}}^{\text{N}}$	-	2.555	2.136	2.049	2.115	2.198	2.315	-	2.412	2.248	2.082
	$Z_{\text{OLS}}^{\text{S}}$	-	3.226	2.401	2.011	2.115	2.277	2.315	-	2.412	2.248	2.082
	Z_{Quad}	-	5.073	2.809	1.977	1.998	2.332	2.389	-	2.382	2.373	2.327
	Z_{LR}	3.860	2.726	2.254	2.037	2.089	2.348	2.670	2.530	2.478	2.360	2.236
	Z_{NA}	4.442	2.552	2.134	2.049	2.159	2.342	2.660	2.424	2.449	2.360	2.342
	Z_{RMS}	3.860	2.726	2.254	2.037	2.032	2.348	2.366	2.539	2.480	2.332	2.172
	$Z_{\text{NA}}(2,t)$	4.442	2.552	2.134	2.049	2.120	2.342	2.277	2.424	2.452	2.323	2.165
Scenario E	Z_{LR}	-	2.586	2.201	2.047	2.059	2.211	2.193	-	2.376	2.231	2.093
	Z_{NA}	-	2.615	2.151	2.047	2.074	2.208	2.250	-	2.238	2.234	2.183
	Z_{RMS}	-	2.586	2.201	2.047	2.041	2.211	2.197	-	2.392	2.237	2.067
	$Z_{\text{NA}}(2,t)$	-	2.615	2.151	2.047	2.042	2.208	2.264	-	2.390	2.250	2.085
	$Z_{\text{OLS Fixed}}$	-	5.141	2.785	1.978	1.954	2.332	2.214	-	2.407	2.301	2.175
	$Z_{\text{OLS}}^{\text{N}}$	-	2.704	2.190	2.042	2.046	2.223	2.266	-	2.387	2.257	2.102
	$Z_{\text{OLS}}^{\text{S}}$	-	3.268	2.397	2.011	2.046	2.278	2.266	-	2.387	2.257	2.102
	Z_{Quad}	-	5.141	2.785	1.978	1.977	2.332	2.325	-	2.341	2.322	2.328
	Z_{LR}	3.571	2.587	2.202	2.047	2.068	2.330	2.393	2.524	2.383	2.308	2.173
	Z_{NA}	4.640	2.615	2.151	2.047	2.076	2.350	2.258	2.343	2.301	2.336	2.279
	Z_{RMS}	3.571	2.587	2.202	2.047	2.043	2.330	2.424	2.528	2.414	2.313	2.162
	$Z_{\text{NA}}(2,t)$	4.640	2.615	2.151	2.047	2.042	2.350	2.173	2.343	2.425	2.320	2.134
Scenario F	Z_{LR}	-	2.418	2.134	2.056	2.123	2.176	2.246	-	2.383	2.245	2.079
	Z_{NA}	-	2.730	2.183	2.041	2.046	2.225	2.125	-	2.161	2.158	2.160
	Z_{RMS}	-	2.418	2.134	2.056	2.206	2.176	2.447	-	2.377	2.269	2.112
	$Z_{\text{NA}}(2,t)$	-	2.730	2.183	2.041	2.041	2.225	2.160	-	2.330	2.245	2.099
	Z_{OLS}	-	4.688	2.666	1.985	1.998	2.324	2.245	-	2.420	2.302	2.197
	$Z_{\text{OLS}}^{\text{N}}$	-	2.951	2.266	2.029	2.073	2.252	2.233	-	2.347	2.304	2.156
	$Z_{\text{OLS}}^{\text{S}}$	-	3.315	2.381	2.012	2.073	2.280	2.233	-	2.347	2.304	2.156
	Z_{Quad}	-	4.688	2.666	1.985	1.995	2.324	2.416	-	2.416	2.361	2.279
	Z_{LR}	3.244	2.421	2.137	2.058	2.128	2.301	2.348	2.485	2.380	2.307	2.153
	Z_{NA}	4.763	2.730	2.183	2.041	2.046	2.361	2.110	2.191	2.244	2.263	2.250
	Z_{RMS}	3.244	2.421	2.137	2.058	2.206	2.301	2.947	2.515	2.406	2.342	2.206
	$Z_{\text{NA}}(2,t)$	4.763	2.730	2.183	2.041	2.041	2.361	2.132	2.191	2.383	2.336	2.209

$Z_{\text{OLS}}^{\text{N}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.
$Z_{\text{OLS}}^{\text{S}}$: Standardized form of the linear composite statistic $U_S/\sqrt{\text{Var}(U_S)}$ where the variance estimator is 2 by time 5.

F.4.6 Results for Recalibrated Boundaries Based on Constrained Boundaries

In Table F.28, the overall Type 1 error rate is inflated mildly for various test statistics when adjusting the boundaries either by approach (a) or (b) based on the observed statistical information under the strong null hypothesis. This is observed, particularly for Scenario A and D when our accrual size is 10,800, to quantify the difference in survival where $S(2) \approx 0.9$. In particular, when the information growth is sufficiently close across consecutive interim analyses, results from Proschan et al. [1992] suggest that this may lead to an inflation of the overall Type 1 error even under the null hypothesis. We observed that previously, the Type 1 error rate based on the constrained boundaries approach, when applied to a Pocock stopping rule, do not necessary conform to the typical 95% CI of 5% Type 1 error of anywhere from 4.58% and 5.43% based on 10,000 simulations.

Occasionally, our nominal Type 1 error rate may be inflated when we monitor the trial on a calendar basis. This can be a consequence of delayed information growth accumulating across calendar time, thus resulting in information growth that is too close apart. However, we can correct for this inflation by choosing to either adjusting the final boundaries, or recalibrating the monitoring boundary according to a different level α, for the monitoring boundary of choice, and re-evaluating this revised boundary using simulations to determine the control of the overall Type 1 error.

We described the latter approach by redefining our boundaries using a smaller α to maintain the overall Type 1 error rate. This is similar to the notion of Neyman-Pearson lemma, where we make statistical adjustments to calibrate the overall Type 1 error, while trying to maintain the desired power under some chosen alternatives. This means that we reapply the constrained boundaries approach to evaluate the true size of the monitoring boundaries after shifting the α level of the boundary. We can then evaluate the weak null hypothesis once our desired overall Type 1 error rate is held at roughly level $\alpha = 5\%$.

Table F.32 and F.33 show the results of such an approach whereby a level α' monitoring boundary is prespecified, and the constrained boundaries approach is applied to the simulated

data to evaluate the overall Type 1 error rate. By adjusting the size of the α' used to construct the monitoring boundary, we can maintain the error rate under the string null scenario so that they are as close to the desired level of $\alpha = 5\%$. This in turn changes the power for the various test statistics when we applied the same level α' monitoring rule respectively to the alternative scenarios A-F (Table F.30 and F.31). Despite being able to control the overall Type 1 error rate under the strong null, there are limitations since there are inherent differences in the rate of information growth for the various test statistics. Thus, this makes it harder to allow us to appropriately compare the boundaries in a fair manner.

Table F.28: Probability of rejecting the strong null hypothesis according to overall, in favor of standard of care (Std), or treatment (Trt) for the constrained boundaries method based on the true information growth monitored on the calendar time scale using the monitoring boundaries. The Type 1 error rate under the strong null are not all within the 95%CI of a typical 5% error rate.

		Contrained Boundaries		Revised Information	
		OBF (a) Both/Std/Trt	OBF (b) Both/Std/Trt	Pocock (a) Both/Std/Trt	Pocock (b) Both/Std/Trt
A	Z_{LR}	5.62/2.88/2.74	5.34/2.73/2.61	5.55/2.83/2.72	5.36/2.74/2.62
	$Z_{\text{NA}}(2,t)$	6.23/3.24/2.99	5.52/2.89/2.63	6.27/3.26/3.01	6.14/3.15/2.99
	$Z^{\text{N}}_{\text{OLS}}$	5.59/2.83/2.76	5.46/2.84/2.62	5.59/2.84/2.75	5.62/2.85/2.77
	$Z_{\text{LR}}[J=4]$	5.62/2.88/2.74	5.38/2.76/2.62	5.46/2.82/2.64	5.45/2.79/2.66
B	Z_{LR}	5.04/2.40/2.64	4.98/2.35/2.63	5.03/2.58/2.45	4.95/2.42/2.53
	$Z_{\text{NA}}(2,t)$	5.47/2.69/2.78	4.93/2.45/2.48	5.67/2.89/2.78	5.56/2.78/2.78
	$Z^{\text{N}}_{\text{OLS}}$	4.97/2.36/2.61	4.89/2.36/2.53	5.03/2.51/2.52	5.08/2.50/2.58
	$Z_{\text{LR}}[J=4]$	5.05/2.41/2.64	5.02/2.37/2.65	4.91/2.53/2.38	4.99/2.44/2.55
C	Z_{LR}	5.29/2.51/2.78	5.14/2.54/2.60	5.01/2.40/2.61	5.12/2.50/2.62
	$Z_{\text{NA}}(2,t)$	5.06/2.55/2.51	4.96/2.48/2.48	5.05/2.53/2.52	5.12/2.53/2.59
	$Z^{\text{N}}_{\text{OLS}}$	5.17/2.58/2.59	5.21/2.65/2.56	4.84/2.45/2.39	5.05/2.49/2.56
	$Z_{\text{LR}}[J=4]$	5.33/2.55/2.78	5.10/2.49/2.61	5.22/2.62/2.60	5.23/2.52/2.71
D	Z_{LR}	5.37/2.79/2.58	5.40/2.96/2.44	5.60/2.94/2.66	5.52/2.89/2.63
	$Z_{\text{NA}}(2,t)$	5.64/2.96/2.68	5.45/2.84/2.61	5.62/2.94/2.68	5.73/3.03/2.70
	$Z^{\text{N}}_{\text{OLS}}$	5.42/2.85/2.57	5.37/2.82/2.55	5.33/2.79/2.54	5.28/2.77/2.51
	$Z_{\text{LR}}[J=4]$	5.38/2.80/2.58	5.39/2.86/2.53	5.57/2.90/2.67	5.61/2.90/2.71
E	Z_{LR}	5.08/2.54/2.54	5.11/2.59/2.52	4.97/2.49/2.48	4.97/2.47/2.50
	$Z_{\text{NA}}(2,t)$	5.12/2.69/2.43	4.69/2.50/2.19	5.46/2.78/2.68	5.44/2.77/2.67
	$Z^{\text{N}}_{\text{OLS}}$	5.05/2.60/2.45	5.00/2.67/2.33	5.25/2.60/2.65	5.17/2.59/2.58
	$Z_{\text{LR}}[J=4]$	5.10/2.56/2.54	5.15/2.57/2.58	5.20/2.65/2.55	5.20/2.62/2.58
F	Z_{LR}	5.42/2.64/2.78	5.43/2.65/2.78	5.24/2.53/2.71	5.32/2.62/2.70
	$Z_{\text{NA}}(2,t)$	4.97/2.51/2.46	5.00/2.51/2.49	4.86/2.50/2.36	4.92/2.45/2.47
	$Z^{\text{N}}_{\text{OLS}}$	4.99/2.37/2.62	4.90/2.43/2.47	5.24/2.63/2.61	5.05/2.52/2.53
	$Z_{\text{LR}}[J=4]$	5.41/2.63/2.78	5.41/2.64/2.77	5.12/2.66/2.46	5.27/2.75/2.52

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite $U_N/\sqrt{\text{Var}(U_N)}$.

Table F.29: Probability of rejecting the weak null hypothesis according to overall, in favor of standard of care (Std), or treatment (Trt) for the constrained boundaries method based on the true information growth monitored on the calendar time scale using the monitoring boundaries. Note that from previous table, the Type 1 error rate under the strong null may not be appropriately controlled

		Contrained Boundaries		Revised Information	
		OBF (a) Both/Std/Trt	OBF (b) Both/Std/Trt	Pocock(a) Both/Std/Trt	Pocock (b) Both/Std/Trt
A	Z_{LR}	95.34/0.17/95.17	68.18/0.27/67.91	96.54/0.10/96.44	94.85/0.26/94.59
	$Z_{\text{NA}}(2,t)$	30.98/0.08/30.90	33.88/0.04/33.84	30.61/0.10/30.51	41.43/0.14/41.29
	$Z^{\text{N}}_{\text{OLS}}$	26.63/0.34/26.29	26.02/0.24/25.78	26.43/0.36/26.07	26.15/0.23/25.92
	$Z_{\text{LR}}[J=4]$	97.11/0.16/96.95	87.36/0.26/87.10	98.94/0.04/98.90	98.66/0.11/98.55
B	Z_{LR}	100/0.00/100	99.97/0.00/99.97	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)$	47.80/0.00/47.80	46.63/0.10/46.53	47.80/0.00/47.80	48.36/0.12/48.24
	$Z^{\text{N}}_{\text{OLS}}$	28.03/0.41/27.62	22.86/0.40/22.46	31.00/0.37/30.63	28.41/0.39/28.02
	$Z_{\text{LR}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
C	Z_{LR}	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
	$Z_{\text{NA}}(2,t)$	95.53/0.00/95.53	95.49/0.53/94.96	95.26/0.00/95.26	95.26/0.53/94.73
	$Z^{\text{N}}_{\text{OLS}}$	49.14/0.17/48.97	43.40/0.17/43.23	65.41/0.17/65.24	65.20/0.17/65.03
	$Z_{\text{LR}}[J=4]$	100/0.00/100	100/0.00/100	100/0.00/100	100/0.00/100
D	Z_{LR}	100/10.52/89.48	99.72/33.39/66.33	100/4.58/95.42	100/5.83/94.17
	$Z_{\text{NA}}(2,t)$	54.85/0.00/54.85	54.35/0.00/54.35	54.28/0.00/54.28	54.74/0.00/54.74
	$Z^{\text{N}}_{\text{OLS}}$	39.86/0.02/39.84	35.60/0.02/35.58	41.55/0.01/41.54	39.45/0.02/39.43
	$Z_{\text{LR}}[J=4]$	100/5.67/94.33	100/9.57/90.43	100/0.38/99.62	100/0.40/99.60
E	Z_{LR}	99.54/0.99/98.55	91.65/2.40/89.25	99.81/0.36/99.45	99.80/0.58/99.22
	$Z_{\text{NA}}(2,t)$	32.02/0.06/31.96	30.54/0.06/30.48	31.02/0.05/30.97	31.56/0.06/31.50
	$Z^{\text{N}}_{\text{OLS}}$	19.52/15.43/4.09	17.55/16.17/1.38	20.80/12.19/8.61	21.18/13.78/7.40
	$Z_{\text{LR}}[J=4]$	99.71/0.48/99.23	98.67/1.16/97.51	99.97/0.03/99.94	99.97/0.04/99.93
F	Z_{LR}	99.39/0.00/99.39	95.37/0.00/95.37	99.64/0.00/99.64	99.46/0.00/99.46
	$Z_{\text{NA}}(2,t)$	5.08/1.93/3.15	5.07/2.01/3.06	4.99/1.88/3.11	5.22/2.05/3.17
	$Z^{\text{N}}_{\text{OLS}}$	96.87/96.87/0.00	96.90/96.90/0.00	95.26/95.20/0.06	95.35/95.29/0.06
	$Z_{\text{LR}}[J=4]$	99.47/0.00/99.47	98.09/0.00/98.09	99.76/0.00/99.76	99.73/0.00/99.73

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite $U_N/\sqrt{\text{Var}(U_N)}$.

Table F.30: Summary of the "calibrated" Type 1 error (strong null) and the probability of rejecting the weak null hypothesis according to overall, in favor of either the standard of care (Std) or treatment (Trt) for the constrained boundaries method based on the true information growth monitored on the calendar time scale using the OBF monitoring boundaries.

		Strong Null					Weak Null				
		$t=2$ Std/Trt	$t=3$ Std/Trt	$t=4$ Std/Trt	$t=5$ Std/Trt	Overall Both/Std/Trt	$t=2$ Std/Trt	$t=3$ Std/Trt	$t=4$ Std/Trt	$t=5$ Std/Trt	Overall Both/Std/Trt
A	Z_{LR}		1.56/1.48	0.67/0.69	0.28/0.26	4.94/2.51/2.43		0.00/96.09	0.00/0.00	0.08/0.00	96.17/0.08/96.09
	$Z_{\text{NA}}(2,t)$		1.30/1.35	0.16/0.16	1.11/1.02	5.10/2.57/2.53		0.03/19.42	0.01/3.27	0.01/4.89	27.63/0.05/27.58
	$Z^{\text{N}}_{\text{OLS}}$		0.90/0.77	1.17/1.29	0.50/0.42	5.05/2.57/2.48		0.30/13.27	0.02/8.49	0.02/3.02	25.12/0.34/24.78
	$Z_{\text{LR}}[J=4]$	1.26/0.93	0.75/0.75	0.36/0.50	0.21/0.22	4.98/2.58/2.40	0.00/98.27	0.00/0.49	0.00/0.00	0.03/0.00	98.79/0.03/98.76
B	Z_{LR}		0.88/0.95	1.07/1.14	0.45/0.55	5.04/2.40/2.64		0.00/100	0.00/0.00	0.00/0.00	100/0.00/100
	$Z_{\text{NA}}(2,t)$		1.44/1.39	0.11/0.09	1.02/1.01	5.06/2.57/2.49		0.00/32.40	0.00/0.60	0.00/13.19	46.19/0.00/46.19
	$Z^{\text{N}}_{\text{OLS}}$		0.56/0.65	1.15/1.25	0.65/0.71	4.97/2.36/2.61		0.34/17.89	0.01/8.52	0.06/1.21	28.03/0.41/27.62
	$Z_{\text{LR}}[J=4]$	0.05/0.06	0.82/0.88	1.09/1.13	0.45/0.57	5.05/2.41/2.64	0.00/100	0.00/0.00	0.00/0.00	0.00/0.00	100/0.00/100
C	Z_{LR}		0.91/0.94	0.96/1.22	0.51/0.48	5.02/2.38/2.64		0.00/100	0.00/0.00	0.00/0.00	100/0.00/100
	$Z_{\text{NA}}(2,t)$		0.56/0.54	0.15/0.03	1.84/1.94	5.06/2.55/2.51		0.00/67.64	0.00/0.35	0.00/27.54	95.53/0.00/95.53
	$Z^{\text{N}}_{\text{OLS}}$		0.13/0.12	1.04/1.13	1.30/1.21	4.93/2.47/2.46		0.16/31.85	0.00/15.04	0.01/0.95	48.01/0.17/47.84
	$Z_{\text{LR}}[J=4]$	0.16/0.09	0.82/0.87	0.94/1.19	0.50/0.48	5.05/2.42/2.63	0.00/100	0.00/0.00	0.00/0.00	0.00/0.00	100/0.00/100
D	Z_{LR}		1.41/1.24	0.75/0.76	0.49/0.43	5.08/2.65/2.43		0.00/95.05	0.23/0.00	4.72/0.00	100/4.95/95.05
	$Z_{\text{NA}}(2,t)$		1.60/1.29	0.00/0.02	1.09/1.04	5.04/2.69/2.35		0.00/34.97	0.00/0.02	0.00/17.34	52.33/0.00/52.33
	$Z^{\text{N}}_{\text{OLS}}$		0.59/0.37	1.42/1.24	0.54/0.70	4.86/2.55/2.31		0.00/20.24	0.00/15.88	0.02/2.19	38.33/0.02/38.31
	$Z_{\text{LR}}[J=4]$	0.99/0.91	0.85/0.71	0.48/0.49	0.34/0.31	5.08/2.66/2.42	0.00/99.51	0.00/0.09	0.04/0.00	0.36/0.00	100/0.40/99.60
E	Z_{LR}		0.48/0.41	1.08/1.12	0.98/1.01	5.08/2.54/2.54		0.00/98.55	0.00/0.00	0.99/0.00	99.54/0.99/98.55
	$Z_{\text{NA}}(2,t)$		0.75/0.85	0.03/0.01	1.91/1.57	5.12/2.69/2.43		0.02/13.93	0.00/0.05	0.04/17.98	32.02/0.06/31.96
	$Z^{\text{N}}_{\text{OLS}}$		0.32/0.27	1.21/1.30	1.07/0.88	5.05/2.60/2.45		0.02/3.31	0.70/0.78	14.71/0.00	19.52/15.43/4.09
	$Z_{\text{LR}}[J=4]$	0.03/0.02	0.47/0.41	1.08/1.11	0.98/1.00	5.10/2.56/2.54	0.00/95.90	0.00/3.33	0.00/0.00	0.48/0.00	99.71/0.48/99.23
F	Z_{LR}		0.69/0.66	1.05/1.21	0.63/0.66	4.90/2.37/2.53		0.00/99.58	0.00/0.01	0.00/0.00	99.59/0.00/99.59
	$Z_{\text{NA}}(2,t)$		0.22/0.19	0.08/0.08	2.21/2.19	4.97/2.51/2.46		0.18/0.76	0.03/0.11	1.72/2.28	5.08/1.93/3.15
	$Z^{\text{N}}_{\text{OLS}}$		0.00/0.00	0.00/0.00	2.37/2.62	4.99/2.37/2.62		0.41/0.00	37.86/0.00	58.60/0.00	96.87/96.87/0.00
	$Z_{\text{LR}}[J=4]$	0.03/0.04	0.63/0.64	1.06/1.21	0.61/0.65	4.87/2.33/2.54	0.00/98.81	0.00/0.93	0.00/0.01	0.00/0.00	99.75/0.00/99.75

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.

Table F.31: Summary of the "calibrated" Type 1 error (strong null) and the probability of rejecting the weak null hypothesis according to overall, in favor of either the standard of care (Std) or treatment (Trt) for the constrained boundaries method based on the true information growth monitored on the calendar time scale using the Pocock monitoring boundaries.

		Strong Null					Weak Null				
		$t=2$ Std/Trt	$t=3$ Std/Trt	$t=4$ Std/Trt	$t=5$ Std/Trt	Overall Both/Std/Trt	$t=2$ Std/Trt	$t=3$ Std/Trt	$t=4$ Std/Trt	$t=5$ Std/Trt	Overall Both/Std/Trt
A	Z_{LR}		1.56/1.48	0.67/0.69	0.28/0.26	4.94/2.51/2.43		0.00/96.09	0.00/0.00	0.08/0.00	96.17/0.08/96.09
	$Z_{\text{NA}}(2,t)$		1.50/1.59	0.12/0.10	0.95/0.80	5.06/2.57/2.49		0.05/20.69	0.01/2.70	0.01/4.05	27.51/0.07/27.44
	$Z^{\text{N}}_{\text{OLS}}$		1.55/1.55	0.75/0.74	0.26/0.16	5.01/2.56/2.45		0.32/18.34	0.02/4.80	0.01/1.65	25.14/0.35/24.79
	$Z_{\text{LR}}[J=4]$	1.26/0.93	0.75/0.75	0.36/0.50	0.21/0.22	4.98/2.58/2.40	0.00/98.27	0.00/0.49	0.00/0.00	0.03/0.00	98.79/0.03/98.76
B	Z_{LR}		1.64/1.37	0.68/0.77	0.26/0.31	5.03/2.58/2.45		0.00/100	0.00/0.00	0.00/0.00	100/0.00/100
	$Z_{\text{NA}}(2,t)$		1.44/1.39	0.11/0.09	1.02/1.01	5.06/2.57/2.49		0.00/32.40	0.00/0.60	0.00/13.19	46.19/0.00/46.19
	$Z^{\text{N}}_{\text{OLS}}$		1.44/1.38	0.71/0.78	0.36/0.36	5.03/2.51/2.52		0.35/26.38	0.01/3.90	0.01/0.35	31.00/0.37/30.63
	$Z_{\text{LR}}[J=4]$	1.05/1.10	0.83/0.66	0.47/0.43	0.18/0.19	4.91/2.53/2.38	0.00/100	0.00/0.00	0.00/0.00	0.00/0.00	100/0.00/100
C	Z_{LR}		1.41/1.52	0.66/0.72	0.23/0.25	4.79/2.30/2.49		0.00/100	0.00/0.00	0.00/0.00	100/0.00/100
	$Z_{\text{NA}}(2,t)$		1.39/1.22	0.03/0.01	1.11/1.29	5.05/2.53/2.52		0.00/78.40	0.00/0.18	0.00/16.68	95.26/0.00/95.26
	$Z^{\text{N}}_{\text{OLS}}$		1.10/0.96	0.71/0.86	0.56/0.47	4.66/2.37/2.29		0.16/59.71	0.00/4.67	0.01/0.07	64.62/0.17/64.45
	$Z_{\text{LR}}[J=4]$	1.09/1.03	0.72/0.84	0.41/0.49	0.24/0.14	4.96/2.46/2.50	0.00/100	0.00/0.00	0.00/0.00	0.00/0.00	100/0.00/100
D	Z_{LR}		1.41/1.24	0.75/0.76	0.49/0.43	5.08/2.65/2.43		0.00/95.05	0.23/0.00	4.72/0.00	100/4.95/95.05
	$Z_{\text{NA}}(2,t)$		1.60/1.29	0.00/0.02	1.09/1.04	5.04/2.69/2.35		0.00/34.97	0.00/0.02	0.00/17.34	52.33/0.00/52.33
	$Z^{\text{N}}_{\text{OLS}}$		1.46/1.19	0.77/0.64	0.26/0.41	4.73/2.49/2.24		0.00/30.31	0.00/8.43	0.00/0.92	39.66/0.00/39.66
	$Z_{\text{LR}}[J=4]$	0.99/0.91	0.85/0.71	0.48/0.49	0.34/0.31	5.08/2.66/2.42	0.00/99.51	0.00/0.09	0.04/0.00	0.36/0.00	100/0.40/99.60
E	Z_{LR}		1.33/1.25	0.74/0.78	0.42/0.45	4.97/2.49/2.48		0.00/99.45	0.00/0.00	0.36/0.00	99.81/0.36/99.45
	$Z_{\text{NA}}(2,t)$		1.50/1.60	0.01/0.01	1.27/1.07	5.46/2.78/2.68		0.02/18.74	0.00/0.03	0.03/12.20	31.02/0.05/30.97
	$Z^{\text{N}}_{\text{OLS}}$		1.34/1.43	0.72/0.80	0.54/0.42	5.25/2.60/2.65		0.11/8.34	0.62/0.27	11.46/0.00	20.80/12.19/8.61
	$Z_{\text{LR}}[J=4]$	1.07/1.01	0.74/0.71	0.51/0.51	0.33/0.32	5.20/2.65/2.55	0.00/99.85	0.00/0.09	0.00/0.00	0.03/0.00	99.97/0.03/99.94
F	Z_{LR}		1.26/1.36	0.74/0.73	0.27/0.40	4.76/2.27/2.49		0.00/99.58	0.00/0.01	0.00/0.00	99.59/0.00/99.59
	$Z_{\text{NA}}(2,t)$		1.19/1.03	0.04/0.04	1.27/1.29	4.86/2.50/2.36		0.87/1.83	0.02/0.04	0.99/1.24	4.99/1.88/3.11
	$Z^{\text{N}}_{\text{OLS}}$		1.18/1.09	0.00/0.00	1.45/1.52	5.24/2.63/2.61		5.82/0.06	37.65/0.00	51.73/0.00	95.26/95.20/0.06
	$Z_{\text{LR}}[J=4]$	1.08/0.83	0.63/0.63	0.52/0.47	0.21/0.28	4.65/2.44/2.21	0.00/98.81	0.00/0.93	0.00/0.01	0.00/0.00	99.75/0.00/99.75

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite statistic $U_N/\sqrt{\text{Var}(U_N)}$.

Table F.32: Overall Type 1 error for the constrained boundaries method after adjusting the OBF boundaries for the test statistics with independent increments. The boundaries are re-calibrated by shifting the α of the group sequential design and recomputing the probability of rejecting the strong null. Each cell is the computed probability of rejecting the strong null analyzed on the calendar time according to overall, in favor of standard of care (Std), or in favor of treatment (Trt).

		OBF						
		$\alpha = 4.0\%$ Both/Std/Trt	$\alpha = 4.25\%$ Both/Std/Trt	$\alpha = 4.35\%$ Both/Std/Trt	$\alpha = 4.4\%$ Both/Std/Trt	$\alpha = 4.5\%$ Both/Std/Trt	$\alpha = 4.75\%$ Both/Std/Trt	$\alpha = 5\%$ Both/Std/Trt
A	Z_{LR}	4.41/2.19/2.22	4.74/2.40/2.34	4.86/2.47/2.39	4.93/2.52/2.41	4.94/2.51/2.43	5.38/2.76/2.62	5.62/2.88/2.74
	$Z_{\text{LR}}[J=4]$	4.40/2.19/2.21	4.69/2.37/2.32	4.84/2.46/2.38	4.92/2.51/2.41	4.98/2.58/2.40	5.38/2.75/2.63	5.62/2.88/2.74
	$Z_{\text{NA}}(2,t)$	5.10/2.57/2.53	5.41/2.74/2.67	5.47/2.77/2.70	5.51/2.79/2.72	5.64/2.89/2.75	5.95/3.09/2.86	6.23/3.24/2.99
	$Z^{\text{N}}_{\text{OLS}}$	4.52/2.35/2.17	4.80/2.46/2.34	4.89/2.49/2.40	4.92/2.51/2.41	5.05/2.57/2.48	5.28/2.69/2.59	5.59/2.83/2.76
B	Z_{LR}	4.06/1.96/2.10	4.35/2.08/2.27	4.48/2.14/2.34	4.51/2.16/2.35	4.53/2.35/2.18	4.86/2.34/2.52	5.04/2.40/2.64
	$Z_{\text{LR}}[J=4]$	4.04/1.95/2.09	4.32/2.06/2.26	4.48/2.14/2.34	4.51/2.16/2.35	4.38/2.23/2.15	4.85/2.34/2.51	5.05/2.41/2.64
	$Z_{\text{NA}}(2,t)$	4.42/2.19/2.23	4.84/2.42/2.42	4.87/2.42/2.45	4.90/2.44/2.46	5.06/2.57/2.49	5.32/2.64/2.68	5.47/2.69/2.78
	$Z^{\text{N}}_{\text{OLS}}$	4.01/1.91/2.10	4.22/2.02/2.20	4.34/2.05/2.29	4.40/2.08/2.32	4.49/2.13/2.36	4.70/2.23/2.47	4.97/2.36/2.61
C	Z_{LR}	4.21/1.97/2.24	4.44/2.08/2.36	4.51/2.12/2.39	4.55/2.12/2.43	4.68/2.26/2.42	5.02/2.38/2.64	5.29/2.51/2.78
	$Z_{\text{LR}}[J=4]$	4.22/1.98/2.24	4.48/2.10/2.38	4.51/2.11/2.40	4.57/2.15/2.42	4.81/2.39/2.42	5.05/2.42/2.63	5.33/2.55/2.78
	$Z_{\text{NA}}(2,t)$	4.08/2.05/2.03	4.34/2.19/2.15	4.45/2.24/2.21	4.48/2.25/2.23	4.53/2.31/2.22	4.81/2.42/2.39	5.06/2.55/2.51
	$Z^{\text{N}}_{\text{OLS}}$	4.10/1.99/2.11	4.49/2.17/2.32	4.57/2.21/2.36	4.63/2.25/2.38	4.72/2.32/2.40	4.93/2.47/2.46	5.17/2.58/2.59
D	Z_{LR}	4.30/2.20/2.10	4.61/2.40/2.21	4.70/2.46/2.24	4.73/2.47/2.26	5.08/2.65/2.43	5.13/2.68/2.45	5.37/2.79/2.58
	$Z_{\text{LR}}[J=4]$	4.30/2.20/2.10	4.61/2.40/2.21	4.70/2.46/2.24	4.73/2.47/2.26	5.08/2.66/2.42	5.12/2.68/2.44	5.38/2.80/2.58
	$Z_{\text{NA}}(2,t)$	4.35/2.32/2.03	4.73/2.49/2.24	4.85/2.53/2.32	4.92/2.57/2.35	5.04/2.69/2.35	5.33/2.79/2.54	5.64/2.96/2.68
	$Z^{\text{N}}_{\text{OLS}}$	4.27/2.27/2.00	4.56/2.42/2.14	4.65/2.47/2.18	4.71/2.50/2.21	4.86/2.55/2.31	5.15/2.68/2.47	5.42/2.85/2.57
E	Z_{LR}	4.10/2.03/2.07	4.32/2.09/2.23	4.47/2.14/2.33	4.50/2.15/2.35	4.49/2.25/2.24	4.87/2.40/2.47	5.08/2.54/2.54
	$Z_{\text{LR}}[J=4]$	4.13/2.04/2.09	4.34/2.10/2.24	4.50/2.15/2.35	4.53/2.16/2.37	4.60/2.28/2.32	4.90/2.41/2.49	5.10/2.56/2.54
	$Z_{\text{NA}}(2,t)$	4.08/2.11/1.97	4.34/2.22/2.12	4.39/2.25/2.14	4.41/2.25/2.16	4.90/2.43/2.47	4.83/2.50/2.33	5.12/2.69/2.43
	$Z^{\text{N}}_{\text{OLS}}$	4.02/2.07/1.95	4.26/2.18/2.08	4.36/2.22/2.14	4.42/2.27/2.15	4.55/2.37/2.18	4.84/2.50/2.34	5.05/2.60/2.45
F	Z_{LR}	4.31/2.05/2.26	4.61/2.21/2.40	4.71/2.26/2.45	4.79/2.28/2.51	4.90/2.37/2.53	5.18/2.53/2.65	5.42/2.64/2.78
	$Z_{\text{LR}}[J=4]$	4.31/2.04/2.27	4.61/2.21/2.40	4.72/2.27/2.45	4.76/2.27/2.49	4.87/2.33/2.54	5.18/2.53/2.65	5.41/2.63/2.78
	$Z_{\text{NA}}(2,t)$	4.21/2.14/2.07	4.43/2.26/2.17	4.52/2.29/2.23	4.57/2.32/2.25	4.62/2.34/2.28	4.77/2.40/2.37	4.97/2.51/2.46
	$Z^{\text{N}}_{\text{OLS}}$	3.96/1.94/2.02	4.24/2.06/2.18	4.34/2.13/2.21	4.42/2.16/2.26	4.51/2.18/2.33	4.73/2.24/2.49	4.99/2.37/2.62

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite $U_N/\sqrt{\text{Var}(U_N)}$.

Table F.33: Overall Type 1 error for the constrained boundaries method after adjusting the Pocock boundaries for the test statistics with independent increments. The boundaries are re-calibrated by shifting the α of the group sequential design and recomputing the probability of rejecting the strong null. Each cell is the computed probability of rejecting the strong null analyzed on the calendar time according to overall, in favor of standard of care (Std), or in favor of treatment (Trt).

		Pocock						
		$\alpha = 4.0\%$ Both/Std/Trt	$\alpha = 4.25\%$ Both/Std/Trt	$\alpha = 4.35\%$ Both/Std/Trt	$\alpha = 4.4\%$ Both/Std/Trt	$\alpha = 4.5\%$ Both/Std/Trt	$\alpha = 4.75\%$ Both/Std/Trt	$\alpha = 5\%$ Both/Std/Trt
A	Z_{LR}	4.46/2.22/2.24	4.70/2.36/2.34	4.76/2.38/2.38	4.81/2.42/2.39	4.94/2.51/2.43	5.19/2.63/2.56	5.55/2.83/2.72
	$Z_{\text{LR}}[J=4]$	4.41/2.28/2.13	4.71/2.44/2.27	4.83/2.50/2.33	4.87/2.52/2.35	4.98/2.58/2.40	5.25/2.71/2.54	5.46/2.82/2.64
	$Z_{\text{NA}}(2,t)$	5.06/2.57/2.49	5.36/2.74/2.62	5.48/2.81/2.67	5.48/2.81/2.67	5.64/2.89/2.75	5.92/3.06/2.86	6.27/3.26/3.01
	$Z^{\text{N}}_{\text{OLS}}$	4.45/2.27/2.18	4.70/2.39/2.31	4.79/2.42/2.37	4.88/2.48/2.40	5.01/2.56/2.45	5.21/2.68/2.53	5.59/2.84/2.75
B	Z_{LR}	4.03/2.10/1.93	4.26/2.20/2.06	4.35/2.25/2.10	4.43/2.28/2.15	4.53/2.35/2.18	4.80/2.47/2.33	5.03/2.58/2.45
	$Z_{\text{LR}}[J=4]$	3.86/1.89/1.97	4.10/2.05/2.05	4.20/2.10/2.10	4.26/2.14/2.12	4.38/2.23/2.15	4.67/2.41/2.26	4.91/2.53/2.38
	$Z_{\text{NA}}(2,t)$	4.38/2.20/2.18	4.73/2.36/2.37	4.92/2.47/2.45	4.93/2.48/2.45	5.06/2.57/2.49	5.37/2.73/2.64	5.67/2.89/2.78
	$Z^{\text{N}}_{\text{OLS}}$	4.02/1.94/2.08	4.29/2.11/2.18	4.37/2.16/2.21	4.41/2.19/2.22	4.54/2.26/2.28	4.79/2.40/2.39	5.03/2.51/2.52
C	Z_{LR}	4.18/2.03/2.15	4.41/2.14/2.27	4.52/2.20/2.32	4.58/2.22/2.36	4.68/2.26/2.42	4.79/2.30/2.49	5.01/2.40/2.61
	$Z_{\text{LR}}[J=4]$	4.31/2.14/2.17	4.56/2.28/2.28	4.65/2.32/2.33	4.68/2.32/2.36	4.81/2.39/2.42	4.96/2.46/2.50	5.22/2.62/2.60
	$Z_{\text{NA}}(2,t)$	4.01/2.05/1.96	4.23/2.15/2.08	4.35/2.22/2.13	4.41/2.26/2.15	4.53/2.31/2.22	4.79/2.43/2.36	5.05/2.53/2.52
	$Z^{\text{N}}_{\text{OLS}}$	3.83/1.99/1.84	4.15/2.15/2.00	4.27/2.21/2.06	4.32/2.22/2.10	4.40/2.24/2.16	4.66/2.37/2.29	4.84/2.45/2.39
D	Z_{LR}	4.44/2.33/2.11	4.78/2.51/2.27	4.88/2.54/2.34	4.91/2.56/2.35	5.08/2.65/2.43	5.39/2.86/2.53	5.60/2.94/2.66
	$Z_{\text{LR}}[J=4]$	4.47/2.35/2.12	4.73/2.47/2.26	4.85/2.55/2.30	4.93/2.59/2.34	5.08/2.66/2.42	5.31/2.76/2.55	5.57/2.90/2.67
	$Z_{\text{NA}}(2,t)$	4.49/2.37/2.12	4.68/2.49/2.19	4.86/2.58/2.28	4.93/2.63/2.30	5.04/2.69/2.35	5.30/2.82/2.48	5.62/2.94/2.68
	$Z^{\text{N}}_{\text{OLS}}$	4.22/2.24/1.98	4.52/2.41/2.11	4.57/2.44/2.13	4.61/2.45/2.16	4.73/2.49/2.24	5.04/2.65/2.39	5.33/2.79/2.54
E	Z_{LR}	4.02/2.04/1.98	4.19/2.12/2.07	4.34/2.18/2.16	4.38/2.21/2.17	4.49/2.25/2.24	4.73/2.38/2.35	4.97/2.49/2.48
	$Z_{\text{LR}}[J=4]$	4.01/2.04/1.97	4.36/2.20/2.16	4.50/2.26/2.24	4.54/2.27/2.27	4.60/2.28/2.32	4.95/2.47/2.48	5.20/2.65/2.55
	$Z_{\text{NA}}(2,t)$	4.18/2.08/2.10	4.48/2.23/2.25	4.62/2.30/2.32	4.65/2.32/2.33	4.90/2.43/2.47	5.18/2.63/2.55	5.46/2.78/2.68
	$Z^{\text{N}}_{\text{OLS}}$	3.98/1.92/2.06	4.31/2.11/2.20	4.44/2.17/2.27	4.50/2.22/2.28	4.58/2.26/2.32	4.89/2.43/2.46	5.25/2.60/2.65
F	Z_{LR}	4.11/1.99/2.12	4.46/2.13/2.33	4.53/2.16/2.37	4.62/2.19/2.43	4.76/2.27/2.49	5.07/2.48/2.59	5.24/2.53/2.71
	$Z_{\text{LR}}[J=4]$	4.19/2.18/2.01	4.46/2.36/2.10	4.49/2.37/2.12	4.52/2.39/2.13	4.65/2.44/2.21	4.92/2.52/2.40	5.12/2.66/2.46
	$Z_{\text{NA}}(2,t)$	3.92/2.00/1.92	4.05/2.06/1.99	4.20/2.13/2.07	4.21/2.14/2.07	4.35/2.20/2.15	4.63/2.36/2.27	4.86/2.50/2.36
	$Z^{\text{N}}_{\text{OLS}}$	4.02/1.92/2.10	4.33/2.12/2.21	4.45/2.20/2.25	4.50/2.22/2.28	4.64/2.30/2.34	4.92/2.47/2.45	5.24/2.63/2.61

$Z^{\text{N}}_{\text{OLS}}$: Non-standardized form of the linear composite $U_N/\sqrt{\text{Var}(U_N)}$.

VITA

William Koh Jen Hoe was born and raised in Singapore. He earned a Bachelor of Science Degree with First Class Honours in Statistics from National University of Singapore in 2004. He previously worked at the Defence Science Organisation in Singapore, Health Promotion Board in Singapore before taking on a 18-months internship with Economic Development Board to learn Statistical Genetics at Eli Lilly. After which, he worked at Lilly Singapore Centre for Drug Discovery (LSCDD) prior to graduate school. Before starting graduate school, he completed 5 marathons before turning 30 and has since stopped running. In 2013, he earned a Masters in Biostatistics from the University of Washington. In 2016, he earned a Doctor of Philosophy from the University of Washington in Biostatistics.