UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Methods for Enriched and Adaptively Randomized Clinical Trials

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Navneet Ram Hakhu

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Professor Bin Nan
Assistant Professor Tianchen Qian
Professor Joshua D. Grill

2024

DEDICATION

To Scott

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to start by thanking the members of my dissertation committee: Bin Nan, Tianchen Qian, and Josh Grill. I appreciate all of your thoughtful questions, insights, and advice that have helped me grow and will continue to propel me as a scientific investigator.

Thank you to Michele Guindani who served as a member of my advancement to candidacy committee. Your enthusiasm and kindness are qualities I strive to embody.

Thank you to David Sultzer who allowed me to participate in the UCI Alzheimer's Disease Research Center weekly Consensus Conference and monthly Clinicopathologic Conference. Your encouragement and support have increased my understanding of clinical practice versus clinical research that informs my approach to addressing scientific questions with statistics.

To the Statistics Department and UCI MIND, thank you to the staff, faculty, instructors, classmates, and peers. It has been an honor to work alongside and learn from everyone and a pleasure to get to know each other as colleagues and friends.

To my family and friends, I am forever thankful for your support from near and far. I cherish that you are a part of my life and that I am a part of yours.

Thank you to the Statistics Department for funding me as a teaching assistant and graduate student researcher; both opportunities helped me grow as a well-rounded statistician. And thank you to the National Institutes of Health for a number of training and research experiences. Notably, the research reported in this dissertation was supported by the National Institute on Aging of the National Institutes of Health under Award Numbers T32AG000096, RF1AG059407, and F31AG077880. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

An extra thank you to my co-advisor Josh Grill who, from the first day we met, has always had my best interests in mind. You have taught me through your actions and advice how to be a diligent and sound clinical trialist and scientific investigator. You have also helped remind me to take a moment at times and think about the bigger picture, including what is important to me in my life. Thank you, Josh.

Thank you to my advisor Dan Gillen. You epitomize what it means to be a statistician, clinical trialist, and collaborator. Your fierce dedication to do what is right while maintaining high levels of consistency serve as my targets in work and life. From your wide-ranging advice, patience, and friendship, you have instilled an extra level of resilience in me that, prior to joining UCI, I did not know I could have. Thank you for introducing me to Josh and for the opportunity to work with you and the Grillen lab. I am eternally grateful. Thank you, Dan.

Finally, to Scott Emerson, without whom I would not be here. You never gave up on me. You have always empowered me, including on July 5, 2010 when you said: "*Do you know at what age I learned to read? No, but all that matters is that I can read now.*" While I will continue to learn, I can read now. Thank you, Scott. I dedicate this dissertation to you.

VITA

Navneet Ram Hakhu

EDUCATION

Doctor of Philosophy in Statistics — **September 2024**
University of California, Irvine — *Irvine, California*

Master of Science in Statistics — **December 2021**
University of California, Irvine — *Irvine, California*

Master of Science in Biostatistics — **December 2014**
University of Washington, Seattle — *Seattle, Washington*

Bachelor of Science in Mathematics — **March 2009**
Santa Clara University — *Santa Clara, California*

RESEARCH EXPERIENCE

Graduate Student Researcher — **October 2023 – September 2024**
University of California, Irvine — *Irvine, California*

ORISE Fellow — **May 2022 – September 2022**
U.S. Food and Drug Administration — *Silver Spring, Maryland*

Graduate Student Researcher — **March 2022 – May 2022**
University of California, Irvine — *Irvine, California*

NIH/NIA Training Grant Fellow — **September 2019 – September 2021**
University of California, Irvine — *Irvine, California*

UCI MIND Associate Specialist III — **September 2018 – September 2019**
University of California, Irvine — *Irvine, California*

Research Assistant — **January 2013 – September 2014**
University of Washington, Seattle — *Seattle, Washington*

Research Assistant — **September 2009 – September 2012**
University of Washington, Seattle — *Seattle, Washington*

Research Assistant — **June 2008 – September 2008**
Santa Clara University — *Santa Clara, California*

TEACHING EXPERIENCE

Instructor of Record — **Summer 2023**
University of California, Irvine — *Irvine, California*

Teaching Assistant — **Spring 2023**
University of California, Irvine — *Irvine, California*

Reader — **Winter 2023**
University of California, Irvine — *Irvine, California*

Teaching Assistant — **Fall 2022**
University of California, Irvine — *Irvine, California*

Teaching Assistant — **Winter 2022**
University of California, Irvine — *Irvine, California*

Teaching Assistant — **Fall 2021**
University of California, Irvine — *Irvine, California*

Statistics Bootcamp Teaching Assistant — **September 2020**
University of California, Irvine — *Irvine, California*

Data Science Initiative Short Course Instructor — **August 2020**
University of California, Irvine — *Irvine, California*

Teaching Assistant — **Fall 2012**
University of Washington, Seattle — *Seattle, Washington*

Grader — **Fall 2008**
Santa Clara University — *Santa Clara, California*

REFEREED JOURNAL PUBLICATIONS

Dyadic Enrollment in a Phase 3 Mild Cognitive Impairment Clinical Trial **2022**
Alzheimer Disease and Associated Disorders

swCRTdesign: An R Package for Stepped Wedge Trial Design and Analysis **2020**
Computer Methods and Programs in Biomedicine

SOFTWARE

swCRTdesign `https://CRAN.R-project.org/package=swCRTdesign`
`R` package for Stepped Wedge Cluster Randomized Trial Design co-developed with Jim Hughes, Emily Voldal, and Fan Xia; version 4.0 available on CRAN.

ABSTRACT OF THE DISSERTATION

Statistical Methods for Enriched and Adaptively Randomized Clinical Trials

By

Navneet Ram Hakhu

Doctor of Philosophy in Statistics

University of California, Irvine, 2024

Professor Daniel L. Gillen, Chair

Randomized controlled clinical trials (RCTs) serve as the gold standard to determine whether a candidate treatment has a favorable benefit-to-risk ratio for a pre-specified target patient population. Enrichment strategies are commonly employed in RCTs to identify the appropriate target population of patients who would likely benefit from a candidate treatment (predictive) and/or have the outcome of interest during the course of the trial (prognostic), including enriching based upon amyloid beta and tau protein levels in Alzheimer's disease (U.S. Food and Drug Administration, 2019b). Currently, there is a gap in the understanding of RCTs using enrichment and adaptations to the randomized treatment assignment allocations (response-adaptive randomization), especially under model misspecification (violation of assumptions) for fixed sample and adaptive RCTs with a repeated measures (e.g., changes in activities of daily living scores) or a censored time-to-event (e.g., time to dementia) primary outcome.

In this dissertation, we focus on valid estimation of estimands (a contrast of summary measures between treatment arms for an appropriate target population) from enriched and adaptively randomized clinical trials. We consider the trial-specific RCT estimand (RCT-E) and a real world estimand (RW-E) for a broader patient population for whom off-label use may be a possibility. Aim 1 quantifies the impact of enrichment in fixed sample pre-post (only

two assessments; one pre- and one post-randomization) RCTs. We show that the application of standard statistical methods, such as the analysis of covariance (ANCOVA) model, yield biased estimates for the RW-E. We propose a novel bias-adjusted estimator of the RCT-E to estimate the RW-E based on an analytic derivation under model misspecification in the multivariate normal data setting. Aim 2 focuses on reliably estimating the RCT-E in fixed sample adaptively randomized time-to-event RCTs that allows for enhanced replicability in the presence of time-varying treatment effects. We propose a novel adaptive randomization censoring-robust estimator that reweights the partial likelihood score à la Boyd et al. (2012) that accounts for differential censoring patterns resulting from adaptive randomization and by incorporating the randomization scheme in the re-weighting, we gain efficiency. Importantly, our proposed estimator consistently estimates a standardized marginal hazard ratio for the RCT-E. Finally, in Aim 3 we examine how to prospectively design and plan for the monitoring of time-to-event group sequential designs that warrant using censoring-robust estimators when targeting a RCT-E. We show how the statistical information of our proposed adaptive randomization censoring-robust estimator is non-linear and has non-independent increments, thus requiring appropriate modifications to the planned timing of interim analyses and the final boundary to maintain statistical operating characteristics and scientific objectives of such trials. Overall, the statistical contributions of this research will aid in the design, conduct, and analysis of enriched and adaptively randomized clinical trials to support efforts during drug development, regulatory review, and clinical decision-making post approval.

Chapter 1

Introduction

1.1 Overview

Medical interventions are designed to prevent, slow the progression of, or treat disease for patients from a specific target population. Defining the target population often depends on several considerations, including: (i) scientific and clinical rationale about the disease; (ii) the mechanism of action of the investigational product under study; (iii) the potential vulnerability of patients from particular populations that may indicate an *a priori* potential safety concern, including having pre-existing conditions such as comorbidities (study exclusion criteria); and (iv) characteristics that define the afflicted patient population (study inclusion criteria, that may be used to enrich study samples).

Clinical trials are experiments conducted on human volunteers. At minimum, the primary aim of a clinical trial is to determine whether a candidate intervention is (ideally, causally) associated with a favorable benefit-to-risk ratio for a clinically meaningful outcome of interest. Benefit corresponds to the efficacy and risk corresponds to the safety of a candidate intervention under study. Typically human volunteers are assigned to one of at least two

interventions to aid in the evaluation of *benefit-risk*: experimental arm and control arm (e.g., placebo or standard of care). Ideally to further clinical understanding and practice, we want to establish a causal association (cause and effect) between an intervention and clinically meaningful outcome.

Randomization, if ethical, is a mechanism used in clinical trials that ensures eligible human volunteers (study participants) are randomly assigned to either the experimental or control arm. Consequently, randomization on average balances intervention arms across all measured and unmeasured confounders—factors that are causally associated with the outcome and also causally associated with the predictor of interest (intervention assignment). Such experiments are called *randomized controlled clinical trials (RCTs)*. A central tenet allowing the possibility of randomization of individuals to a control arm (even if placebo) is that prior to the start of the RCT there is inconclusive evidence suggesting benefit or harm for the experimental arm versus control arm—termed *clinical equipoise*. However, if after the start of the RCT, reliable evidence emerges from the monitoring of this RCT or from external sources that invalidates the clinical equipoise assumption, it would no longer be ethical to at least randomize newly enrolled patients to one of the two arms. In such a situation, a Data Monitoring Committee (DMC) — a group of typically three to five experts with collective expertise and experience in clinical trials, clinical practice, therapeutic areas, ethics, and biostatistics, independent from the trial sponsor, tasked with monitoring study conduct and emerging safety and efficacy data as a trial is ongoing — would typically convene and formulate a recommendation to the sponsor taking the newly emerged information into consideration as it pertains to the trial participants they have been monitoring.

Furthermore, to protect against spurious results and data-driven analyses, *pre-specification* of statistical analyses of the pre-specified primary and secondary outcomes (that correspond to scientific questions of interest) must be documented and decided upon before looking at the data. Ideally, the statistical analyses should be determined before the start of a

study. Accordingly, both the within-group measure of treatment effect and the between-group comparison of treatment effect need to be determined to evaluate the potential for benefit.

Overall, evaluation of the benefit-risk ratio requires valid estimation of treatment efficacy coupled with a careful assessment of the safety profile (e.g., treatment-emergent adverse events). In this dissertation, we focus on valid estimation of treatment efficacy. Two phenomena that can complicate obtaining valid estimation of treatment efficacy in RCTs are: (1) effect modification (i.e., treatment heterogeneity across subpopulations); and (2) time-varying treatment effects. In the presence of effect modification, the selection of the full analysis set population and/or recruitment strategies (or lack thereof) of certain subpopulations of patients may alter the estimated treatment efficacy in a trial. In the presence of time-varying treatment effects, the length of follow-up of enrolled (and randomized) trial participants may alter the estimated treatment efficacy in a trial. These phenomena can arise in the fixed sample RCT setting. Furthermore, altering the estimation of treatment efficacy can be compounded for RCTs designed with a sequential testing plan, referred to as adaptive design RCTs. Sequential testing can lead to early stopping of a trial for efficacy, safety/harm, or futility, reducing the length of follow-up among trial participants and up- or down-weighting early signals in favor of or against the experimental treatment. For these reasons, it is therefore of importance that the design of every RCT should include collaborative discussions among trial leadership about the potential for and relevance of effect modification and time-varying effects as it pertains to the trial's pre-specified primary question of interest. To facilitate this process, examination of the potential impact of both phenomena in the estimation of treatment efficacy for a given RCT design (both fixed sample and adaptive) is necessary for trial sponsors and scientific review boards to make well-informed decisions prior to starting trial recruitment.

Ultimately, an important two-fold objective for any RCT is to infer results from the trial

that are replicable and that generalize to the appropriate target population. Replicability of results can impact the regulatory evaluation of candidate drugs and biologics. Generalizability can impact clinical practice by way of the label of approved drugs and biologics for commercial use by regulatory agencies (e.g., U.S. Food and Drug Administration, FDA). An important sequela of regulatory approval of a drug or biologic is the potential for off-label use — when an intervention is approved for a specific target population that is indicated on the drug label, but ends up being used more broadly among individuals not meeting the criteria of the patient population on the packet insert label.

In this dissertation, as is ubiquitously done for clinical trials in practice, we consider both the target of inference (contrast of summary measures between treatment arms) and the appropriate target population of interest, together known as the *estimand*. Estimands are used throughout statistical inference to answer scientific questions, whether explicitly mentioned or not. Simply put, an estimand is defined as the quantity of interest to be estimated, known as the target of inference — often times in RCTs this is a between-group comparison of within-group summary measures (e.g., difference in means; ratio of hazards) — for a particular target population of interest. One target population of interest corresponds to the population that comprises the enrolled RCT sample. We know, however, that often we are not simply interested in results pertaining to only those individuals who participate in trials. There are a variety of reasons for this, including the lack of representation of clinical trial samples for the intended target population of interest. Representativeness of the sample is vital not only in RCTs but in all studies (e.g., observational studies and sample surveys). This is a concept emphasized repeatedly in introductory statistics courses at the undergraduate level, yet in research settings and practice, we sometimes lose sight of this basic, albeit extremely critical concept.

To this end, estimands form the basis of our investigations in this dissertation to: (i) quantify the impact of enrichment strategies and adaptive randomization in RCTs; and (ii) develop

new statistical methods to reduce bias and improve generalizability of results to the appropriate target populations. We focus on two types of estimands: the RCT estimand (RCT-E) and a real world estimand (RW-E) corresponding to the trial-specific population and broader patient population (that may include off-label use), respectively.

1.2 Enrichment

The choice of target population relates to the scientific objectives and questions set to answer. Because of possible heterogeneity of treatment effects across subpopulations of patients (e.g., due to health disparities), medical interventions are not a panacea for the entire patient population. Enrichment strategies using baseline patient characteristics (e.g., biomarker status or surrogate outcomes) are used to decide which patients to randomize — those likely to benefit from treatment (prognostic) or have the outcome during the trial (predictive; to permit shorter, smaller trials).

The FDA guidance document on enrichment strategies (2019b) summarizes the relevance and commonality of enrichment strategies in an array of disease areas. Predictive enrichment examples include: systolic or diastolic dysfunction in congestive heart failure; high renin hypertension for assignment to beta-blockers or ACE inhibitors; responders based on protein or genetic markers for breast cancer (HER2), lung cancer (epidermal growth factor receptor), and cystic fibrosis (transmembrane conductance regulator mutation). Prognostic enrichment is divided into event- vs. progression-based to identify high-risk patients. Examples include: ACE inhibitors in heart failure (enalapril trials); history of heart disease and high cholesterol (statin trials); high-risk for breast cancer women using the Gail model (adjuvant therapy trials of tamoxifen); history of exacerbation in chronic pulmonary obstructive disease; MRI findings for multiple sclerosis; and amyloid beta and tau proteins in AD.

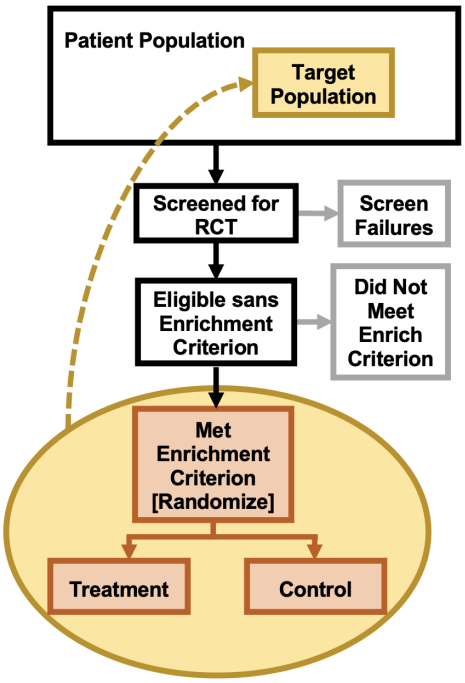


Figure 1.1: An example of a fixed enrichment RCT.

Enrichment criteria either remain the same for the duration of recruitment (fixed enrichment; see Figure 1.1) or are modified (adaptive enrichment). In either case, the generalizability of trial results may be complicated in settings where only the "enriched" patients are randomized, but in clinical practice "non-enriched" patients are likely to receive treatment off-label. Since the signing of the Prescription Drug User Fee Act (PDUFA) VI (U.S. Food and Drug Administration, 2016; Dabrowska and Thaul, 2018), the FDA's pilot program on complex innovative design clinical trials for new drug applications is a mechanism to examine and better understand adaptive designs that play important roles in drug discovery to address scientific questions in timely, feasible, and ethical ways.

Implementing enrichment strategies in RCTs aligns with the National Plan to Address Alzheimer's Disease Strategy 1.B (2020) to expand research to develop disease-modifying treatments. Examples of enrichment strategies in early AD RCTs include: (i) fixed enrichment: the UC Cures Nicotinamide as an Early AD Treatment (NEAT) trial (2021,

ClinicalTrials.gov identifier NCT03061474), a phase 2 proof-of-concept pre-post (only two assessments: one pre- and one post-randomization) RCT where patients are randomized to nicotinamide or placebo if they meet biomarker criteria based on cerebral spinal fluid amyloid beta level or a ratio of total tau to amyloid beta; and (ii) response-adaptive enrichment (response-adaptive randomization and enrichment): BAN2401-G000-201 (2024, ClinicalTrials.gov identifier NCT01767311), a Bayesian adaptive phase 2b proof-of-concept longitudinal RCT where response adaptive randomization was used to assign patients to one of five doses of lecanemab or placebo. Additionally, during the conduct of the BAN2401-G000-201 trial changes were made to the studied patient population as requested by the European Medicines Agency due to concerns of amyloid-related imaging abnormalities-edema (ARIA-E) among apolipoprotein E $\varepsilon 4$ carriers (AlzForum, nd). Such changes to the sample alters the pre-specified target population and further complicates the estimation of treatment effect. While enrichment strategies can expedite efforts to identify candidate treatments, understanding the impact of enrichment on the statistical analysis of treatment efficacy in RCTs when statistical assumptions are violated is critical for obtaining reliable evidence used to evaluate benefit-risk, with potential downstream consequences for providers and patients.

1.3 Adaptive randomization

In an effort to accelerate drug development and potentially enhance the ethical conduct of trials, adaptations to traditional RCT designs may be considered. Adaptive designs constitute a broad class of RCT designs in which adaptations have ideally been pre-specified. The FDA guidance document (2019a) classifies adaptive designs into two types: well-understood and less well-understood. In particular, when multiple adaptations are planned or considered in a RCT, the typical statistical operating characteristics to evaluate such designs become complicated and require simulation studies to assess. Hence, designs with multiple adapta-

tions, such as adapting patient allocation and adapting enrichment, are currently classified as less well-understood. This is particularly true with respect to the impact on estimated treatment effects and the degree to which estimates may be biased for such designs. Prior research examining adaptive enrichment (Simon and Simon, 2013) has focused primarily on controlling the type I error rate as opposed to possible bias due to violation of assumptions.

Another type of adaptation to a trial is changing the treatment assignment allocation (randomization ratios). In particular, altering the randomization ratios based on accumulating outcome (response) data is called response-adaptive randomization has garnered attention in recent years based on the work by Wei and Durham (1978) on randomized Play-the-Winner. Prior work on response-adaptive randomization has looked at frequentist methods (Simon and Simon, 2013; Proschan and Evans, 2020, 2021) and Bayesian methods (Thall, 2021). Proponents of response-adaptive randomization (Thall, 2021; Rosenberger et al., 2012) argue that patients in a clinical trial are more likely to receive the treatment that is beneficial compared to a fixed randomization scheme. On the other hand, others argue that response-adaptive randomization creates issues that are not present with fixed randomization, such as inefficiency compared to fixed randomization (Korn and Freidlin, 2022) and biased estimates of treatment effect in the presence of temporal trends (Proschan and Evans, 2020, 2021); for the latter, it has been suggested to use a block randomization scheme and block-stratified analysis when the outcome of interest is binary (Korn and Freidlin, 2011). While there are mixed reviews of the merits of response-adaptive randomization in practice, there are limitations in the understanding of the impact of response-adaptive randomization on statistical operating characteristics, whether they be frequentist or Bayesian. Without understanding the operating characteristics of candidate designs we run the risk of exposing increased numbers of trial participants to treatments that are ineffective or harmful because we may not obtain accurate and precise treatment effect estimates for the target population to that we wanted to generalize to.

Furthermore, prior research (Proschan and Evans, 2020, 2021) suggests that bias is induced for estimated treatment effects when using response-adaptive randomization and when there are temporal trends — notably, the bias would be towards earlier occurring treatment effects. That said, previous work has assumed a correctly specified model. There are, however, several ways that a violation of assumptions may arise with time-to-event outcomes (Ye and Shao, 2020). Examples of misspecification with time-to-event outcomes include dependent censoring, non-proportional hazards, and time trends. Furthermore, while time-to-event outcomes are commonly specified in RCTs, yet little attention has been paid to time-varying effects with adaptive randomization.

1.4 Dissertation aims

In order to ensure reproducible and reliable RCT results, *a priori* specification of statistical methods to estimate treatment effects is essential (and required). Violation of assumptions for statistical methods may result in bias (tendency to systematically over- or under- estimate treatment effects). Biased estimates can lead to approval of less effective therapies, in the best case, and approval of potentially harmful or ineffective therapies or missing an effective therapy, in the worst case, as a consequence of over- or under-estimating treatment effects.

Currently, there is a gap in the understanding of enriched pre-post RCTs and adaptively randomized time-to-event trials because little attention has been paid to violation of statistical assumptions with respect to the mean-variance relationship and time trends, respectively. Because biased estimates alter the benefit-risk ratio, there is a critical unmet need to evaluate the impact of enriched and adaptively randomized clinical trials for AD and other diseases.

To this end, the three aims of this dissertation are:

Aim 1. To develop methodology for bias-adjusted estimation of the real world estimand

(RW-E) from a fixed sample pre-post RCT with a continuous primary outcome and a fixed pre-randomization enrichment strategy.

Aim 2. To extend methods for censoring-robust estimation of the RCT estimand (RCT-E) in fixed sample time-to-event trials with adaptive randomization in the presence of time-varying treatment effects.

Aim 3. To evaluate and quantify non-linear statistical information growth of censoring-robust estimators of the RCT-E in adaptively randomized time-to-event group sequential trials that will facilitate the design, monitoring, and timing of interim analyses.

The rationale for this dissertation is to identify scenarios in which enrichment and adaptive randomization yield biased estimates, develop methods to correct for the bias induced by such designs, and ensure valid inference can be obtained to better inform all stakeholders (e.g., patients, sponsors, regulatory).

The remainder of this dissertation begins with Chapter 2, providing an overview of pre-post designs, time-to-event analyses, and group sequential designs, serving as supplementary background material for Aims 1-3. Chapter 3 examines the use of enrichment in pre-post RCTs for a RW-E (Aim 1). Chapter 4 transitions to the RCT-E and the role of adaptive randomization in time-to-event RCTs in the fixed sample setting (Aim 2). The methodology developed there will then be used in Chapter 5 that extends our investigation of censoring-robust estimators of the RCT-E to the time-to-event group sequential setting and how to design and implement a sequential monitoring plan using our adaptive randomization censoring-robust estimator (Aim 3). Finally, Chapter 6 includes a summary and future directions.

Chapter 2

Background

2.1 Pre-Post designed randomized trials

A pre-post designed RCTs typically consists of two outcome assessment times: one at pre-randomization and one at post-randomization. Often in a pre-post design the corresponding target of inference or estimand, denoted by θ, is the difference in mean change from baseline (pre-randomization) comparing a subpopulation of patients randomly assigned to treatment to a subpopulation of patients randomly assigned to control (e.g., placebo). The two-sample t-test, analysis of covariance (ANCOVA), and paired change are three linear regression models commonly used to estimate θ. In the typical RCT setting, each of these three models consistently estimate θ and have the same interpretation for θ. The ANCOVA model, however, is more efficient, and hence preferred, in settings when the correlation is at least 0.5 between pre and post assessments.

The RCT-E for a fixed pre-post design, denoted by θ, is the change from baseline comparing two subpopulations of patients, those randomized to treatment versus those randomized to control. In the RCT setting, three typical analytic approaches can be used to estimate θ in

the RCT setting.

- Two-sample t-test: $E[Post|Tx] = \beta_0^* + \beta_1 Tx$
- ANCOVA: $E[Post|Tx, Pre] = \beta_0 + \beta_1 Tx + \beta_2 Pre$
- Paired Change: $E[(Post - Pre)|Tx] = \beta_0^{**} + \beta_1 Tx$

With no enrichment, RCT-E $\theta = \beta_1$ where β_1 corresponds to the marginal population parameter for patients from a broader patient population (RW-E).

ANCOVA model

Consider the outcome of interest is a continuous variable measured at the final (post-randomization) time point, denoted by Y. Let the predictor of interest X be binary taking values $X = 1$ and $X = 0$ for treatment and control, respectively, assigned randomly in a 1:1 fashion. The baseline (pre-randomization) measurement of the outcome assessed at the final time point, denoted by W, is a precision variable in a RCT. In this setting, Z is associated with the outcome of interest Y as it is a measurement of the outcome at baseline. However, W is not associated with treatment assignment X, on average, because of randomization implying that the correlation between W and X is zero (i.e., $r_{WX} = 0$). Hence, W is a precision variable. Accordingly, the ANCOVA model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i$$

where independent $\epsilon_i \sim (0, \sigma^2_{Y|X,W}(\mu_i))$ with $\sigma^2_{Y|X,W}(\mu_i)$ a function of the mean-variance relationship for patients $i = 1, \ldots, n$. The ANCOVA model can also be written as

$$E[Y|X, W] = \beta_0 + \beta_1 X + \beta_2 W.$$

In matrix notation, the ANCOVA model can be written as $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where $\mathbf{X} = [\vec{1}\ \ \vec{X}\ \ \vec{W}]_{n\times 3}$ and $\vec{\beta} = [\beta_0\ \ \beta_1\ \ \beta_2]^T$. The (ordinary) least squares estimate of $\vec{\beta}$ is $\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$ with $Var[\widehat{\vec{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var[\vec{Y}|\vec{X},\vec{W}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$. The variance of the treatment effect is constant if $Var[Y|X,W]$ is constant. When non-constant (i.e., in the presence of a mean-variance relationship) the variance cannot be written in closed form unless the distribution of W is known. Since the distribution of X is fixed by design, we could compute an expectation with respect to X, however, the closed form expression of the variance of treatment effect will change depending on distribution of W.

The estimated treatment effect based on the randomized sample is

$$\hat{\beta}_1 = r_{XY} \cdot \frac{S_Y}{S_X}$$

where the sample correlation between X and Y is r_{XY}, the marginal sample variance of Y is $S_Y^2 \equiv \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2 \equiv \frac{1}{n-1} SS_{YY}$ and, the marginal sample variance of X is $S_X^2 \equiv \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \equiv \frac{1}{n-1} SS_{XX}$. Assuming homoscedasticity (constant $Var[Y|X,W]$), the variance of the estimated treatment effect can be written in closed form as

$$Var[\hat{\beta}_1] = \frac{S_Y^2(1 - r_{XY}^2 - r_{WY}^2)}{SS_{XX}} = \frac{S_Y^2(1 - r_{XY}^2 - r_{WY}^2)}{n \cdot \frac{\phi}{(\phi+1)^2}}$$

where $\phi : 1$ is the randomization allocation. When $\phi = 1$,

$$Var[\hat{\beta}_1] = \frac{S_Y^2(1 - r_{XY}^2 - r_{WY}^2)}{n/4}.$$

Under heteroscedasticity (non-constant $Var[Y|X,W]$), $Var[\hat{\beta}_1]$ does not have a simplified closed form expression without specifying or assuming the distributions for X and W. In

particular, the post-randomization assessment variance conditional on treatment arm is

$$\begin{aligned} Var[Y|X] &= E_W\{Var[Y|X,W]\} + Var_W\{E[Y|X,W]\} \\ &= E_W\{Var[Y|X,W]\} + \beta_2^2 \cdot Var[W] \end{aligned}$$

and and the marginal post-randomization assessment variance is

$$\begin{aligned} Var[Y] &= E_X\{Var[Y|X]\} + Var_X\{E[Y|X]\} \\ &= E_X\left(E_W\{Var[Y|X,W]\} + \beta_2^2 \cdot Var[W]\right) + \beta_1^2 \cdot Var[X] \\ &= E_X\left(E_W\{Var[Y|X,W]\}\right) + \beta_2^2 \cdot Var[W] + \beta_1^2 \cdot Var[X] \\ \implies S_Y^2 &= E_X\left(E_W\{Var[Y|X,W]\}\right) + r_{WY}^2 \cdot S_Y^2 + r_{XY}^2 \cdot S_Y^2 \end{aligned}$$

The estimate of treatment effect, $\hat{\beta}_1$, can be impacted by choice of enrichment strategy and mean-variance relationship. In a RCT, the baseline measurement W is not a potential confounding variable; it is instead a precision variable: associated with the outcome (final measurement) and not with the predictor of interest (randomly assigned treatment group). Furthermore in a RCT, X and as a result S_X are fixed according to the randomization allocation ratio. Therefore, $\hat{\beta}_1$ is a function r_{XY} and S_Y, both of which are functions of the final measurement Y. In particular, since W is associated with Y and W is impacted by enrichment, Y is expected to be impacted by enrichment. Furthermore, S_Y may also be influenced by a mean-variance relationship. This relates to the note above, namely that $Var[Y|X]$ may not be constant with respect to X and $Var[Y|X,W]$ may not be constant with respect to W. That is, $Var[Y]$ depends on the mean-variance relationship for $Var[Y|X,W]$. Figure 2.1 graphically depicts pre-post trajectories for a design with no enrichment (left column) and an enrichment criterion of pre-randomization assessment values among the top 10% of those observed (right column), varying by mean-variance relationship (constant, proportional, and inversely proportional from top to bottom, respectively).

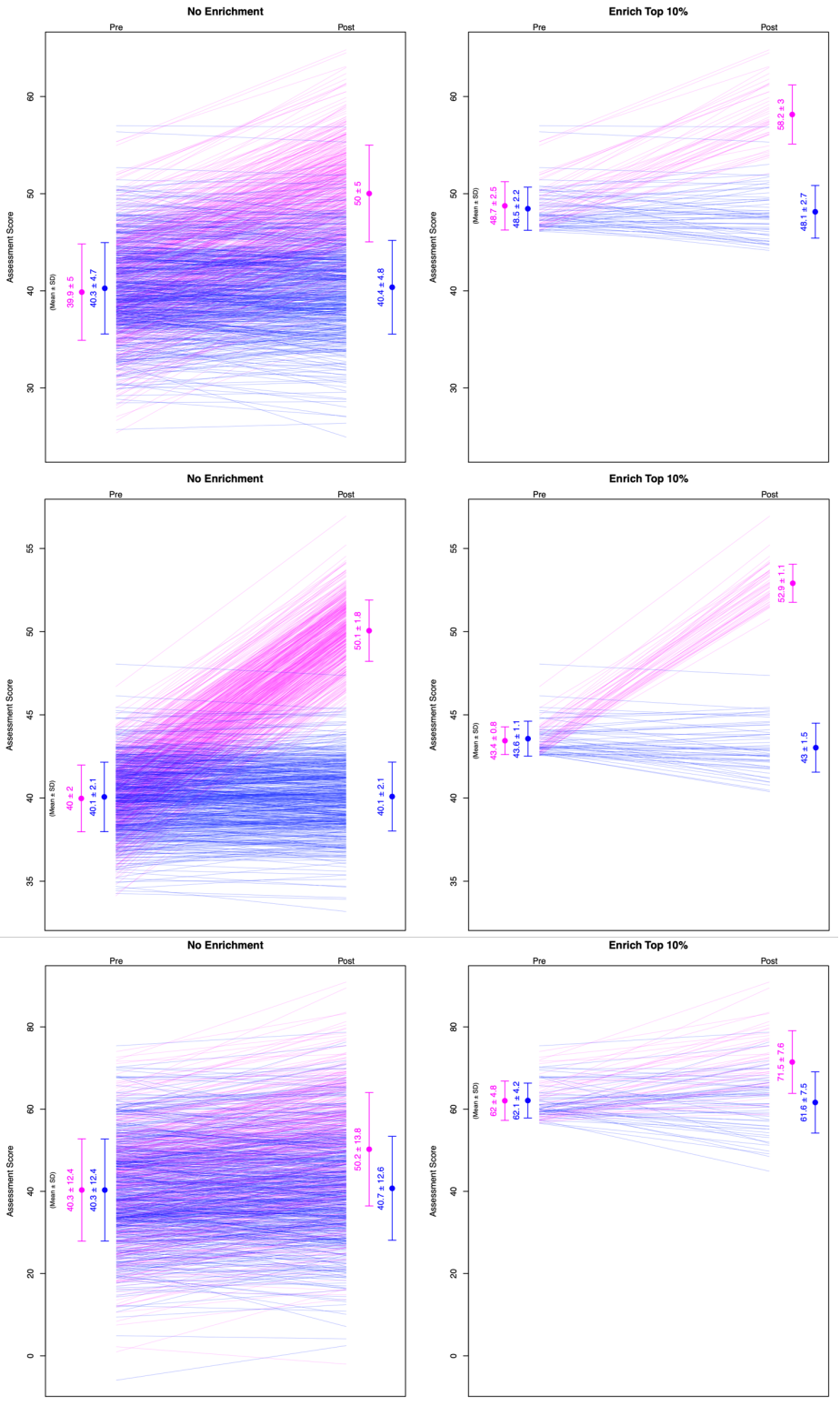


Figure 2.1: Spaghetti plots illustrating pre-post data for treatment (pink) and control (blue). Left column assumes no enrichment and right column assumes enrichment based on the top 10% of pre-randomization scores. Rows correspond to a constant, proportional, and inversely proportional mean-variance relationship, respectively. RW-E $\beta_1 = 10$ for all scenarios.

2.2 Right censored time-to-event outcomes

For RCTs with a time-to-event outcome where participants may be *censored* because they did not have the event of interest during their time on study, often the instantaneous hazard function is the within group measure of treatment effect and the marginal hazard ratio (relative difference in hazards) is the between group comparison of treatment effect. For example, in oncology trials common endpoints (primary, co-primary, or secondary) are overall survival and progression-free survival. Overall survival is defined by all-cause mortality (i.e., the time from randomization to death due to any cause as the event of interest). Progression-free survival is defined by time to death or progression of disease (which may involve additional efforts to determine such as using the Response Evaluation Criteria in Solid Tumours, RECIST (Eisenhauer et al., 2009)), whichever occurs first, as the event of interest.

Let T_i denote the actual time to the event of interest and C_i denote the actual time to censoring for subject i. Then, define $X_i = \min\{T_i, C_i\}$ as the observed time from the origin (in RCTs, this is time of randomization). Let $\delta_i = I(T_i \leq C_i)$ denote the event indicator. Furthermore, let $f_T(t)$ density function of T, $F_T(t)$ cumulative distribution function of T, and $S_T(t) = 1 - F_T(t)$ survival distribution function of T.

The conditional failure rate, also referred to as the hazard function, is defined by

$$\lambda(t) = \lim_{h \downarrow 0} \frac{\Pr[t \leq T < t + h \,|\, T \geq t]}{h}.$$

The multiplicative hazards model with $p \times 1$ covariate vector Z, baseline hazard $\lambda_0(t)$, and $g(\dot{})$ is a relative risk function such that $\lambda(t) = \lambda_0(t)\; g(\,\beta(t)Z(t)\,)$. In the typical RCT setting

without treatment crossover, the Cox proportional hazards model 1972 is

$$\lambda(t) = \lambda_0(t)\ \exp(\ \beta Z\ )$$

where $g(\cdot) = \exp(\cdot)$, $Z(t) = Z\ \{0, 1\}$ and a constant treatment effect $\beta(t) = \beta$ for all t.

Let (X_i, δ_i, Z_i) be independent observations $(i = 1, \ldots, n)$. Then, the likelihood and score functions are

$$L(\beta) = \prod_{i=1}^{n} f_T(X_i|Z_i)^{\delta_i} S_T(X_i|Z_i)^{1-\delta_i}$$
$$U(\beta) = \sum_{i=1}^{n} \left(\delta_i \frac{\lambda'(X_i)}{\lambda(X_i)} - \int_0^{X_i} \lambda'(u)du \right)$$

respectively. The partial likelihood by Cox (1972) yields the estimating equation for β as a stochastic integral

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left(Z_i - \frac{S^{(1)}(t, \beta)}{S^{(0)}(t, \beta)} \right) dN_i(t)$$

where $S^{(r)}(t, \beta) = n^{-1} \sum_{j=1}^{n} Y_j(t) Z_j^r \exp(\beta Z_j)$ and $N_i(t)$ is the number of events in $(0, t)$ for ith individual. Under proportional hazards, Tsiatis (1981) proved the asymptotic theory of the Cox PH regression estimator. Consequently, the asymptotic distribution of the Cox PH estimator is used to determine 'sample size' calculations for the number of events to observe in a time-to-event trial. When assuming a fixed 1:1 randomization,

$$\hat{\beta} \dot{\sim} \mathcal{N}(\beta, 4/D)$$

where D is the total number of events.

Model misspecification: Non-proportional hazards

Proportional hazards is a key assumption for valid interpretation of the Cox model. On the other hand, non-proportional hazards is a common phenomenon. In clinical trials, non-proportional hazards can arise in a variety of ways. One such example of non-proportional hazards is with a surgical intervention vs. non-surgical intervention. In such a setting it may not be too surprising that patients who have a surgical procedure tend to have a higher hazard of dying (or lower probability of survival) during and immediately following the surgery due to possible complications compared to patients who received a non-surgical intervention, but for those who survive past this time, the hazard may lower for those who had the surgery compared to the hazard for those who did not and have a worse prognosis at a later time.

In the presence of time-varying treatment effects (e.g. non-proportional hazards), additional care must be taken regarding the scientific question of interest and what the target of inference and target population are for inference. This relates to the concept of an *estimand* that has received explicit attention since 2017 with the initial development of the International Conference on Harmonisation (ICH) E9 (R1) Addendum on estimands and sensitivity analyses (U.S. Food and Drug Administration, 2021). As noted previously, to define an estimand one needs to specify the target population of interest and the quantity of interest to estimate (i.e., the between-group summary measure for a pre-specified analysis population).

Reliance upon the proportional hazards assumption to obtain a consistent estimate of the hazard ratio (HR) and valid inference can be a problem in settings where it is known *a priori* or *a posteriori* that the hazards are not proportional (i.e., non-proportional hazards). The basis for such investigations stems from Struthers and Kalbfleisch (1986) where they showed the solution to the partial likelihood score equation (solving for the parameter β that represents the marginal hazard ratio) that yields the Cox PH estimator depends on

the censoring distribution. This β corresponds to the solution of the following estimating equation

$$\int_0^\infty E_Z \left\{ f_T(t|Z) S_C(t|Z) \left[Z - \frac{E_Z \left\{ Z S_T(t|Z) S_C(t|Z) e^{\beta Z} \right\}}{E_Z \left\{ S_T(t|Z) S_C(t|Z) e^{\beta Z} \right\}} \right] \right\} dt = 0$$

that depends upon censoring patterns, $S_C(t|Z)$.

Since the censoring patterns are not typically of scientific interest, and can be study design induced, an estimand that depends on a trial's censoring patterns can inhibit replicability and comparisons across trials. While Xu and O'Quigley (2000) and van Houwelingen et al. (2005) offer solutions under non-proportional hazards, their approaches that involve re-weighting the hazard ratio estimate to a standard censoring distribution requires independent censoring.

Differential censoring patterns, however, come up in practice. Two such classifications include intentional design-based strategies and unintended treatment effects (Boyd et al., 2012). Within intentional design-based strategies, differential censoring can be induced through the use of historical data such as from historical controls, as the authors illustrate with an example from a trial on brain metastases. Other design-based strategies that induce differential censoring include levels of stratification variables added during course of trial, new sites or regions added (or starting up at different times) during trial, and outcome adaptive randomization.

Boyd et al. (2012) focus on possible differences in the censoring distributions based on a single binary predictor of interest, randomized treatment assignment (treatment or control), in a fixed sample size RCT. Considering ways in which the censoring patterns can differ by treatment groups motivates the need to reassess a common assumption when using survival analysis: independent censoring. They note that the requirement for independent censoring (i.e., the time to censoring C is independent of time to event T) if two criteria are met: (i) a functional of the distribution of T does not require specifying C; and (ii) the distribution of

the binary predictor of interest Z is independent of the distribution of C. Another commonly assumed form of censoring is conditionally independent censoring (CIC) where C and T are independent conditional on covariate Z.

Under independent censoring, $S_C(t|Z) = S_C(t)$,

$$\int_0^\infty E\left\{ f_T(t|Z) S_C(t) \left[Z - \frac{E\left\{ Z S_T(t|Z) e^{\beta Z} \right\}}{E\left\{ S_T(t|Z) e^{\beta Z} \right\}} \right] \right\} dt = 0$$

$\hat{\beta}_{PH}$ depends on $S_C(t)$ as a weighted average. This, however, cannot be simplified under CIC. Hence, under model misspecification (i.e., non-proportional hazards) $\hat{\beta}_{PH}$ depends on the censoring distribution. Similarly, the log-rank test and weighted log-rank statistics also depend on the censoring distribution (Gillen, 2003; Gillen and Emerson, 2005). To remove the dependence on the censoring distribution for the estimate and interpretation of the marginal hazard ratio, Boyd et al. (2012) defined the weight function for subject j at time t

$$W_j(t) = \left\{ S_C(t|Z_j) \right\}^{-1}.$$

Then, the re-weighted estimating equation is

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty W_i(t) \left(Z_i - \frac{S_W^{(1)}(t,\beta)}{S_W^{(0)}(t,\beta)} \right) dN_i(t) = 0$$

where

$$S_W^{(r)}(t,\beta) = n^{-1} \sum_{j=1}^{n} W_j(t) Y_j(t) Z_j^r \exp(\beta Z_j),$$
$$W_j(t) = \left\{ \hat{S}_C^{KM}(t|Z_j) \right\}^{-1},$$

and $\hat{S}_C^{KM}(t|Z_j)$ is the left-continuous version of Kaplan and Meier (1958) estimate of censoring distribution. The solution to the re-weighted estimating equation yields $\hat{\beta}_{CIC}$ that is

consistent for β_{CIC} defined as the solution to

$$\int_0^\infty E\left\{ f_T(t|Z)\left[Z - \frac{E\left\{ Z S_T(t|Z) e^{\beta Z} \right\}}{E\left\{ S_T(t|Z) e^{\beta Z} \right\}} \right] \right\} dt = 0$$

which also works under IC and is the same as $\hat{\beta}_{IC}$ (Xu and O'Quigley, 2000) where the weight function from the combined censoring distribution (common weight) is $W(t) = \{S_C(t)\}^{-1}$.

To this end, the estimand framework is useful to critically think about the quantity to estimate and the methodology that will yield the desired estimates. This approach relates to the time-to-event setting, specifically in the presence of non-proportional hazards. As already mentioned, pre-specification of the analytic plan is required for clinical trials. Censoring-dependent estimation procedures can make comparisons of results across even similarly designed trials difficult. Hence, to protect against the issues arising from time-varying treatment effects, targeting a marginal hazard ratio standardized to a common censoring distribution, referred to as *censoring-robust estimation*, seems an appropriate step to obtain a reliable estimate of the RCT-E and enhance replicability across trials.

2.3 Group sequential designs

A fixed sample design RCT consists of a single pre-specified primary outcome used in the pre-specified primary analysis to answer the pre-specified primary question of scientific interest. Limiting the number of ineffective interventions approved is consistent with the goals in medicine and public health. Pre-specification (ideally, before the trial begins or at least before looking at the data) of the primary scientific question (aim) of interest and correspondingly the primary endpoint and primary analysis helps to control the false positive rate (type I error rate). Typical outcomes in fixed RCTs, measured at a single time point or longitudinally, include: binary (e.g., relapsed with first 5 years after completion of cancer treatment); count

(e.g., number of metastatic lesions); continuous (e.g., change from baseline in a quality of life assessment score); or (censored) time-to-event (e.g., overall survival, an objective measure of time to death due to any cause).

Duration of a fixed sample design RCT varies depending on the disease, intervention, purpose of the trial, including the target population and availability (or lack thereof) of existing interventions for treatment or prevention. Accordingly, in certain settings ethical and feasibility reasons may warrant formal interim analyses of the primary endpoint before observing the primary endpoint for all pre-planned study participants. A formal interim analysis based on partial information prior to a pre-planned primary analysis on full information requires use of *sequential testing* due to multiple comparisons (multiple looks of the efficacy data). Not accounting for sequential testing will inflate the type I error rate if at least two analyses are conducted (e.g., interim analysis and final (full) analysis).

A *group sequential design (GSD)* (Jennison and Turnbull, 1999) allows for modification of the RCT design by incorporating pre-specified interim analyses at pre-determined time points (e.g., proportion of information observed). Consequently, the sampling distribution of a chosen test statistic (based on the estimand of scientific interest pertaining to the pre-specified primary aim for the trial) changes for a GSD compared to a fixed design RCT. Emerson et al. (2007) discuss how the sampling distribution of test statistics are not simply a location shift between null and alternative hypotheses because of the possibility of stopping the trial "early" due to (overwhelmingly) adequate and reliable evidence of a clinically meaningful favorable benefit-to-risk ratio.

Typically in a fixed sample RCT or GSD enrollment is staggered. That is, not all patients are enrolled and randomized at the start of the trial so patients have the same time zero but different calendar start times. Consequently, it is possible and sometimes desirable for ethical reasons to have at least one interim analysis before the trial completes enrollment. If an (early) potential safety signal appears or a conduct issue arises, the DMC may recom-

mend (in a blinded manner) to the sponsor to temporarily suspend enrollment or make other modifications to protect the interests of the human volunteers and trial integrity. Emerson et al. (2007) define the group sequential test statistic with respect to a chosen stopping rule (also called a monitoring guideline or group sequential boundary) as a pair of two quantities: the first (interim) analysis time when the estimated partial sum statistic "crosses" the boundary value at a given time (denoted by M) and the corresponding partial sum statistic at M (denoted by S). Then, the sampling density of (M, S), assuming independent increments of information between times $j-1$ and j, is defined recursively (Armitage et al., 1969). They also note that the sampling density is dependent on the choice of group sequential boundaries; two common boundaries include O'Brien and Fleming (1979) and Pocock (1977).

Partial sum statistic

Let $X_1, \ldots, X_{N_J} \sim_{iid} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known and μ unknown. Consider analyses conducted when $N_1, \ldots, N_J$ participants have the outcome of interest recorded. Note that $N_1 < \ldots < N_J$. Furthermore, define $n_1 = N_1$ and $n_j = N_j - N_{j-1}$ for $j = 2, \ldots, J$. Then the partial sum statistic is

$$S_j = \sum_{i=1}^{N_j} X_i \ \sim \ \mathcal{N}(N_j\mu, N_j\sigma^2)$$

where $Cov(S_j, S_{j+1}) = N_j\sigma^2$ and $Cov(S_j, S_{j+1} - S_j) = Cov(S_j, S_{j+1}) - Var(S_j) = N_j\sigma^2 - N_j\sigma^2 = 0$ (independent increments structure).

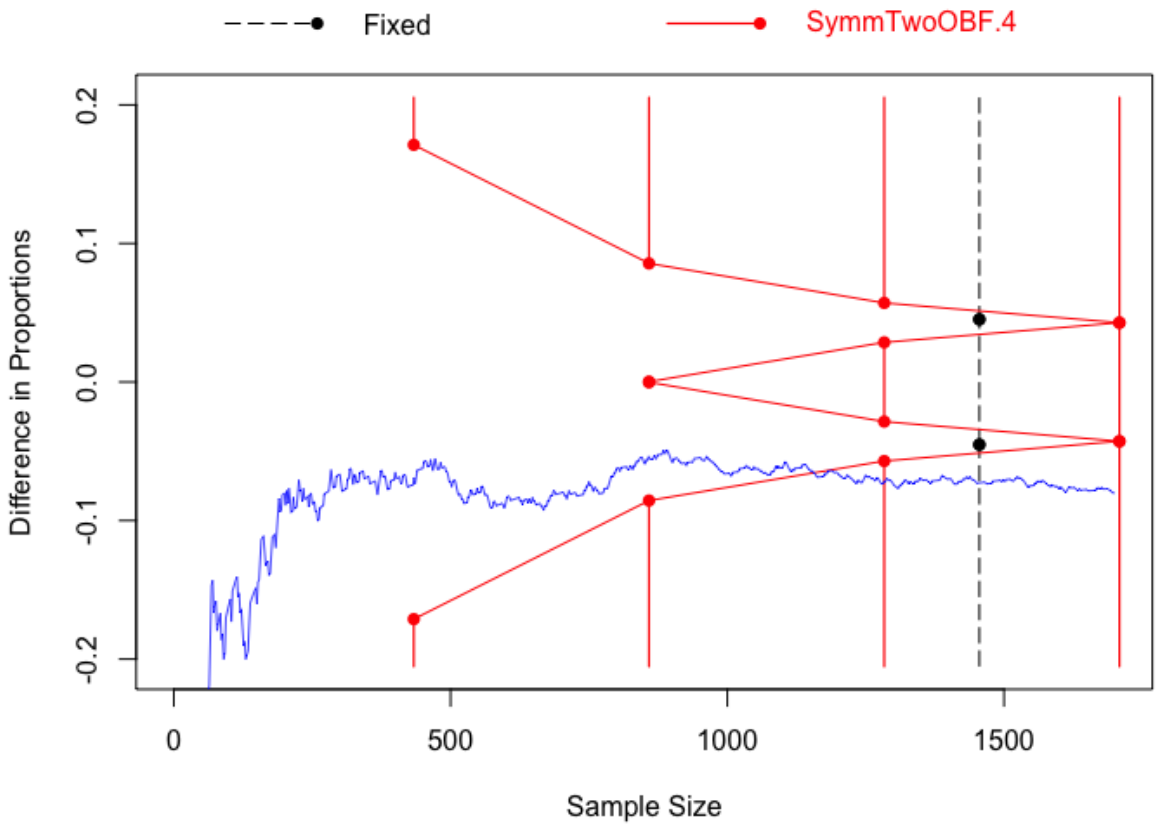


Figure 2.2: Sample path of the estimated difference in binomial proportions (blue line) with the fixed sample RCT (black vertical boundary) and a symmetric two-sided O'Brien-Fleming GSD with up to four total analyses (red vertical boundaries) overlaid.

Continuation set

Define $\mathcal{C}_{S_j} = (a_{S_j}, b_{S_j}] \cup [c_{S_j}, d_{S_j})$ such that the trial continues to the next analysis time if $S_j \in \mathcal{C}_{S_j}$, and stops after the analysis time j for which $S_j \notin \mathcal{C}_{S_j}$. Figure 2.2 illustrates an example observed sample path of a statistic in a fixed sample design and a GSD.

Group sequential test statistic (M, S)

In a group sequential design (Jennison and Turnbull, 1999), the test statistic is bivariate, consisting of $M = \min\left\{1 \leq j \leq J \ : \ S_j \notin \mathcal{C}_{S_j}\right\}$, the analysis at which the study stopped, and $S = S_M$ is the statistic at the Mth analysis. Note that Chang (1989) showed that $(M = m, S = s)$ is sufficient statistic for unknown normal mean μ.

Sequential sampling density

Armitage et al. (1969), under independent increments, recursively defined the sequential sampling density for $(M = m, S = s_m)$ by

$$p(m, s; \mu) = \begin{cases} f(m, s; \mu) & \text{if } s \notin \mathcal{C}_{S_m} \\ 0 & \text{otherwise} \end{cases}$$

where

$$f(1, s; \mu) = \frac{1}{\sqrt{n_1}\sigma} \ \phi\left(\frac{s - n_1\mu}{\sqrt{n_1}\sigma}\right)$$

$$f(j, s; \mu) = \int_{\mathcal{C}_{S_{j-1}}} \frac{1}{\sqrt{n_j}\sigma} \ \phi\left(\frac{s - u - n_j\mu}{\sqrt{n_j}\sigma}\right) \ f(j-1, u; \mu) du$$

where $n_j = N_j - N_{j-1}$ for $j = 2, \ldots, m$ and $\phi(\cdot)$ is the standard normal density function. Emerson and Fleming (1990) noted the following identity

$$f(j, s; \mu) = f(j, s; 0) \ \exp\left(\frac{s\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} N_j\right)$$

allows one to easily compute the sampling density under different values of μ after computing for $\mu = 0$, thus computing confidence intervals without much computational effort. Figure

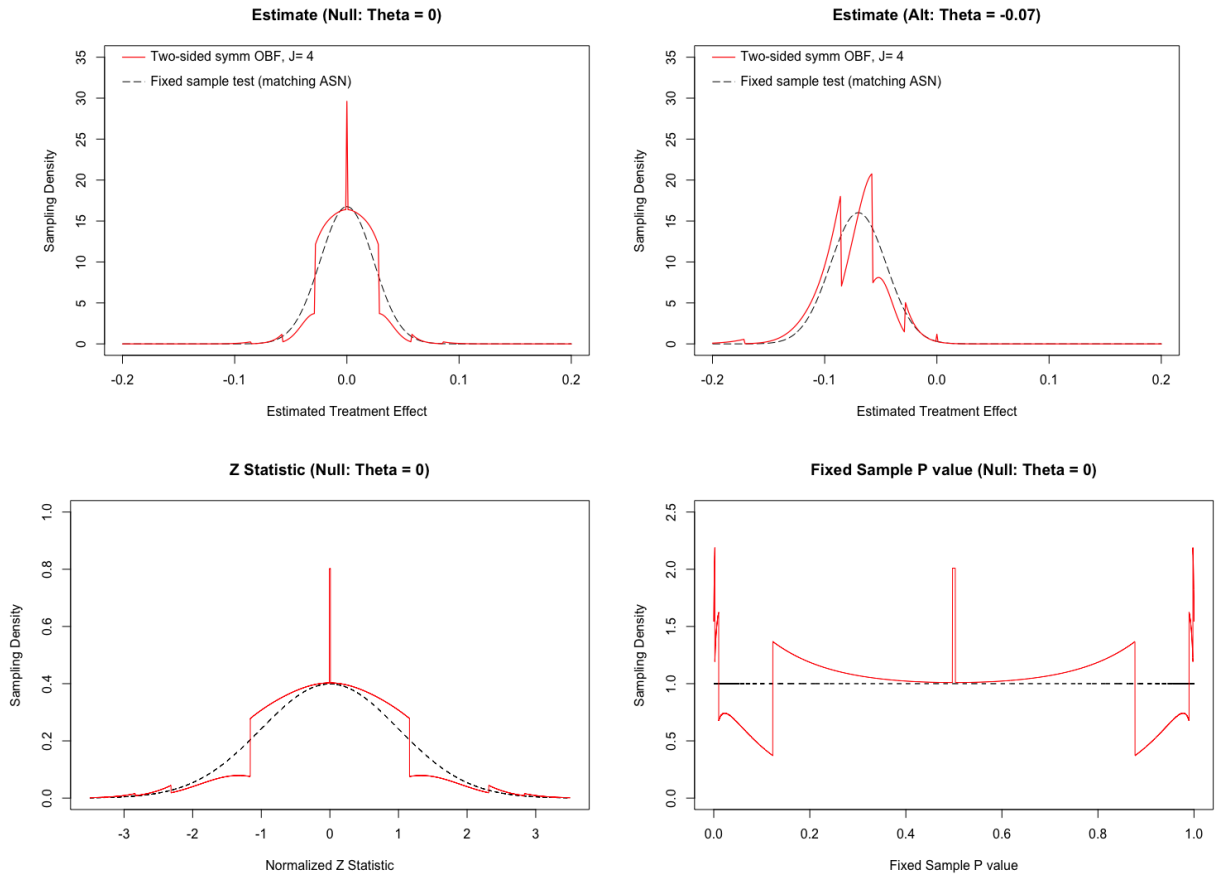


Figure 2.3: Sequential sampling density under a null for the estimated treatment effect (top left), normalized Z statistic (bottom left), and fixed sample P value (bottom right), and under an alternative (top right).

2.3 illustrates how the sequential sampling density is not normally distributed under the null (neither on the estimated treatment effect scale nor the normalized Z statistic scale) nor under the alternative. Furthermore, in the sequential testing setting, the distribution of the fixed sample P values under the null is no longer uniformly distributed on $(0, 1)$ as they are in the fixed sample size testing setting.

Repeated significance tests alters the sampling distributions of statistics, including those that in a fixed sample setting would be approximately normally distributed. Such sequential densities have a form that can be expressed via recursion (Armitage, McPherson, and Rowe, 1969). Specialized software is needed to compute the sequential sampling density $p(m, s_m; \theta)$.

Then, integrating over the sequential sampling density allows computation of confidence intervals and p-values, and statistical operating characteristics to evaluate group sequential designs.

Group sequential boundaries and timing of analyses

For a group sequential design, the timing of interim analyses ($j = 1, ..., J-1$) and the planned final analysis J as well as the boundaries are based on the proportion of maximal information at the j-th analysis, $\Pi_j = \mathcal{I}_j/\mathcal{I}_J$. Kittelson and Emerson (1999) constructed the unified family of group sequential designs that encompasses many commonly known and used designs, as well as an infinite number of designs that may warrant consideration for a specific scientific and clinical context. In their specification, let $*$ represent one of the a, b, c, or d boundaries described previously. Then, the boundary function is

$$f_*(\theta_*, g(\Pi_j; A_*, P_*, R_*, G_*))$$

with boundary shape function

$$g(\Pi_j; A_*, P_*, R_*, G_*) = \left\{A_* + \Pi_j^{-P_*}(1-\Pi_j)^{R_*}\right\} G_*$$

where A, P, and R are user-specified and G is calculated to attain certain operating characteristics. This parameterization of group sequential designs includes

- Pocock (1977): $A = 0$, $R = 0$, $P = 0.5$ (constant on the Z scale)
- O'Brien-Fleming (1979): $A = 0$, $R = 0$, $P = 1$ (constant on the partial sum scale)
- Whitehead-Stratton triangular (1983): $A = 1$, $R = 0$, $P = 1$
- Wang-Tsiatis (1987): $A = 0$, $R = 0$, any P

Furthermore, Emerson et al. (2007) noted that group sequential boundaries can be expressed on different scales as there is a 1-to-1 mapping between the partial sum statistics, crude estimate of treatment effect, normalized Z statistic, fixed sample P-value statistic, error spending statistic, Bayesian posterior probabilities, conditional power statistics, and predictive power statistics scales. Inspite of this relationship across the scales, the ease by which these scales map to a clinically meaningful interpretation varies. Using the treatment effect scale (e.g., hazard ratio with a time-to-event endpoint) can aid in scrutinizing candidate group sequential designs and boundaries, including the timing of the analyses.

For time-to-event group sequential trials, when proportional hazards holds, the proportion of maximal information is equivalent to the proportion of maximal events. Recall that under proportional hazards, the Cox PH estimator $\hat{\beta}_{PH} \dot{\sim} \mathcal{N}(\beta, 4/D)$ where D is the total number of events. This implies that the statistical information of the corresponding score statistic (i.e., the logrank test) is proportional to the number of events (i.e., linear information growth). Hence, many times in practice, monitoring of time-to-event group sequential trials are based on the number of events observed (or when they are projected to occur). However, under non-proportional hazards, Gillen and Emerson (2005) showed that under non-proportional hazards, the information growth of weighted logrank statistics is non-linear with their dependence on the censoring distribution as the primary contributor. This will prove an important intuition for characterizing the information growth of censoring-robust estimators (which we noted previously are weighted statistics) in Chapter 5. For statistics with a non-linear information growth, regardless of whether proportional hazards holds or not, we cannot base the timing of interim analyses naively assuming $\Pi_j \propto D_j/D$ where D_j is the number of events observed by analysis j and D is the maximal number of events to observe during the trial. Furthermore, suppose the study planning stage begins with a GSD requiring D maximal events under proportional hazards. If, through collaborative discussions, a time-varying treatment effect is possible, a censoring-robust estimation may be warranted for the pre-specified primary analysis. In a scenario where proportional hazards

approximately holds a censoring-robust estimator will be less efficient. In this case, at the study planning stage, consideration for observing a $D* > D$ may be necessary to maintain trial operating characteristics (e.g., statistical power to detect the design alternative) and thus valid statistical inference.

Chapter 3

On the Use of Enrichment in Fixed Sample Pre-Post Randomized Trials

SUMMARY: Enrichment in a pre-post randomized clinical trial (RCT) often consists of an inclusion criterion on a pre-randomization assessment (e.g., biomarker or surrogate outcome). The target of inference is commonly the difference in mean change from baseline comparing treatment to control for the RCT target population: the RCT estimand (RCT-E). In clinical practice, however, health care providers may prescribe an approved drug on- or off-label to patients who belong to a broader real-world (RW) target population, requiring the RW estimand (RW-E). Here, we quantify the impact of enrichment in pre-post RCTs when estimating the RW-E from the RCT sample. Specifically, we show that regression to the mean can induce a biased estimator for the RW-E. We analytically derive this bias term under normality and a heteroscedastic mean-variance relationship. We propose a bias-adjusted estimator for the RW-E and establish its operating characteristics via Monte Carlo simulation.

3.1 Introduction

A pre-post randomized clinical trial (RCT) typically consists of two correlated continuous outcome assessments: one pre-randomization and one post-randomization. Often, the within group summary measure is the mean change from baseline (post – pre) that we denote by

$$\Delta_k = \mu_{2k} - \mu_{1k}$$

where μ_{jk} is the mean outcome at time j (1 for pre, 2 for post) and treatment arm k (1 for treatment, 0 for control). Then, the target of inference is the difference in mean change from baseline for a subpopulation of individuals randomly assigned to the treatment arm and mean change from baseline for a subpopulation of individuals randomly assigned to the control arm. We denote this target of inference by

$$\theta = \Delta_1 - \Delta_0 = \mu_{21} - \mu_{20}. \tag{3.1}$$

The latter equality in (3.1) holds because $\mu_{11} = \mu_{10}$, on average, from randomizing treatment assignment.

The RCT target population corresponds to individuals who would meet the specific trial's eligibility criteria. Together, a target of inference and a corresponding target population are key attributes of an *estimand*. Formally, ICH E9 R1 Addendum (U.S. Food and Drug Administration, 2021) requires specifying the following attributes to define an estimand:

(i) the treatment conditions;

(ii) the target patient population of interest;

(iii) the endpoint (or variable); and

(iv) handling of intercurrent events.

In this chapter, θ is a trial-specific estimand for a population-level comparison, assuming no intercurrent events, that we refer to as the RCT estimand (RCT-E) for a pre-post design.

There are a number of analysis methods to estimate the RCT-E θ in the pre-post setting (Liang and Zeger, 2000; Yang and Tsiatis, 2001; Senn, 2006; O'Connell et al., 2017; Wan, 2021). Three common approaches are the two-sample t-test, analysis of covariance (ANCOVA), and paired change. In the pre-post RCT setting, assuming no missing data, these three methods yield consistent and unbiased estimates of θ. The ANCOVA model, however, is often preferred when the correlation between pre and post assessments is at least 0.5 because it is more efficient (Feldt, 1958) (i.e., yields smaller standard error estimates corresponding to smaller confidence interval widths and higher statistical power).

In an effort to accelerate the drug development process, enrichment strategies based on demographic information, biomarkers, or a surrogate outcome are employed to identify individuals likely to benefit from the candidate treatment (prognostic) or to identify individuals likely to have the outcome of interest during the trial (predictive) (U.S. Food and Drug Administration, 2019b) that may result in smaller and shorter trials. The U.S. Food and Drug Administration guidance document on enrichment strategies (2019b) summarizes the relevance and commonality of enrichment strategies in an array of disease areas.

An enrichment strategy considered in the pre-post RCT setting is an added trial inclusion criterion in which otherwise eligible individuals must meet some threshold based on a biomarker or (surrogate) outcome assessment. For example, in the University of California Cures Nicotinamide as an Early Alzheimer's disease Treatment (NEAT) phase 2 proof-of-concept pre-post RCT (2021, ClinicalTrials.gov identifier NCT03061474), individuals were randomized to nicotinamide or placebo if they met biomarker criteria based on cerebral spinal fluid amyloid beta 1-42 not exceeding 600 pg/mL or a ratio of total tau to amyloid

beta 1-42 of at least 0.39. Only individuals meeting this enrichment criterion and the rest of the trial's inclusion/exclusion criteria were randomized. The primary endpoint was change in cerebrospinal fluid phosphorylated tau (p-tau231) 12 months post-randomization, the target of inference was the difference in these pre-post changes between nicotinamide and placebo arms, and the RCT-E target population most directly corresponded to individuals with mild Alzheimer's disease or mild cognitive impairment due to Alzheimer's disease who met the trial eligibility criteria, including the enrichment criterion.

Often, we use clinical trials to determine whether a candidate treatment is causally associated with a favorable benefit-to-risk ratio for a pre-specified target population of interest. Generalizability of results beyond the enrolled RCT sample is important to have utility in clinical practice. While an approved drug has a particular indication on the drug label that corresponds most often to the RCT-E, there can be potential for off-label use from a broader population once the drug is available on the market. Healthcare providers may elect to prescribe a drug off-label in keeping to their Hippocratic Oath. In such instances, the inference on the RCT-E may not provide adequate information for a provider to make an informed decision. Instead, a real-world estimand (RW-E) may be of importance. While we contend that there can be many types of RW-Es, in this chapter, we restrict our attention to a RW-E β_1 that corresponds to patients who may or may not have met the pre-randomization enrichment cutoff, but who otherwise met the eligibility criteria of the trial.

Figure 3.1 illustrates a number of potential target populations among the population of potential users of an intervention (e.g., drug or biologic) where 'eligible' refers to those meeting all inclusion/exclusion criteria without regard for any enrichment criterion. In this chapter, the RCT-E will correspond to the solid gold-colored 'Screened / Eligible / Enrich' population of potential users whereas the RW-E will correspond to both the solid gold-colored and thick patterned blue-colored boxes, 'Screened / Eligible' (regardless of meeting the enrichment criterion). These RCT-E and RW-E target populations would also correspond

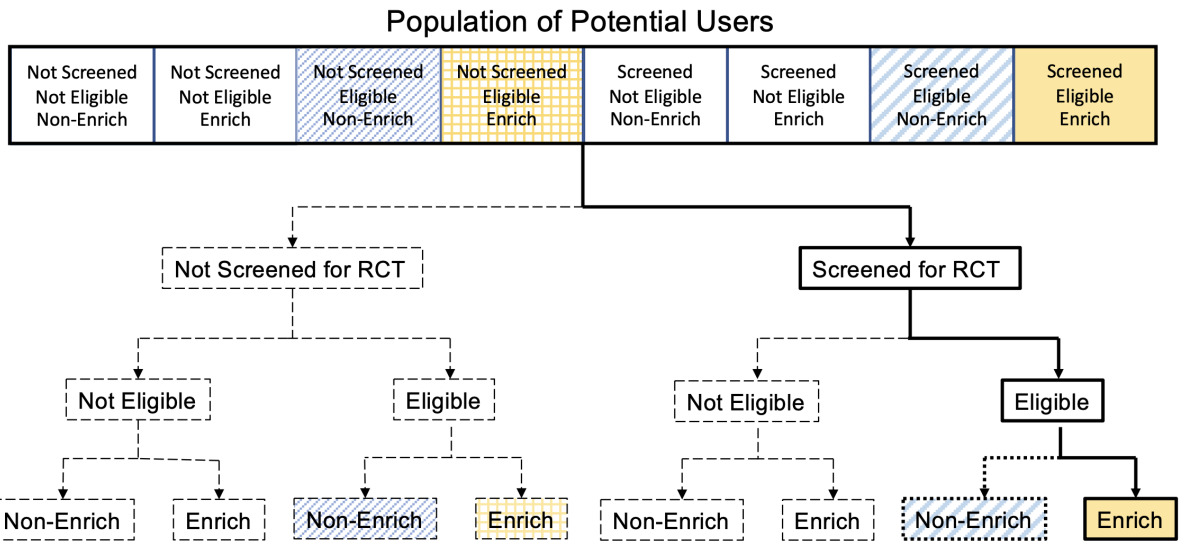


Figure 3.1: A breakdown of the population of potential users partitioned into 8 potential target populations where eligible means having met all trial inclusion/exclusion criteria without regard for an enrichment criterion and enrich means met the enrichment criterion.

to the most direct target population for the UC Cures NEAT enriched pre-post RCT and a potential broader population of population users, respectively.

Interested in inference for a broader target population, we conjectured that RCT-E θ and RW-E β_1 may differ. To our knowledge, there is a gap in understanding whether there are differences in these estimands and when for enriched pre-post RCTs. Hence, in this chapter, we quantified the impact of enrichment in pre-post RCTs on the estimation of the RW-E (that we will denote by β_1) when using an estimate of the RCT-E, denoted by $\hat{\theta}$, as a proxy measure. In Section 3.2 we show, based on analytical derivation, that bias of the RCT-E estimator with respect to the RW-E arises as a consequence of regression to the mean for normally distributed pre-post data with a heteroscedastic mean-variance relationship. We then propose a bias-adjusted estimator for the RW-E as a function of the RCT-E estimator and quantities from the RW target population. In Section 3.3 we present Monte Carlo simulation studies examining operating characteristics for our proposed estimator when all components are known. We conclude with a discussion of our findings in the context of trial design and generalizing beyond just the RCT sample in Section 3.4.

3.2 Methods

3.2.1 Notation

Consider a pre-post RCT in which N individuals are screened, and of those there are n participants randomized to one of two treatments (treatment or control). Randomized participants will complete two assessments, either the same or that are (highly) correlated with each other (pre at baseline and post at some *a priori* determined clinically meaningful time point). Suppose there are n_1 participants randomly assigned to treatment and n_0 to placebo (i.e., $n = n_1 + n_0$). Let Y_{ijk} denote an assessment for participant i ($i = 1, ..., n_k$), at time j ($j = 1, 2$ where 1=pre, 2=post) who was assigned to intervention k ($k = 0, 1$ where 0=control, 1=treatment). Let X_i denote an indicator for whether participant i was randomized to treatment ($X_i = 1$) or placebo ($X_i = 0$). Let $\vec{Y}_2$ denote a vector of the post-randomization outcome assessments for all randomized participants. Let $\vec{Y}_1$ denote a covariate vector for the baseline (pre-randomization) assessments for all randomized participants. Let $\vec{X}$ denote the treatment indicator vector for all randomized participants.

Let θ continue to denote the trial-specific estimand, RCT-E, and β_1 to denote a broader patient population estimand, RW-E, that may include individuals who receive drug off label. Recall that the three common linear models used to estimate RCT-E θ from the enriched pre-post RCT sample using the current notation for pre and post assessments are

- Two-sample t-test: $E[\,\vec{Y}_2 \,|\, \vec{X}\,] = \beta_0^{(t)} + \theta\,\vec{X}$
- Analysis of covariance (ANCOVA): $E[\,\vec{Y}_2 \,|\, \vec{X}, \vec{Y}_1\,] = \beta_0^{(a)} + \theta\,\vec{X} + \beta_2^{(a)}\,\vec{Y}_1$
- Paired change: $E[\,(\vec{Y}_2 - \vec{Y}_1) \,|\, \vec{X}\,] = \beta_0^{(p)} + \theta\,\vec{X}$

To distinguish between the enriched RCT sample and a sample from a broader, real world patient population, here we denote corresponding 'real world' quantities by a superscripted

asterisk (except for the already defined RW-E β_1). Then, the three linear models will be expressed as

- Two-sample t-test: $E[\,\vec{Y}_2^* \,|\, \vec{X}^*\,] = \beta_0^{*(t)} + \beta_1 \vec{X}$
- Analysis of covariance (ANCOVA): $E[\,\vec{Y}_2^* \,|\, \vec{X}^*, \vec{Y}_1^*\,] = \beta_0^{(*a)} + \beta_1 \vec{X}^* + \beta_2^{(*a)}\, \vec{Y}_1^*$
- Paired change: $E[\,(\vec{Y}_2^* - \vec{Y}_1^*) \,|\, \vec{X}^*\,] = \beta_0^{(*p)} + \beta_1 \vec{X}^*$

In practice, when employing an enriched pre-post RCT design, post-randomization information is not collected for individuals who fail to meet the enrichment criterion. Here, we assume all screened individuals have pre-randomization assessments (one or more depending on the the study design) collected, and all among those meeting the enrichment criterion have post-randomization assessments collected. To this end, in this chapter we consider estimation of the RW-E β_1 based upon the RCT-E estimator $\hat{\theta}$, and restrict attention to superiority trials with enrichment based on pre assessments. Here, we focus attention on testing for efficacy. For our setup, this corresponds to testing the null hypothesis $H_0 : \beta_1 = 0$ versus the one-sided (greater) alternative hypothesis $H_A : \beta_1 > 0$, and motivation for focusing on estimation for $\beta_1 > 0$.

3.2.2 Analytic form of the bias resulting from an enriched pre-post RCT design

Consider a pre-post randomized controlled clinical trial (RCT) design. Define the RCT-E θ to be the difference in mean change from baseline comparing those randomized in the trial to treatment versus control. Suppose the real world estimand (RW-E) β_1 corresponds to the same contrast for a broader patient population. Let $\vec{Y}_{ik}$ denote a 2×1 vector of pre ($j = 1$) and post ($j = 2$) assessments for subject i ($i = 1, ..., n_k$) assigned to treatment k ($k = 0, 1$)

and

$$\begin{pmatrix} Y_{i1k} \\ Y_{i2k} \end{pmatrix} \sim \mathcal{N}_2 \left(\vec{\mu}_k \equiv \begin{pmatrix} \mu_{1k} \\ \mu_{2k} \end{pmatrix}, \Sigma_{jk} \equiv \begin{pmatrix} v_{1k} & \rho_k \sqrt{v_{1k} v_{2k}} \\ \rho_k \sqrt{v_{1k} v_{2k}} & v_{2k} \end{pmatrix} \right)$$

where $\mu_{1k} = \beta_0 + \gamma$ for $\beta, \gamma \in (-\infty, \infty)$, $\mu_{2k} = \beta_0 + \beta_1 \cdot \mathbf{1}_{[k=1]}$, and $v_{jk} = f(\mu_{jk}) = \sigma^2 \cdot |\mu_{jk}|^\delta$ with $\sigma \in (0, \infty)$ and where $\delta < 0$ corresponds to an inversely proportional, $\delta = 0$ to a constant, and $\delta > 0$ to a proportional mean-variance relationship. Proposition 3.1 describes the form of the bias of $\hat{\theta}$ with respect to β_1.

Proposition 3.1: *For a pre-post designed RCT with enrichment, define β_1 as the real world estimand (RW-E) and θ as the RCT estimand (RCT-E). Then, the RW-E can be decomposed in terms of the RCT-E and a bias term*

$$\beta_1 = \theta + \left(\rho_1 \sqrt{v_{21}} \cdot \frac{\int_{-\infty}^{c_{11}^*} z \cdot \phi(z) dz}{1 - \Phi\left(c_{11}^*\right)} - \rho_0 \sqrt{v_{20}} \cdot \frac{\int_{-\infty}^{c_{10}^*} z \cdot \phi(z) dz}{1 - \Phi\left(c_{10}^*\right)} \right)$$

where $c_{1k}^* = (c - \mu_{1k})/\sqrt{v_{1k}}$, the enrichment cutoff based on pre assessments $c = \Phi^{-1}(p_{enrich})$ with p_{enrich} denoting the enrichment proportion, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density function and distribution function, respectively. (See Appendix A.1 for a derivation.)

From Proposition 3.1, the bias term is zero when: (i) there is no enrichment (i.e., $c = -\infty$ implying that $\psi(-\infty, \rho, \mu_1, v_1) = 0$) irrespective of the form of the mean-variance relationship; (ii) the pre-post assessments are uncorrelated within each treatment arm (i.e., $\rho_1 = \rho_0 = 0$); or (iii) there is enrichment based on the pre-randomization assessment and the product of the pre-post correlation and post variance is equal for the two treatment arms (i.e., $\rho_1 \sqrt{v_{21}} = \rho_0 \sqrt{v_{20}}$).

With the derived analytic form of the bias in a pre-post RCT when assuming normally distributed data, in Figure 3.2 we characterize the analytic bias as a function of the enrichment cutoff criterion (i.e., value of the probability corresponding to the quantile for the

pre-randomization assessment cutoff) in two ways. The plot on the left is meant to examine how the correlation between pre and post assessments changes the bias given a particular mean-variance relationship. From the analytic form of the bias in the Methods section, we know that a constant mean-variance relationship and the same pre-post correlation for both treatment arms will zero out the bias term, irrespective of the value of the correlation. Notably, for a proportional mean-variance relationship based on $\delta = 0.5$ (i.e., the variance is a square root of the mean, up to a fixed scaling term, which we denoted earlier by σ). The plot on the right illustrates how the magnitude of the bias in the proportional mean-variance setting is markedly higher for larger values of δ. Highly (positively) correlated pre-post assessments yield larger bias, as illustrated in the plot on the left.

3.2.3 Bias-adjusted estimator for the RW-E

We propose the following bias-adjusted estimator for the RW-E β_1

$$\hat{\beta}_1^{\text{BiasAdj}} = \hat{\theta}_{\text{RCT-E}} + \widehat{\text{Bias}}[\,\hat{\theta}_{\text{RCT-E}}, \beta_1\,]$$

with $\hat{\theta}_{\text{RCT-E}}$ estimated using any of the three analytic methods described earlier for estimating the RCT-E θ (though ANCOVA may be preferred for possible efficiency gains after bias correction) and empirical estimates for the population quantities in the expression of $\text{Bias}[\,\hat{\theta}, \beta_1\,]$. We conjecture that this estimator will at least account for the bias induced from enrichment on the basis of our analytic derivation when pre-post data for the broader patient population follow a multivariate normal distribution.

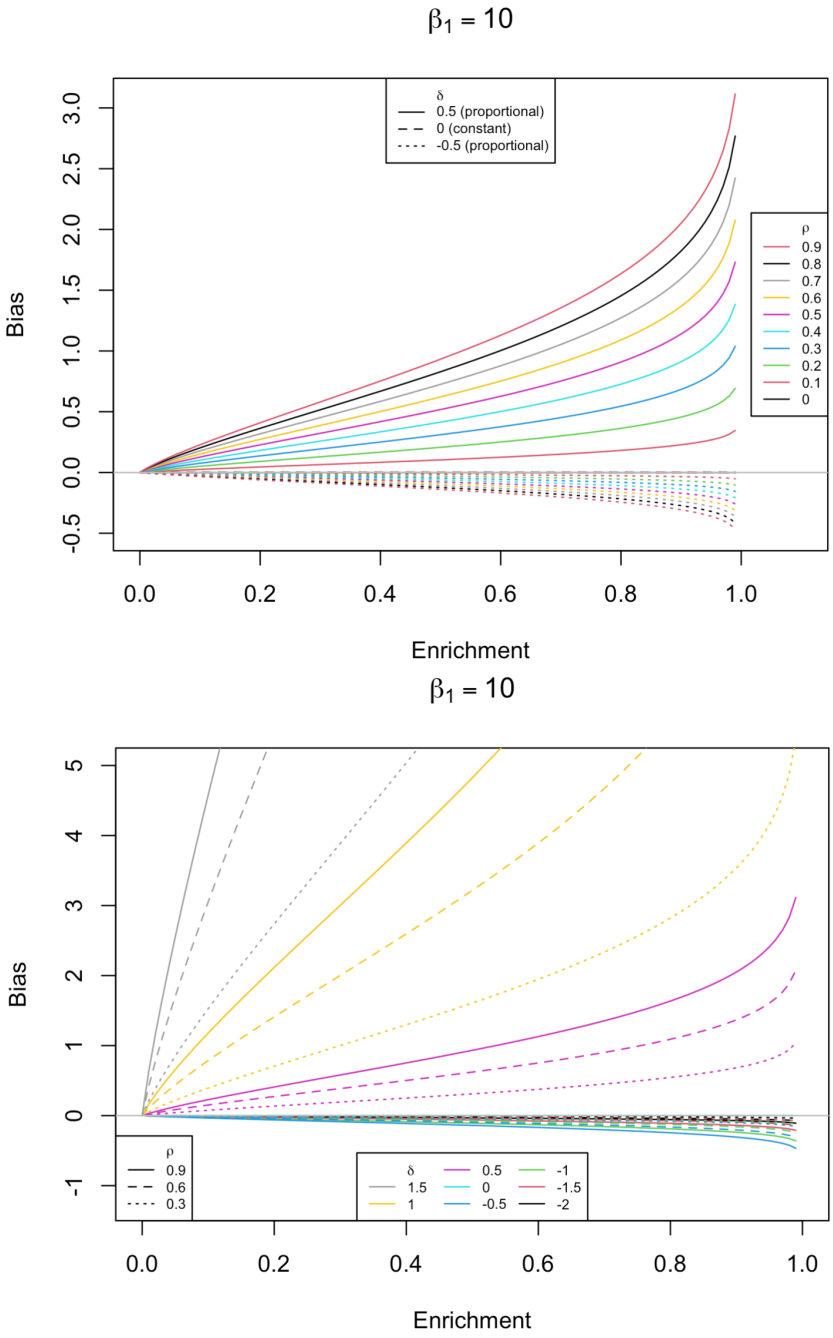


Figure 3.2: Analytic bias (derived under the multivariate normal pre-post data setting) of the RCT-E estimator $\hat{\theta}$ with respect to the RW-E β_1 (here assumed to be 10) as a function of enrichment (value of the probability corresponding to the quantile for the pre-randomization assessment cutoff, to randomize individuals with higher pre values), the mean-variance relationship ($\sigma^2|\mu|^\delta$, with $\sigma = 9$ and pre-randomization $\mu = 40$ for both treatment and control arms), and the pre-post correlation (ρ, same for both treatment and control arms). The top plot considers three mean-variance relationships (varying by line type), each having the pre-post correlation vary by color. The bottom plot considers three pre-post correlations (varying by line type), each having the mean-variance relationship vary by color.

3.3 Simulation studies evaluating our proposed bias-adjusted estimator for the RW-E

We considered a pre-post RCT design in which N screened individuals completed a pre-randomization assessment after meeting all other trial eligibility criteria. We further supposed that each subject's pre-randomization assessment could be expressed as a linear model with a systematic component (in this case, a linear combination of age as a continuous covariate centered around the sample mean of all ages and female sex as a binary covariate) and a random component. For screened subject i, where $i = 1, \ldots, N$, we expressed the pre-randomization assessment as

$$\text{PRE}_i = \alpha_0 + \alpha_1(\text{AGE}_i - \overline{\text{AGE}}) + \alpha_2 \text{FEMALE}_i + \epsilon_{\text{PRE},i}$$

where

$$(\text{AGE}_i - \overline{\text{AGE}}) \sim \mathcal{N}(0, \sigma^2_{\text{AGE}})$$
$$\text{FEMALE}_i \sim \text{Bernoulli}(p_{\text{FEMALE}})$$
$$\epsilon_{\text{PRE}_i} \sim \mathcal{N}(0, f(\mu_{\text{PRE}_i}))$$

with a mean-variance relationship defined as

$$\text{Mean: } \mu_{\text{PRE}_i} = \text{PRE}_i - \epsilon_{\text{PRE},i}$$
$$\text{Variance: } v_{\text{PRE}_i} \equiv f(\mu_{\text{PRE}_i}) = \sigma^2 \left|\mu_{\text{PRE}_i}\right|^{\delta}$$

for $\sigma \in (0, \infty)$ and $\delta \in (-\infty, \infty)$. Negative values of δ correspond to an inversely proportional mean-variance relationship, positive values of δ correspond to a proportional mean-variance relationship, and $\delta = 0$ corresponds to a constant mean-variance relationship. We are

interested in estimating the difference in mean change from baseline between *all* individuals assigned to treatment versus placebo, denoted by β_1 (the RW-E). We defined the true mean and variance of post-randomization assessments for subject i as

$$\mu_{\text{POST}_i} = \beta_{0i} + \beta_1 \text{TX}_i + \beta_2 \text{PRE}_i$$
$$v_{\text{POST}_i} \equiv f(\mu_{\text{POST}_i}) = \sigma^2 \left|\mu_{\text{POST}_i}\right|^{\delta}$$

where $\beta_{0i} \equiv \mu_{\text{PRE},i}$ and PRE_i is a binary indicator of experimental treatment arm assignment (value of 1) versus control or placebo (value of 0). Together, the joint distribution of pre- and post-randomization assessments for subject i

$$\begin{pmatrix} \text{PRE}_i \\ \text{POST}_i \end{pmatrix} \Bigg| \, \text{TX}_i \;\sim\; \mathcal{N}_2\left(\vec{\mu}_i,\; \boldsymbol{\Sigma}_i\right)$$

$$\vec{\mu}_i \equiv \begin{pmatrix} \mu_{\text{PRE}_i} \\ \mu_{\text{POST}_i} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_i \equiv \begin{pmatrix} v_{\text{PRE}_i} & \rho_{\text{TX}_i}\sqrt{v_{\text{PRE}_i} v_{\text{POST}_i}} \\ \rho_{\text{TX}_i}\sqrt{v_{\text{PRE}_i} v_{\text{POST}_i}} & v_{\text{POST}_i} \end{pmatrix}$$

follows a bivariate normal distribution with the above-specified mean-variance relationship where $\rho_{\text{TX}_i} \equiv Corr[\text{PRE}_i, \text{POST}_i \,|\, \text{TX}_i]$.

In our simulation studies, the enrichment criterion was a cutoff based on a probability corresponding to a quantile, $p_{\text{enrich}} \in (0, 1)$, of all pre-randomization assessments from all N screened, otherwise eligible individuals, denoted by $c \equiv \hat{F}^{-1}_{\text{PRE}}(p_{\text{enrich}})$ and $\hat{F}_{\text{PRE}}$ is the empirical distribution function of the observed pre-randomization assessments. We defined an enrichment indicator by $\text{ENRICH}_i = 1$ if $\text{PRE}_i \geq c$, and zero otherwise. We assumed 1:1 randomization allocation of treatment versus control assignment. We generated samples of individuals meeting and not meeting the enrichment criterion, referred to as "enriched" and

"non-enriched" samples, respectively.

Our proposed bias-adjusted estimator requires marginal quantities for pre mean and post variances for the broader patient population (i.e., including both patients who met and did not meet the enrichment criterion). Here, we assume we can reliably estimate these quantities to compute the estimated bias for the RW-E and adjust the estimate for the RW-E.

We conducted simulation studies for the estimation of the RW-E β_1 using our proposed estimator in pre-post RCTs with 1000 screened and 100 randomized participants (in a 1:1 fashion, treatment:control) under four scenarios: (i) enrichment based upon a single pre-randomization (PRE) assessment for screened individuals (no bias correction); (ii) enrichment based upon a single pre-randomization (PRE) assessment for screened individuals (with our proposed bias correction); (iii) enrichment based upon the true mean PRE value for screened individuals; and (iv) no enrichment criterion. For (i)-(iii), individuals were randomized who met the enrichment criterion. We considered an enrichment criterion for which individuals with the top 10%, top 25%, and top 50% of pre-randomization assessments scores were randomized. Statistical operating characteristics include the average estimate of $\hat{\theta}$ (Avg Est: $E[\hat{\theta}]$), estimated bias (and percent relative bias) of $\hat{\theta}$ with respect to β_1, empirical standard deviation (SD), average model-based (MB) or robust (Rob) standard error (SE), and corresponding confidence interval coverage (CI cvrg) probability assuming β_1 is the truth. Additional simulation parameters included: pre-post correlation of 0.9 for each treatment arm; a constant mean-variance relationship (MVR) of the form $\sigma^2|\mu|^\delta$ where $\sigma = 2$ and $\delta = 1$; the true functional form of pre assessments modeled with a binary covariate with 0.5 probability equal to one and a linear continuous covariate (with mean=0 and SD=5), corresponding to a specified covariate parameter vector $\vec{\alpha} = (40, 1, 1)$.

Table 3.1 summarizes results from simulations where the RW-E $\beta_1 = 3$ and enrichment varied (top 10%, 25% and 50% of pre-randomization assessment scores, respectively) while assuming a proportional mean-variance relationship between pre and post assessment scores.

Table 3.1: Estimating RW-E $\beta_1 = 3$ assuming a proportional mean-variance.

(10,000 simulations)	Target	Avg	Est Bias (%)	Emp	Avg SE		CI Cvrg for β_1	
Randomized $n = 100$	Estimand	Est	wrt β_1	SD	MB	Rob	MB	Rob
Enrich with top 10% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.68	0.68 (22.5)	1.63	1.63	1.61	0.93	0.93
w/ Proposed Bias Adj	β_1	2.99	-0.01 (-0.3)	1.59	NA	NA	0.96	0.95
Enrich True Mean PRE	β_1	3.01	0.01 (0.3)	2.87	2.87	2.84	0.95	0.95
No Enrichment	β_1	3.00	0.00 (-0.1)	2.75	2.77	2.74	0.95	0.95
ANCOVA (model SE)								
Enrich Single PRE	θ	3.66	0.66 (22.0)	1.19	1.18	1.16	0.91	0.91
w/ Proposed Bias Adj	β_1	2.97	-0.03 (-0.9)	1.20	NA	NA	0.95	0.94
Enrich True Mean PRE	β_1	3.03	0.03 (0.9)	1.24	1.25	1.23	0.95	0.95
No Enrichment	β_1	3.02	0.02 (0.6)	1.14	1.13	1.12	0.95	0.95
Paired Change (model SE)								
Enrich Single PRE	θ	3.66	0.66 (22.0)	1.19	1.18	1.17	0.92	0.91
w/ Proposed Bias Adj	β_1	2.97	-0.03 (-0.8)	1.20	NA	NA	0.94	0.94
Enrich True Mean PRE	β_1	3.03	0.03 (0.9)	1.26	1.26	1.25	0.95	0.95
No Enrichment	β_1	3.02	0.02 (0.6)	1.15	1.15	1.13	0.95	0.95
Enrich with top 25% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.45	0.45 (14.9)	1.76	1.76	1.74	0.94	0.94
w/ Proposed Bias Adj	β_1	2.96	-0.04 (-1.4)	1.77	NA	NA	0.95	0.95
Enrich True Mean PRE	β_1	3.00	0.00 (0.0)	2.84	2.82	2.79	0.95	0.95
No Enrichment	β_1	2.95	-0.05 (-1.8)	2.76	2.77	2.74	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.46	0.46 (15.2)	1.17	1.17	1.15	0.93	0.93
w/ Proposed Bias Adj	β_1	2.97	-0.03 (-1.1)	1.35	NA	NA	0.91	0.91
Enrich True Mean PRE	β_1	3.00	0.00 (-0.1)	1.21	1.22	1.20	0.95	0.95
No Enrichment	β_1	2.98	-0.02 (-0.5)	1.14	1.14	1.12	0.95	0.95
Paired Change								
Enrich Single PRE	θ	3.46	0.46 (15.4)	1.17	1.17	1.15	0.93	0.93
w/ Proposed Bias Adj	β_1	2.97	-0.03 (-1.0)	1.36	NA	NA	0.91	0.90
Enrich True Mean PRE	β_1	3.00	0.00 (-0.1)	1.23	1.23	1.22	0.95	0.95
No Enrichment	β_1	2.99	-0.01 (-0.4)	1.16	1.15	1.14	0.95	0.95
Enrich with top 50% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.34	0.34 (11.3)	1.99	1.97	1.95	0.94	0.94
w/ Proposed Bias Adj	β_1	3.03	0.03 (1.1)	2.02	NA	NA	0.94	0.94
Enrich True Mean PRE	β_1	2.96	-0.04 (-1.2)	2.76	2.77	2.74	0.95	0.95
No Enrichment	β_1	2.97	-0.03 (-1.1)	2.80	2.77	2.74	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.31	0.31 (10.2)	1.16	1.15	1.14	0.94	0.94
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.0)	1.52	NA	NA	0.87	0.86
Enrich True Mean PRE	β_1	3.01	0.01 (0.2)	1.19	1.19	1.17	0.95	0.95
No Enrichment	β_1	3.00	0.00 (0.0)	1.15	1.14	1.12	0.95	0.94
Paired Change								
Enrich Single PRE	θ	3.31	0.31 (10.2)	1.16	1.16	1.14	0.94	0.94
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.0)	1.54	NA	NA	0.87	0.86
Enrich True Mean PRE	β_1	3.01	0.01 (0.3)	1.20	1.20	1.19	0.95	0.95
No Enrichment	β_1	3.00	0.00 (0.0)	1.16	1.15	1.14	0.95	0.94

Table 3.2: Estimating RW-E $\beta_1 = 3$ assuming an inversely proportional mean-variance.

(10,000 simulations)	Target	Avg	Est Bias (%)	Emp	Avg SE		CI Cvrg for β_1	
Randomized $n = 100$	Estimand	Est	wrt β_1	SD	MB	Rob	MB	Rob
Enrich with top 10% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.10	0.10 (3.5)	0.75	0.74	0.73	0.94	0.94
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.1)	0.77	NA	NA	0.94	0.94
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	1.14	1.14	1.13	0.95	0.95
No Enrichment	β_1	3.00	0.00 (-0.1)	1.41	1.42	1.41	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.10	0.10 (3.4)	0.47	0.46	0.46	0.95	0.94
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.0)	0.54	NA	NA	0.91	0.90
Enrich True Mean PRE	β_1	3.01	0.01 (0.3)	0.46	0.47	0.46	0.95	0.95
No Enrichment	β_1	3.01	0.01 (0.2)	0.45	0.45	0.44	0.95	0.95
Paired Change	β_1							
Enrich Single PRE	θ	3.10	0.10 (3.4)	0.46	0.46	0.46	0.94	0.94
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.0)	0.54	NA	NA	0.91	0.90
Enrich True Mean PRE	β_1	3.01	0.01 (0.3)	0.47	0.47	0.47	0.95	0.95
No Enrichment	β_1	3.01	0.01 (0.2)	0.45	0.45	0.45	0.95	0.95
Enrich with top 25% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.05	0.05 (1.6)	0.82	0.82	0.82	0.95	0.95
w/ Proposed Bias Adj	β_1	2.98	-0.02 (-0.8)	0.90	NA	NA	0.93	0.93
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	1.17	1.16	1.15	0.95	0.95
No Enrichment	β_1	2.97	-0.03 (-0.9)	1.41	1.43	1.41	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.07	0.07 (2.2)	0.46	0.46	0.45	0.95	0.94
w/ Proposed Bias Adj	β_1	2.99	-0.01 (-0.2)	0.67	NA	NA	0.83	0.83
Enrich True Mean PRE	β_1	3.00	0.00 (0.0)	0.46	0.46	0.46	0.95	0.95
No Enrichment	β_1	2.99	-0.01 (-0.2)	0.45	0.45	0.44	0.95	0.95
Paired Change								
Enrich Single PRE	θ	3.07	0.07 (2.2)	0.46	0.46	0.45	0.95	0.95
w/ Proposed Bias Adj	β_1	3.00	0.00 (-0.2)	0.67	NA	NA	0.83	0.83
Enrich True Mean PRE	β_1	3.00	0.00 (0.0)	0.47	0.47	0.46	0.95	0.95
No Enrichment	β_1	3.00	0.00 (-0.1)	0.46	0.45	0.45	0.95	0.95
Enrich with top 50% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.05	0.05 (1.8)	0.94	0.95	0.94	0.95	0.95
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.1)	1.03	NA	NA	0.93	0.93
Enrich True Mean PRE	β_1	2.98	-0.02 (-0.5)	1.21	1.20	1.19	0.95	0.95
No Enrichment	β_1	2.99	-0.01 (-0.3)	1.43	1.43	1.41	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.04	0.04 (1.5)	0.45	0.46	0.45	0.95	0.95
w/ Proposed Bias Adj	β_1	2.99	-0.01 (-0.2)	0.75	NA	NA	0.77	0.77
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	0.46	0.46	0.45	0.95	0.95
No Enrichment	β_1	3.00	0.00 (0.0)	0.46	0.45	0.44	0.95	0.94
Paired Change								
Enrich Single PRE	θ	3.04	0.04 (1.5)	0.45	0.46	0.45	0.95	0.95
w/ Proposed Bias Adj	β_1	2.99	-0.01 (-0.2)	0.75	NA	NA	0.77	0.76
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	0.46	0.46	0.46	0.95	0.95
No Enrichment	β_1	3.00	0.00 (0.0)	0.46	0.45	0.45	0.95	0.95

Table 3.3: Estimating RW-E $\beta_1 = 3$ assuming a constant mean-variance.

(10,000 simulations)	Target	Avg	Est Bias (%)	Emp	Avg SE		CI Cvrg for β_1	
Randomized $n = 100$	Estimand	Est	wrt β_1	SD	MB	Rob	MB	Rob
Enrich with top 10% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.00	0.00 (0.0)	0.48	0.47	0.47	0.95	0.95
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.1)	0.54	NA	NA	0.91	0.91
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	0.57	0.57	0.57	0.95	0.95
No Enrichment	β_1	3.00	0.00 (-0.1)	1.07	1.08	1.07	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.01	0.01 (0.2)	0.18	0.18	0.18	0.95	0.94
w/ Proposed Bias Adj	β_1	3.01	0.01 (0.4)	0.35	NA	NA	0.68	0.68
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	0.18	0.18	0.17	0.95	0.95
No Enrichment	β_1	3.00	0.00 (0.1)	0.18	0.18	0.18	0.95	0.95
Paired Change								
Enrich Single PRE	θ	3.01	0.01 (0.2)	0.18	0.18	0.18	0.95	0.94
w/ Proposed Bias Adj	β_1	3.01	0.01 (0.4)	0.35	NA	NA	0.68	0.68
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	0.18	0.18	0.18	0.95	0.95
No Enrichment	β_1	3.00	0.00 (0.1)	0.18	0.18	0.18	0.95	0.95
Enrich with top 25% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.00	0.00 (0.2)	0.55	0.55	0.55	0.95	0.95
w/ Proposed Bias Adj	β_1	3.00	0.00 (0.0)	0.67	NA	NA	0.90	0.89
Enrich True Mean PRE	β_1	3.00	0.00 (0.1)	0.64	0.63	0.63	0.95	0.95
No Enrichment	β_1	2.99	-0.01 (-0.5)	1.06	1.08	1.07	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
w/ Proposed Bias Adj	β_1	3.00	0.00 (-0.1)	0.48	NA	NA	0.54	0.53
Enrich True Mean PRE	β_1	3.00	0.00 (0.0)	0.18	0.18	0.17	0.95	0.95
No Enrichment	β_1	3.00	0.00 (-0.1)	0.18	0.18	0.18	0.95	0.95
Paired Change								
Enrich Single PRE	θ	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
w/ Proposed Bias Adj	β	3.00	0.00 (-0.1)	0.48	NA	NA	0.54	0.53
Enrich True Mean PRE	β_1	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
No Enrichment	β_1	3.00	0.00 (-0.1)	0.18	0.18	0.18	0.95	0.95
Enrich with top 50% pre-randomization assessment scores								
Two-sample t-test								
Enrich Single PRE	θ	3.00	0.00 (0.0)	0.67	0.67	0.66	0.95	0.95
w/ Proposed Bias Adj	β	3.00	0.00 (-0.1)	0.78	NA	NA	0.91	0.91
Enrich True Mean PRE	β	2.99	-0.01 (-0.2)	0.74	0.72	0.72	0.95	0.95
No Enrichment	β	3.00	0.00 (0.0)	1.08	1.08	1.07	0.95	0.95
ANCOVA								
Enrich Single PRE	θ	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
w/ Proposed Bias Adj	β	3.00	0.00 (-0.1)	0.53	NA	NA	0.49	0.48
Enrich True Mean PRE	β	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
No Enrichment	β	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.94
Paired Change								
Enrich Single PRE	θ	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
w/ Proposed Bias Adj	β	3.00	0.00 (-0.1)	0.53	NA	NA	0.48	0.48
Enrich True Mean PRE	β	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.95
No Enrichment	β	3.00	0.00 (0.0)	0.18	0.18	0.18	0.95	0.94

For a pre-post design with enrichment based on the top 10% of single pre-randomization assessment scores, we overestimated the RW-E β_1 by approximately 22% using the two-sample t-test, ANCOVA, or paired change models. After applying our proposed bias adjustment, removing the impact of regression to the mean in order to target β_1 instead of θ, we obtained approximately unbiased estimates for β_1. For enrichment of the top 10%, our proposed bias adjustment applied to each of the t-test, ANCOVA, and paired change yielded coverage probabilities at approximately the nominal 0.95 level. As the enrichment criterion becomes less restrictive, however, we found that while the bias adjustment applied to the two-sample t-test maintain approximately nominal coverage, those for ANCOVA and paired change became anti-conservative (between 0.86 to 0.91). We found similar results, with further reductions in coverage probabilities when considering inversely proportional and constant mean-variance relationships (Table 3.2 and Table 3.3, respectively).

Furthermore, we considered two alternative pre-post designs: enrichment based on the true mean pre-randomization value of each screened participant and no enrichment. For both of these designs, we obtained approximately unbiased estimates for β_1 and achieved nominal coverage probabilities across enrichment cutoffs and mean-variance relationships. Not surprisingly, as compared to top 10%, results for top 25% and 50% enrichment were qualitatively similar where the amount of bias with $\hat{\theta}$ attenuated as the enrichment criterion became less restrictive (i.e., regression to the mean became less of an issue).

3.4 Discussion

Enrichment in a pre-post RCT can result in biased estimates of treatment efficacy depending upon the estimand of interest. We have demonstrated, analytically and via simulation studies, that a single assessment enrichment-based inclusion criterion using a pre-randomization assessment can yield biased estimates of treatment efficacy for a broader patient population

RW-E β_1 when using the trial-specific RCT-E estimator $\hat{\theta}$ without any adjustment. Simply relying upon using a robust standard error estimate, such as White (1980), that only applies a post-hoc fix to the variance of an estimate does not remove the bias when the variance of pre and post assessments is a function of the respective mean (i.e., a mean-variance relationship). Furthermore, this suggests that simply increasing the planned sample size for an enriched pre-post RCT will not aid in validly estimating the RW-E — analogous to obtaining a more precise biased estimate by inflating the planned sample size of a clinical trial in the attempts of trying to account for having (non-ignorable) missing data (Fleming, 2011). Hence, an alternative pre-specified approach, such as what we proposed, was needed to correct for the bias that may arise from an enriched pre-post RCT when estimating the RW-E.

We showed scenarios in which a proportional mean-variance relationship for pre and post assessments can lead to overestimation bias for the RW-E from an enriched pre-post RCT. Issues can thus arise with RCTs where the primary or co-primary endpoints are measured on continuous scales with restricted ranges. For instance, in Alzheimer's disease trials measures of cognition (e.g., the Alzheimer's Disease Assessment Scale-Cognitive Subscale, ADAS-Cog (Rosen et al., 1984; Mohs et al., 1997)) and function (e.g., the Alzheimer's Disease Cooperative Study Activities of Daily Living for Mild Cognitive Impairment, ADCS-ADL-MCI scores range from 0 to 53 with lower scores indicating greater functional impairment) have restricted ranges. To this end, encountering non-constant mean-variance relationships of assessments used in RCTs seem likely. Furthermore, the magnitude of the bias may depend upon the form of the mean-variance relationship, the pre-post correlation (based on the choice of biomarker or surrogate outcome measure selected), and the threshold used to define meeting the enrichment criterion. The enrichment threshold is ideally selected for valid scientific and ethical reasons. If, however, there is interest to learn about a subpopulation of individuals who may not meet the enrichment criterion, careful consideration should be taken at the trial design stage to decide whether to include or exclude such a subpopulation.

We proposed a bias-adjusted estimator for the RW-E, $\hat{\beta}_1^{\text{BiasAdj}}$, that uses the RCT-E estimator $\hat{\theta}$ and empirical estimates to estimate the bias term we derived under the multivariate normal pre-post data setting, and empirically demonstrated reduction in the bias compared to the naive application of the RCT-E estimator $\hat{\theta}$ without any adjustment. From our investigation, we found that applying our bias adjustment to the two-sample t-test estimate of the RCT-E yielded an estimate for the RW-E that was approximately unbiased and yielded confidence intervals approximately achieving the nominal coverage probability level. On the other hand, we found applying our bias adjustment to the ANCOVA and paired change estimates of the RCT-E, while resulting in approximately unbiased estimates for the RW-E, yielded marked drops in coverage probabilities, especially as the enrichment criterion was less restrictive. This is a limitation of our proposed approach. One explanation for this may be that the bias term we derived under the normality assumption can more directly correspond to the formulation of the two-sample t-test estimation. Deriving possible specific ANCOVA-based and paired-change-based bias adjustments may result in both unbiased estimation and nominal coverage probabilities, while drawing upon ANCOVA's typically more efficient design. In the interim, however, we also found that enrichment based on the true mean pre-randomization assessment for individuals alleviates issues with estimating RW-E from the RCT sample. In practice, this would correspond to an enrichment period with multiple pre assessments over time to remove the impact of regression to the mean when a single assessment is taken that could be randomly higher or lower than one's true pre-randomization mean assessment value.

Another limitation of our proposed bias-adjusted estimator for the RW-E is that it was derived under the assumption of multivariate normality of pre-post data and a known mean-variance relationship. In practice, however, knowing the exact form of the mean-variance relationship may be difficult. It is therefore incumbent to estimate the mean-variance relationship before collection of the post-randomization data. This can be done by leveraging data from the available pre-randomization assessments for all screened individuals of the

trial under study or using an auxiliary data source (e.g., from a pilot study or phase 2 trial) to estimate post variance and pre-post correlation for the control arm. Then, to obtain corresponding marginal estimates for the treatment arm, select a prediction model for the pre-randomization assessments based on pre-randomization covariates (e.g., one continuous covariate, such as age, and one binary covariate, such as biological sex) by minimizing Akaike's Information Criterion (AIC; Akaike, 1974). To flexibly estimate the form of the mean-variance relationship empirically, one approach is to fit a generalized additive model (GAM; Hastie and Tibshirani, 1987) by regressing the squared residuals against the corresponding fitted values from the selected prediction model. Then calculate the predicted value from the GAM of the post-randomization marginal variance for the treatment arm by computing the squared residual value corresponding to the sample mean of pre-randomization assessments among all individuals (those meeting and not meeting the enrichment criterion) plus the trial-specific estimated treatment effect among enriched individuals, $\hat{\theta}$. An additional proposed fix to the bias induced in this setting is to use the estimated form of the mean-variance relationship and plug in the corresponding estimates of the squared residuals from the above mentioned GAM and perform iteratively re-weighted least squares. When it is of interest to allow for the possibility of a mean-variance relationship in enriched pre-post trials, weighted least squares (WLS) seems reasonable à la generalized linear models (GLMs). Obtaining a WLS estimator in this setting would allow relaxing the normality assumption for pre-post data that our analytic fix relies upon. Based on the Gauss-Markov theorem, this WLS estimator would be the best linear unbiased estimator.

Understanding which estimand is of interest (e.g., the RCT-E vs. the RW-E) is important for clinical decision making, but also an important consideration by trialists at the design stage prior to start of a new trial. This is important for settings where individuals not meeting a trial's enrichment criterion, and hence become ineligible to be randomized, are not studied in RCTs. In clinical practice, this subpopulation may comprise individuals who may not meet a similar criterion on an approved drug label's indication, and may receive

drug off-label, especially if there are no other alternatives. One concern with off-label use is that the benefit-to-risk ratio may differ even if the safety profile is assumed to be similar for subpopulations of patients meeting versus not meeting the enrichment criterion indicated on a drug label. Trial design is therefore an important stage for sponsors in consultation with review boards and regulators to carefully consider the pros and cons of the design, conduct, and analysis choices prior to a trial's initiation and enrollment. Our work is a step forward to facilitate well-informed *a priori* decision making about what may be an issue in terms of generalizability of trial results beyond the studied RCT sample.

Chapter 4

Censoring-Robust Estimation in Fixed Sample Time-to-Event Clinical Trials with Adaptive Randomization

SUMMARY: Adaptive randomization is a clinical trial design feature used to modify treatment allocation probabilities during accrual. In time-to-event trials, the impact of adaptive randomization is less well-understood for estimating treatment efficacy in the presence of time-varying effects (e.g., relative risk of progression to AIDS or death changes over time). Here, we focus on time-to-event trials where the scientific estimand is a marginal hazard ratio in the absence of intermittent censoring over the support of observed times. We analytically show that adaptive randomization alters censoring patterns and illustrate via Monte Carlo simulations that the Cox proportional hazards estimator can yield biased estimates. As a remedy, we propose a censoring-robust estimator based on reweighting the partial likelihood score by treatment-specific censoring distributions that account for adaptive randomization. We derive the asymptotic properties of the proposed estimator and evaluate its finite sample operating characteristics via simulation. Finally, we apply our proposed method using data

from the Community Programs for Clinical Research on AIDS Trial 002.

4.1 Introduction

Reliably estimating treatment efficacy in randomized clinical trials is imperative to adequately assess benefit-risk of a candidate intervention (U.S. Food and Drug Administration, 2019a, 2023). Reliable estimation requires pre-specifying an estimand, study design, and corresponding statistical analysis method. In randomized clinical trials with a right censored time-to-event primary endpoint, the estimand is often parametrized by a hazard ratio. The Cox proportional hazards model (1972) is ubiquitously pre-specified to estimate a hazard ratio. When proportional hazards and independent censoring hold, a correctly specified Cox model has a semiparametric efficient estimator consistent for an estimand that is a constant hazard ratio over the observed support.

In the presence of time-varying treatment effects, however, prior work has shown that the underlying estimand from a misspecified Cox model is a marginal hazard ratio that can depend on the trial's censoring patterns, marginally under independent censoring (Xu and O'Quigley, 2000) or conditionally by treatment arms under covariate dependent censoring (Struthers and Kalbfleisch, 1986; Boyd et al., 2012). Such a dependence on censoring is often not of scientific interest and will restrict inference. To this end, censoring-robust estimators have been proposed (Xu and O'Quigley, 2000; Boyd et al., 2012) because they target an estimand that is a hazard ratio in the absence of intermittent censoring over the support of observed times, denoted here by $\theta^* \equiv \exp(\beta^*)$. These approaches are an important step towards reliably estimating treatment efficacy. Yet, the role of randomization scheme (fixed versus adaptive) and impact on the underlying censoring distributions remain less well-understood in time-to-event trials.

Compared to fixed randomization, adaptive randomization allows for changing allocation probabilities for newly enrolled trial participants that may accelerate information accrual on the treatment arm. Types of adaptive randomization include restricted, covariate-adaptive, response-adaptive, and covariate-adjusted response-adaptive randomization (Hu and Rosenberger, 2006; Rosenberger et al., 2012) that can be based on comparative data (U.S. Food and Drug Administration, 2019a). For time-to-event clinical trials, Ye and Shao (2020) have investigated covariate-adaptive randomization and Zhang and Rosenberger (2007) and Korn and Freidlin (2011, 2017, 2022) have examined response-adaptive randomization. Covariate-adjusted response-adaptive randomization may also have utility in time-to-event trials for late-stage disease where the event can occur quickly and accrual for most of the trial's duration Rosenberger et al. (2012).

In this chapter, we show that adaptive randomization can alter follow-up times and thus censoring patterns in a trial. We propose a censoring-robust estimator (CRE) that incorporates adaptive randomization into the weight used to reweight the partial likelihood score (à la Boyd, Kittelson, and Gillen (BKG), 2012) in Section 4.2. We then examine frequentist operating characteristics of our proposed adaptive randomization CRE, the BKG CRE, and the Cox proportional hazards estimator (the latter serving as the referent estimator used in practice) via simulation studies in Section 4.3. In Section 4.4 we apply our method after inducing an adaptive randomization scheme to data from the Community Programs in Clinical Research on AIDS Trial 002 (Abrams et al., 1994), a time-to-event non-inferiority trial with a fixed 1:1 randomization ratio that exhibited time-varying treatment effects. We finish with a discussion and conclusions in Section 4.5.

4.2 Methods

We consider a time-to-event clinical trial that enrolls n eligible participants from trial start until the end of the accrual period, $\tau_{Accrual}$. After introducing notation for fixed versus adaptive randomization schemes and time-to-event outcomes, we review methods to estimate a hazard ratio under proportional and non-proportional hazards, emphasizing how the estimand can differ between these settings. Then, we analytically show how adaptive randomization alters treatment-specific censoring patterns and propose a censoring-robust estimator that incorporates the adaptive randomization rule.

4.2.1 Fixed versus adaptive randomization

For fixed randomization, these participants are randomly assigned to treatment $Z = z$ with probability $\pi_{Z=z} \equiv \Pr(Z = z)$. For adaptive randomization, we define accrual subperiods according to when treatment allocation probabilities change. Let K be the number of non-overlapping accrual subperiods that form a finite partition of $(0, \tau_{Accrual})$. Let A be the index of an accrual subperiod k taking values $\{1, 2, ..., K\}$. For accrual subperiod k, let n_k be the number of enrolled participants and $\pi_{A=k} \equiv \Pr(A = k) = n_k/n$ denote the proportion enrolled among all n participants. We then define treatment allocation probabilities according to the accrual subperiod, $\pi_{Z=z|A=k} \equiv \Pr(Z = z|A = k)$. While our notation can be used for multi-arm clinical trials and those with different types of adaptive randomization (e.g., covariate-adaptive or response-adaptive), in this article we focus on two-arm clinical trials with either fixed randomization or adaptive randomization based on non-comparative data. Figure 4.1 illustrates the four randomization schemes (one fixed and three adaptive) we consider that were pre-determined according to accrual calendar time.

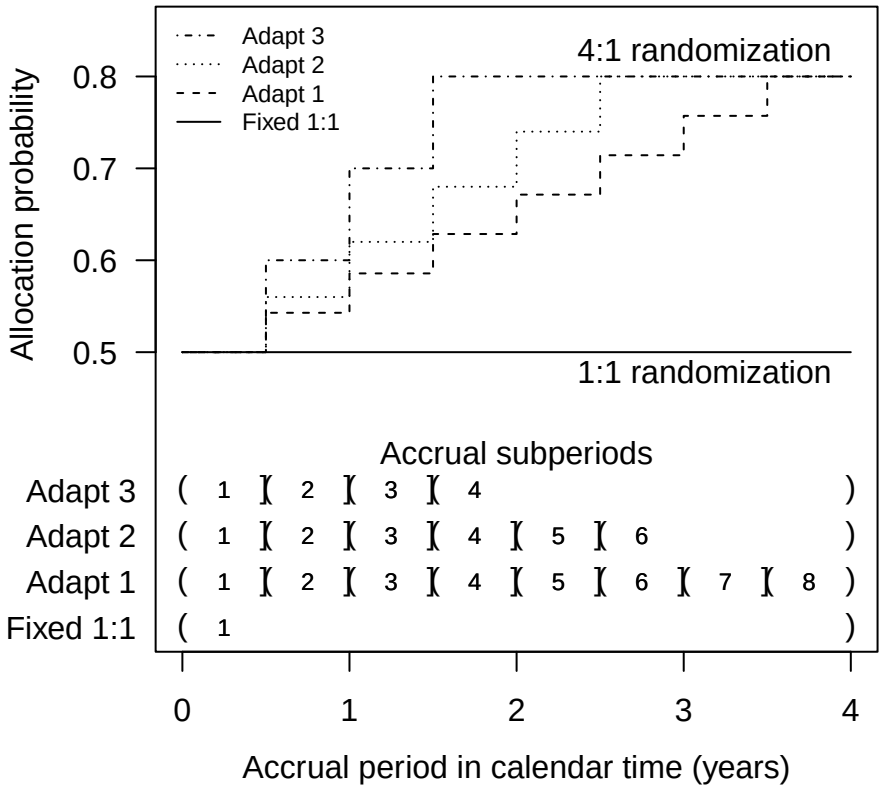


Figure 4.1: Four randomization schemes of treatment allocation probabilities over trial accrual.

4.2.2 Hazard ratio estimands under non-proportional hazards

Let T_i and C_i denote the time to an event and censoring, respectively, for individual i with observed time $X_i = \min\{T_i, C_i\}$ and event indicator $\Delta_i = I(T_i \leq C_i)$. Also, $f(t)$, $F(t)$, $S(t) = 1 - F(t)$, and $\lambda(t) = \lim_{h \downarrow 0} \Pr[t \leq T < t+h | T \geq t]/h$ denote the density, distribution, survivor, and hazard functions.

The Cox proportional hazards model (1972) parametrizes the hazard by

$$\lambda(t) = \lambda_0(t) \exp(\beta Z) \tag{4.1}$$

with baseline hazard $\lambda_0(t)$, p-dimensional covariate vector Z, and p-dimensional parameter vector β. As referenced above, in this article we assume $p = 1$ in the two-arm trial context. The corresponding partial likelihood estimating equation for β is

$$\mathcal{U}_{CoxPL}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left(Z_i - \frac{S^{(1)}(t, \beta)}{S^{(0)}(t, \beta)} \right) dN_i(t) \tag{4.2}$$

where $S^{(r)}(t) = n^{-1} \sum_{j=1}^{n} Y_j(t) Z_j^r \exp(\beta Z_j)$, $Y_j(t)$ is an at-risk indicator at time t, and $N_i(t)$ is the number of events that occurred in $(0, t)$ for individual i.

Under proportional hazards, the Cox proportional hazards estimator $\hat{\beta}_{CoxPH}$ is consistent for the value of β that solves $\mathcal{U}_{PL}(\beta) = 0$, denoted by β_{PH}. When either independent censoring ($T_i \perp\!\!\!\perp C_i$ and $C_i \perp\!\!\!\perp Z_i$) or conditionally independent censoring ($T_i \perp\!\!\!\perp C_i | Z_i$) hold, the interpretation of the $\exp(\beta_{PH})$ estimand is a constant hazard ratio in the absence of intermittent censoring over the support of observed times.

In the presence of a time-varying effect (e.g., non-proportional hazards), fitting the Cox proportional hazards model is an example of a misspecified model. Struthers and Kalbfleisch (1986) show that, under a misspecified Cox model, $\hat{\beta}_{CoxPH}$ is consistent for the value of β

that solves

$$\int_0^\infty E_Z \left\{ f_T(t|Z) S_C(t|Z) \left[Z - \frac{E_Z \left\{ Z S_T(t|Z) S_C(t|Z) e^{\beta Z} \right\}}{E_Z \left\{ S_T(t|Z) S_C(t|Z) e^{\beta Z} \right\}} \right] \right\} dt = 0 \tag{4.3}$$

denoted by $\beta_{S_C(\cdot|Z)}$ because of its notable dependence on the treatment-specific censoring distributions.

Because censoring patterns are not often of scientific interest, but often instead a byproduct of a trial's design, we contend that the scientific estimand should not depend on a trial's intermittent censoring patterns. No intermittent censoring can only arise when all individuals who do not experience an event are followed up to the same maximum follow-up time τ. In such a setting, $S_C(t|Z) = 1$ for all t and (4.3) reduces to

$$\int_0^\infty E_Z \left\{ f_T(t|Z) \left[Z - \frac{E_Z \left\{ Z S_T(t|Z) e^{\beta Z} \right\}}{E_Z \left\{ S_T(t|Z) e^{\beta Z} \right\}} \right] \right\} dt = 0 \tag{4.4}$$

with the corresponding solution denoted by β^*. In this article, we consider $\theta^* \equiv \exp(\beta^*)$ as the scientific estimand, interpreted as a marginal hazard ratio comparing treatment to control in the absence of intermittent censoring over the support of observed times.

While it is not always feasible to achieve no intermittent censoring in a trial, we contend that $\theta^* \equiv \exp(\beta^*)$ remains the scientific estimand. Under independent censoring (IC) where $S_C(t|Z) = S_C(t)$, Xu and O'Quigley (2000) show that the solution to (4.3) still depends on the marginal censoring distribution $S_C(t)$ and proposed reweighting the partial likelihood score by point-wise Kaplan-Meier (1958) estimates of the marginal censoring distribution, $S_C(t)$. Their estimator $\hat{\beta}_{\text{XO}}$ is consistent for $\beta^* \equiv \beta_{IC}$.

In most clinical trials, however, independent censoring may be violated due to differential adverse events profiles for treatment and control inducing covariate-dependent censoring. Instead, conditionally independent censoring (CIC) may be a more appropriate assumption.

Boyd et al. (2012) use the result from (4.3) in the context of a two-arm clinical trial and propose reweighting the partial likelihood score by point-wise Kaplan-Meier estimates of the treatment-specific censoring distributions, $S_C(t|Z)$. Their estimator, $\hat{\beta}_{BKG}$ is consistent for $\beta^* \equiv \beta_{CIC}$. We adopt the nomenclature of censoring-robust estimation (Boyd et al., 2012) as a framework to examine the role of adaptive randomization on estimating the scientific estimand $\theta^* \equiv \exp(\beta^*)$.

4.2.3 An adaptive randomization censoring-robust estimator

For any clinical trial using adaptive randomization, the adaptive randomization rule should be pre-specified. But even if the adaptive randomization rule is not pre-specified, it should be documented what the changes were to $\pi_{Z=z|A=a}$ and what was $\pi_{A=a}$. In most practical settings, these will be known. As such, it would behoove us to incorporate that information into the weight used for re-weighting the estimating equation to remove the dependence on intermittent censoring patterns for the maximal observed follow-up in a trial. To this end, we extend the work by Boyd et al. (2012) for the censoring-robust estimation framework, under conditionally independent censoring, by incorporating knowledge of the adaptive randomization rule.

Consider an adaptive randomization rule $\mathcal{AR} = \{\pi_{Z=z|A=k} \text{ for all } z \text{ and } k = 1, ..., K\}$ denoting the set of allocation probabilities for the K subperiods partitioning $(0, \tau_{Accrual})$. We derived that a treatment-specific censoring distribution can be expressed as a weighted average with respect to accrual subperiods

$$S_C(t|Z = z) = \sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) S_C(t|Z = z, A = k) \tag{4.5}$$

with treatment-subperiod weight

$$w_k^{\mathcal{AR}}(z) \equiv \frac{\pi_{Z=z|A=k}\pi_{A=k}}{\sum_{a=1}^{K}\pi_{Z=z|A=a}\pi_{A=a}}$$

(see Appendix B.1 for derivation). Based on the decomposition in (4.5), we propose an estimator for $S_C(t|Z)$ that incorporates the adaptive randomization rule and uses a left-continuous Kaplan-Meier estimate of the treatment-subperiod-specific censoring distribution at time t as a plug-in estimate for $S_C(t|Z = z, A = k)$. Using this result, we propose an adaptive randomization censoring-robust estimator, denoted by $\hat{\beta}_{CRE}^{\mathcal{AR}}$, that incorporates an adaptive randomization rule in the weights $\hat{W}_i^{\mathcal{AR}}(t) = 1/\hat{S}_C^{\mathcal{AR}}(t|Z = z)$ used to reweight the partial likelihood score. Solving the following reweighted estimating equation for β

$$\mathcal{U}_{CRE}^{\mathcal{AR}}(\beta) = \sum_{i=1}^{n}\int_0^{\infty} W_i^{\mathcal{AR}}(t)\left(Z_i - \frac{S_{\mathcal{AR}}^{(1)}(t,\beta)}{S_{\mathcal{AR}}^{(0)}(t,\beta)}\right)dN_i(t) \equiv 0 \tag{4.6}$$

where $S_{\mathcal{AR}}^{(r)}(t) = n^{-1}\sum_{j=1}^{n} W_j^{\mathcal{AR}}(t)Y_j(t)Z_j^r\exp(\beta Z_j)$ yields a consistent estimate for β^* (see Proposition 4.1).

Proposition 4.1: *For a known adaptive randomization rule and under conditionally independent censoring, the estimator that solves (4.6) has* $\sqrt{n}(\hat{\beta}_{CRE}^{\mathcal{AR}} - \beta^*) \to_d \mathcal{N}(0, V(\beta^*))$ *as* $n \to \infty$. (See Appendix B.2 for a proof.)

4.3 Simulations

In this section we present simulation studies evaluating frequentist operating characteristics of the Cox proportional hazards (PH) estimator, Boyd-Kittelson-Gillen (BKG) censoring-robust estimator (CRE), and our proposed adaptive randomization CRE in time-to-event clinical trials with adaptive randomization. Because we showed in the previous section how

adaptive randomization induces covariate-dependent censoring, we do not consider the Xu-O'Quigley CRE as the independent censoring assumption is violated. This is often true in clinical trials because adverse event profiles differ by arm leading to treatment-specific censoring patterns.

4.3.1 Simulations scenarios

We considered three data generating models: constant relative benefit, delayed benefit, and waning benefit; see the left panel of Figure 4.2. For each data generating model, the corresponding estimand of interest is a marginal hazard ratio over the observed support in the absence of intermittent censoring, denoted by $\theta^* \equiv \exp(\beta^*)$.

To obtain an approximate value for θ^* for a given data generating model, we generated a single dataset with a sample size of 100,000 participants per treatment arm. We assumed instantaneous accrual and 1:1 randomization. Participants were only administratively censored at 4 years if they had not already experienced the event.

For each data generating model, we considered four randomization schemes. The first was a fixed 1:1 randomization over the entire accrual period $(0, 4)$ years. The remaining three were adaptive randomization schemes that all started with 1:1 randomization for the first six months followed by increasing the probability of being assigned to treatment ($Z = 1$) from 0.5 to 0.8 over one year (*Adapt 1*), two years (*Adapt 2*), and three years (*Adapt 3*); see Figure 4.1. For example with *Adapt 1*, we partitioned the accrual period into four subperiods: $(0, 0.5)$, $[0.5, 1)$, $[1, 1.5)$, and $[1.5, 4)$, with treatment ($Z = 1$) allocation probabilities of 0.5 (i.e., 1:1 randomization), 0.6, 0.7, and 0.8 (i.e., 4:1 randomization), respectively. The scientific estimand for each data generating mechanism remained the same across accrual-randomization scenarios considered.

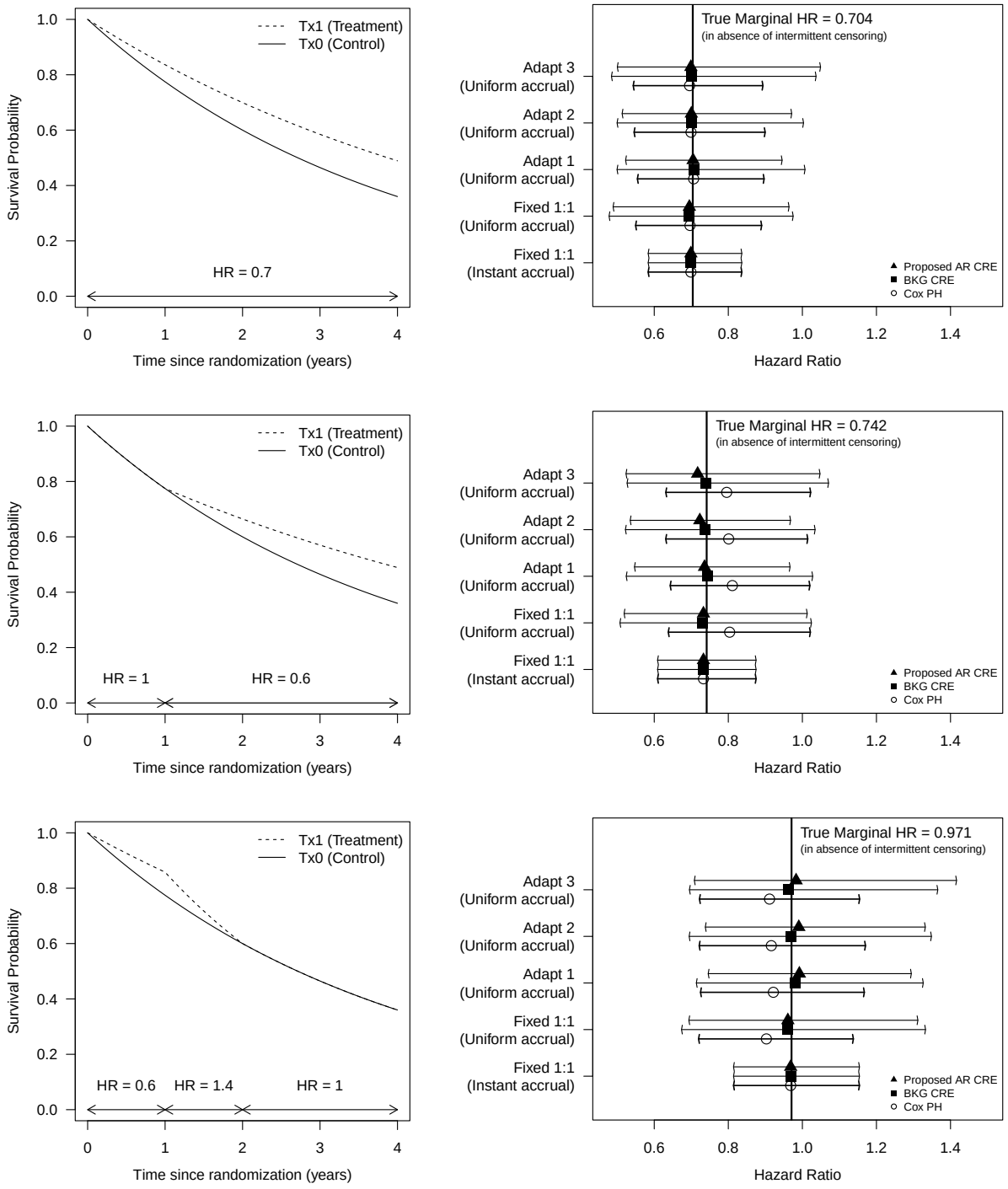


Figure 4.2: Data generating models for constant relative benefit (top left), delayed benefit (middle left), and waning benefit (bottom left) along with corresponding forest plots (right) of estimated hazard ratios and 95% confidence intervals from simulation studies.

We considered two accrual settings. For the instantaneous accrual setting, we generated true event times according to whether an individual was randomly assigned to treatment or control in a 1:1 fashion using piece-wise exponential or exponential distributions. For the remaining settings, we considered a four-year time-to-event clinical trial that accrued participants for the entire four-year period. First, we generated entry times into the trial for each participant, $E_i \sim \text{Uniform}(0, 4)$. We did not assume any other censoring occurred except for administrative censoring. Here, we considered each of the randomization schemes in Figure 4.1. For the adaptive randomization schemes, the censoring time (with respect to time since randomization) depended on their entry into the study (with respect to calendar time), denoted by $C_i = 4 - E_i$.

We compared the Cox PH estimator, the BKG CRE, and our proposed adaptive randomization CRE against the marginal log hazard ratio in absence of intermittent censoring over the maximal support β^*. In all scenarios, we used uniform accrual over the four-year study. Each scenario consisted of 1000 Monte Carlo simulations. We calculated: bias (average of the estimated log hazard ratios across simulations minus β^*); relative bias with respect to β^*; average of the standard error estimates across simulations; empirical standard deviation of the estimated log hazard ratios across simulations; confidence interval coverage probability.

4.3.2 Simulation results

The right panel of Figure 4.2 displays forest plots of estimated hazard ratios with 95% confidence intervals (CI) for constant relative benefit (top row), delayed benefit (middle row), and waning benefit (bottom row). For each forest plot, the horizontal axis displays the hazard ratio scale and the vertical axis includes each accrual-randomization trial design: Instantaneous accrual with fixed 1:1 randomization; and uniform accrual with fixed 1:1, *Adapt 1*, *Adapt 2*, and *Adapt 3* randomization. For each accrual-randomization design, we

display the estimated hazard ratio (HR) and corresponding 95% CI for each estimator (Cox PH, BKG CRE, and our proposed adaptive randomization (AR) CRE). The instantaneous accrual scenarios had no intermittent censoring by design. In these instances, both censoring-robust estimators are equivalent to the Cox PH estimator, thus yielding the same results under proportional hazards and non-proportional hazards. Under uniform accrual and fixed randomization, the Cox PH estimator yields biased estimates for $\theta^* \equiv \exp(\beta^*)$ while both censoring-robust estimators yield approximately unbiased estimates. Similar estimates are observed under uniform accrual and adaptive randomization; however, because our proposed estimator accounts for the adaptive randomization rule, it had higher precision compared to the BKG CRE. Table 4.1 summarizes additional operating characteristics for these data generating mechanisms. Of note, the coverage probabilities for the Cox PH estimator are anti-conservative under non-proportional hazards whereas both censoring robust estimators achieve the nominal CI coverage probability level. Because the Cox PH estimator yielded biased estimates under the non-proportional hazards scenarios investigated, the corresponding coverage probabilities (which were anti-conservative) have less relevance when the goal is to reliably estimating treatment efficacy.

4.4 Application to data from Community Programs for Clinical Research on AIDS Trial 002

We consider the Community Programs for Clinical Research on AIDS (CPCRA) Trial 002 (Abrams et al., 1994) of individuals infected with human immunodeficiency virus who had not received benefit from first-line treatment with zidovudine. CPCRA Trial 002 enrolled 467 participants from December 1990 to September 1991 and, using fixed 1:1 randomization, assigned 237 participants to the experimental treatment, zalcitabine (ddC), and 230 to the standard of care, didanosine (ddI). This was a non-inferiority trial in which the ob-

Table 4.1: Simulation results for three data generating mechanisms comparing the Cox proportional hazards estimator, the Boyd-Kittleson-Gillen (BKG) censoring-robust (CR) estimator, and our proposed CR estimator against the marginal log hazard ratio in absence of intermittent censoring over the maximal support β^*. See Figure 4.2 for details of the randomization schemes considered. In all scenarios, we used uniform accrual over the 4 year study. Each scenario consisted of 1000 simulations.

Constant relative benefit ($\beta^* = -0.353$)

	Cox PH Estimator					BKG CR Estimator					Our Proposed Adapt. Rand. CR Estimator				
Rand.	Est.	Bias (Rel. Bias)	SE	SD	CP	Est.	Bias (Rel. Bias)	SE	SD	CP	Est.	Bias (Rel. Bias)	SE	SD	CP
Fixed 1:1	-0.362	-0.008 (2.4%)	0.125	0.128	95.5	-0.364	-0.011 (3.1%)	0.167	0.181	93.4	-0.364	-0.011 (3.1%)	0.167	0.181	93.4
Adapt 1	-0.348	0.006 (-1.6%)	0.126	0.125	95.4	-0.347	0.006 (-1.7%)	0.168	0.181	95.6	-0.349	0.005 (-1.3%)	0.148	0.149	95.5
Adapt 2	-0.358	-0.004 (1.2%)	0.128	0.125	95.0	-0.356	-0.002 (0.6%)	0.170	0.180	95.0	-0.357	-0.003 (0.9%)	0.151	0.152	94.6
Adapt 3	-0.363	-0.010 (2.8%)	0.131	0.128	95.7	-0.356	-0.003 (0.8%)	0.171	0.187	94.7	-0.360	-0.006 (1.8%)	0.157	0.163	94.6

Delayed benefit ($\beta^* = -0.305$)

	Cox PH Estimator					BKG CR Estimator					Our Proposed Adapt. Rand. CR Estimator				
Rand.	Est.	Bias (Rel. Bias)	SE	SD	CP	Est.	Bias (Rel. Bias)	SE	SD	CP	Est.	Bias (Rel. Bias)	SE	SD	CP
Fixed 1:1	-0.219	0.087 (-28.4%)	0.121	0.125	87.5	-0.313	-0.008 (2.5%)	0.164	0.177	94.1	-0.313	-0.008 (2.5%)	0.164	0.177	94.1
Adapt 1	-0.210	0.096 (-31.4%)	0.123	0.121	88.6	-0.297	0.008 (-2.6%)	0.163	0.176	95.6	-0.288	0.017 (-5.6%)	0.144	0.144	95.5
Adapt 2	-0.222	0.084 (-27.4%)	0.125	0.120	91.1	-0.304	0.001 (-0.5%)	0.164	0.171	95.9	-0.297	0.009 (-2.8%)	0.147	0.146	95.6
Adapt 3	-0.229	0.077 (-25.1%)	0.129	0.123	92.2	-0.302	0.003 (-1.1%)	0.166	0.177	95.7	-0.300	0.006 (-1.9%)	0.153	0.155	96.0

Waning benefit ($\beta^* = -0.030$)

	Cox PH Estimator					BKG CR Estimator					Our Proposed Adapt. Rand. CR Estimator				
Rand.	Est.	Bias (Rel. Bias)	SE	SD	CP	Est.	Bias (Rel. Bias)	SE	SD	CP	Est.	Bias (Rel. Bias)	SE	SD	CP
Fixed 1:1	-0.102	-0.072 (2.4-fold)	0.118	0.119	91.4	-0.040	-0.009 (30.1%)	0.156	0.168	93.9	-0.040	-0.009 (30.1%)	0.156	0.168	93.9
Adapt 1	-0.082	-0.051 (1.7-fold)	0.121	0.121	92.8	-0.019	0.011 (-36.2%)	0.158	0.164	95.6	-0.025	0.006 (-18.4%)	0.140	0.140	95.3
Adapt 2	-0.088	-0.057 (1.9-fold)	0.123	0.123	92.5	-0.032	-0.001 (4.7%)	0.160	0.168	95.6	-0.033	-0.003 (9.4%)	0.143	0.143	95.3
Adapt 3	-0.093	-0.063 (2.1-fold)	0.127	0.126	93.3	-0.037	-0.007 (22.6%)	0.161	0.173	94.6	-0.039	-0.008 (27.8%)	0.149	0.153	94.4

Bias = average of the estimated log hazard ratios across simulations minus $\hat{\beta}^*$; Relative bias with respect to $\hat{\beta}^*$ reported for values in (-100,100) and "-fold" (higher) for values >100; SE = average of the standard error estimates across simulations; SD = empirical standard deviation of the estimated log hazard ratios across simulations; CP = confidence interval coverage probability.

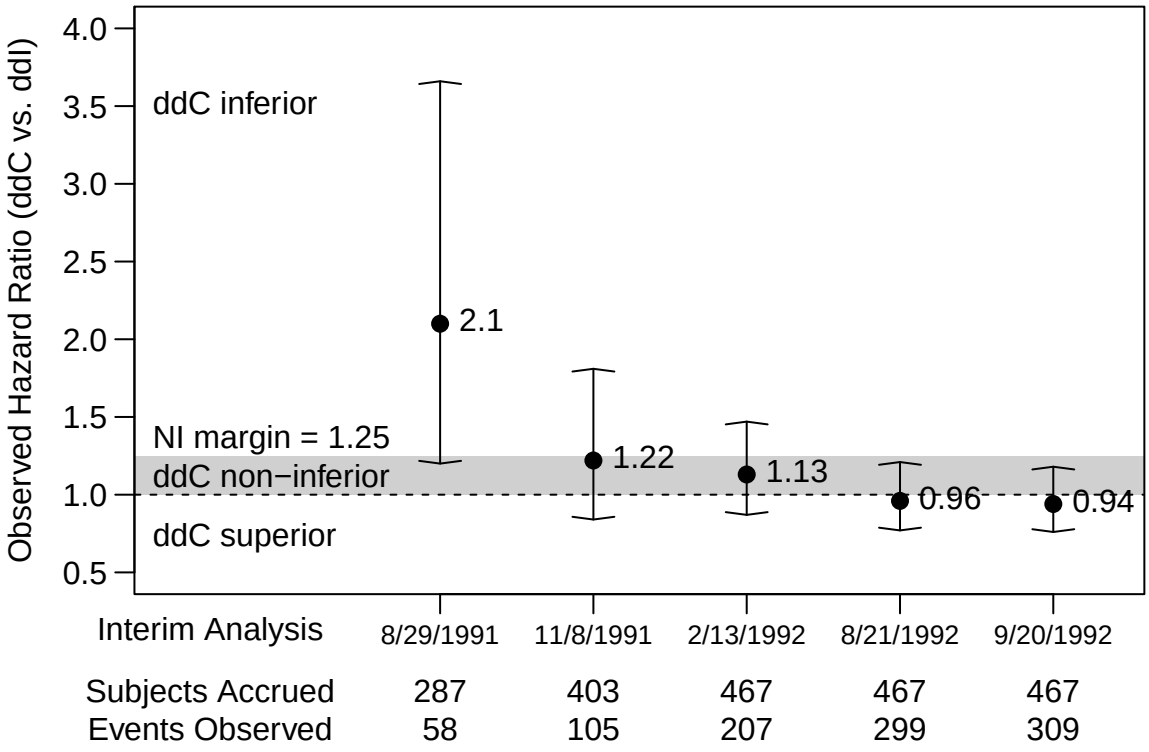
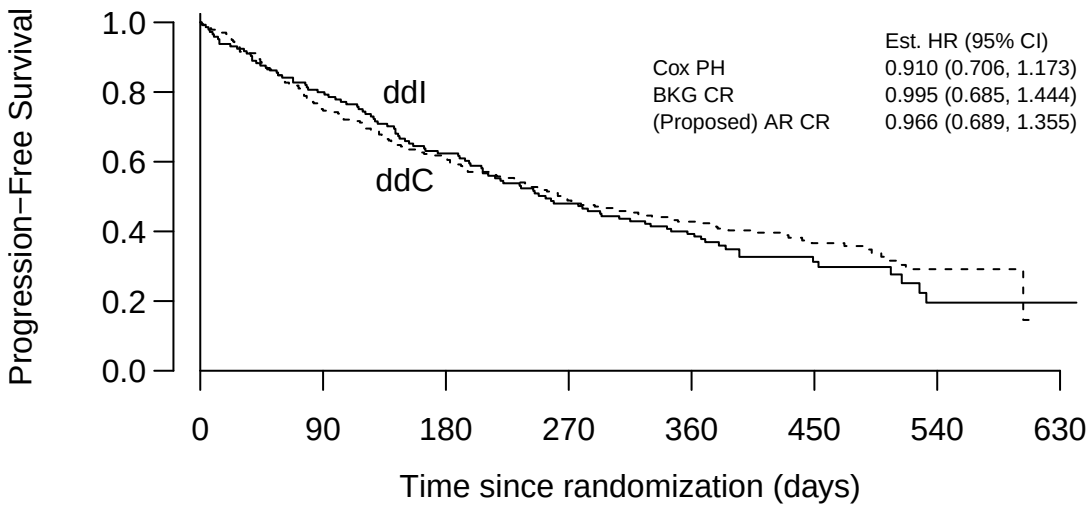


Figure 4.3: Estimated hazard ratio with 95% confidence interval at each interim analysis illustrating a time-varying effect on progression-free survival for zalcitabine (ddC) versus didanosine (ddI) in CPCRA Trial 002.

jective was to determine whether ddC did not exceed a 25% higher risk of progression to acquired immunodeficiency syndrome (AIDS) or death (i.e., progression-free survival, PFS) compared to ddI. A data monitoring committee monitored this trial for safety and efficacy at pre-specified interim analyses. The estimated hazard ratio varied across interim analyses indicating non-proportional hazards (see Figure 4.3).

To illustrate using censoring-robust methods compared to the Cox model in a time-to-event trial with adaptive randomization that exhibits time-varying effects, we induced an adaptive randomization scheme using the CPCRA Trial 002 data. We split the accrual period into two subperiods at the time that the 58th PFS event occurred. The first 247 participants were allocated to ddC versus ddI with 2:1 randomization (accrual subperiod 1) and the



N At Risk (Events)

ddI	145 (0)	88 (54)	54 (86)	6 (99)
ddC	237 (0)	142 (92)	98 (134)	14 (151)
Total	382 (0)	230 (146)	152 (220)	20 (250)

Figure 4.4: Analysis of CPCRA Trial 002 data after inducing an adaptive randomization scheme. Kaplan-Meier estimates for the progression-free survival distributions by treatment arm along with the estimated hazard ratio (HR) and corresponding 95% confidence interval (CI) for the Cox proportional hazards (PH) estimator, the Boyd-Kittelson-Gillen (BKG) censoring-robust (CR) estimator, and our proposed adaptive randomization (AR) CR estimator. As in the original trial, here the non-inferiority margin is 1.25.

remaining 135 participants were allocated with 1:1 randomization (accrual subperiod 2). In total, 382 participants were randomized (237 to ddC, 145 to ddI) and 251 had a PFS event. Figure 4.4 displays the PFS curves by accrual period and treatment arm, indicating non-proportional hazards. We estimated the hazard ratio comparing ddC versus ddI using the Cox proportional hazards (PH) estimator to be 0.910, the Boyd-Kittelson-Gillen (BKG) censoring-robust estimator (CRE) to be 0.995, and our proposed adaptive randomization (AR) CRE to be 0.966. The upper bound of the 95% confidence interval for each CRE is greater than the non-inferiority margin of 1.25. Hence, in this scenario, we cannot rule out an excess of 25% risk of progression to AIDS or death for ddC compared to ddI.

4.5 Discussion

We examined the impact of adaptive randomization on the underlying censoring distributions in time-to-event clinical trials where a scientific estimand for treatment efficacy is a marginal hazard ratio in the absence of intermittent censoring over the support of observed times, $\theta^* \equiv \exp(\beta^*)$. First, we showed that adaptive randomization can alter censoring patterns and can lead to biased estimates for θ^* in the presence of time-varying effects. Next, we proposed a novel censoring-robust estimator that extends existing methods by accounting for an adaptive randomization rule and removes dependence on censoring patterns, thus yielding a consistent estimate for θ^*. Finally, using data from the CPCRA Trial 002 where we induced an adaptive randomization scheme, we illustrated how adaptive randomization can yield different estimates of treatment efficacy on progression-free survival that depended on the estimation procedure used. In the scenario considered, we found that the Cox proportional hazards estimator overestimated efficacy compared to censoring-robust estimation approaches.

Time-varying treatment effects can arise in superiority (as shown in our simulation studies) and non-inferiority designs (as in the CPCRA Trial 002). Model misspecification, such as fitting a Cox proportional hazards model in the presence of time-varying effects, leads to hazard ratio estimates that depend on censoring patterns. This restricts the generalizability of inference from a trial and can make comparisons of hazard ratio estimates across trials challenging as censoring patterns are likely to differ across trials. By considering an estimand that does not depend on a trial's censoring patterns (e.g., θ^*), this can allow more fair comparisons provided that the estimation procedure yields consistent estimates for the scientific estimand.

We found that the extent of bias depended on the estimand and the data generating mechanism. In practice, we are unable to know the true data generating mechanism for time-to-

event endpoints. Instead, it behooves us to pre-specify statistical methods that can yield reliable estimates of treatment efficacy with as few assumptions as possible. Censoring-robust estimation supports this objective. Because we need to pre-specify the estimand, study design, and statistical analysis methods, it would not be possible to know the direction of bias we may have *a priori*. Furthermore, bias in either direction is not ideal. Overestimating a treatment effect can lead to an approval of a non-efficacious intervention, while underestimating can result in missing out on approving an intervention with truly favorable benefit-risk.

One limitation of our investigation is that we did not explicitly consider adaptive randomization as would be used in practice (e.g., restricted, covariate-adaptive, response-adaptive, or covariate-adjusted response-adaptive randomization). Instead, we assumed that the adaptive randomization rule was pre-specified according to accrual time (or after a number of events had been observed) that did not depend comparative data. This simplification of an adaptive randomization scheme was done to isolate the impact of adaptive randomization on estimating a treatment effect. We showed that even in this scenario, adaptive randomization affects censoring patterns and hence estimation of the scientific estimand θ^*. In practice, incorporating comparative data in the adaptive randomization rule would only exacerbate differential censoring patterns, further warranting consideration of censoring-robust estimation. A limitation of our estimator is reliance on the treatment-subperiod censoring distribution $S_C(t|Z = z, A = k)$. When adaptations occur more frequently the number of events within each treatmeant subperiod may be small. In such cases, the nonparametric Kaplan-Meier estimate may be highly variable. One alternative would be to use the BKG estimator. Since the BKG estimator uses the Kaplan-Meier estimates of $S_C(t|Z = z)$ over the entire support, the estimate of the weight would likely be more stable while still targeting θ^*.

Censoring-robust estimation can facilitate reliably estimating treatment efficacy in fixed

sample time-to-event trials for evaluating benefit-risk of candidate interventions. In practice, as with the CPCRA Trial 002, a sequential monitoring plan may be pre-specified to allow for the possibility of stopping a trial early for efficacy, futility, or harm. This can affect the timing of analyses, estimation of treatment efficacy, and inference. Investigating the role of censoring-robust estimators in this case is a focus of the next chapter.

Chapter 5

Information Growth of Censoring-Robust Estimators in Group Sequential Time-to-Event Clinical Trials with Adaptive Randomization

SUMMARY: Adaptive randomization (AR) allows for changes to the treatment allocation ratio during accrual of a clinical trial. In fixed sample time-to-event trials, AR alters treatment arm censoring patterns that can lead to biased efficacy estimates under time-varying treatment effects. Censoring-robust estimators (CREs) have been proposed to remove such dependence by reweighting the Cox proportional hazards estimating equation. CREs are consistent for an average hazard ratio over the support of observed times in the absence of intermittent censoring. For maximal information group sequential designs (GSDs) that may stop early for efficacy, futility, or harm, the estimand targeted by a CRE also depends on

the length of follow-up at an interim analysis (IA), as determined by the fraction of planned maximal information. Accurately estimating information growth is key to logistically plan the timing of IAs. In this chapter we quantify information growth of CREs, illustrating non-linear relationships with the observed number of events in GSDs with AR. We also show how to account for non-independent increments in such GSDs that together with the information growth allows for properly planning IAs while maintaining overall type I error rate.

5.1 Introduction

Group sequential designs (GSDs; Jennison and Turnbull, 1999) are frequently used to balance scientific, ethical, logistical, and statistical constraints to efficiently obtain statistically persuasive and clinically relevant evidence for answering a clinical trial's primary question. GSDs tend to be classified as "well-understood" by the U.S. Food and Drug Administration (2019a). Yet, with efforts to accelerate drug development through complex innovative designs (U.S. Food and Drug Administration, 2016; Dabrowska and Thaul, 2018) there is an increasing need by sponsors, data monitoring committees, scientific review boards, and regulators to understand the impact of such designs. As we saw in Chapter 4, a seemingly innocuous design feature as changing the treatment allocation probabilities based on an adaptive randomization scheme can change the target estimand unless appropriate methods (e.g., censoring-robust estimation) are pre-specified and employed. This was investigated in the previous chapter assuming a fixed sample time-to-event randomized clinical trial (RCT). In this chapter, we examine how to design a time-to-event GSD, with or without an adaptive randomization scheme, when the pre-specified analysis method calls for censoring-robust estimation to alleviate potential issues that arise in the presence of a time-varying treatment effect (i.e., non-proportional hazards).

GSDs allow for the possibility of stopping a trial prior to the maximal planned sample size

(events in a time-to-event trial) through repeated significance testing via analyses performed at a select number of times over the duration of a trial. Hence, the sequential sampling density requires modification compared to that for a fixed sample design. Consider a group sequential trial with up to J analyses. Let J denote the final planned analysis occurring at the maximal sample size and analyses $j = 1, ..., J-1$ be interim analyses. Let S_j denote the score statistic at analysis j where $S_j \dot{\sim} \mathcal{N}(\theta\mathcal{I}_j, \mathcal{I}_j)$ and $\mathcal{I}_j$ is the information at analysis j. Define the continuation set at analysis j, here based on the scale of the score statistic, by

$$\mathcal{C}_{S_j} = (a_{S_j}, b_{S_j}] \cup [c_{S_j}, d_{S_j})$$

where a, b, c, and d are the group sequential boundaries determined based on a combination scientific relevance and attaining trial design operating characteristics. In general, values between b and c boundaries represent equivalence. In this chapter, we consider superiority trials; hence no stopping for equivalence (i.e., $b = c$). Furthermore, we focus on testing a one-sided lower hypothesis where a hazard ratio less than one is favorable for the experimental treatment. That is, a boundaries correspond to efficacy and d boundaries to futility and the continuation set at analysis j expressed as a single interval, $\mathcal{C}_{S_j} = (a_{S_j}, d_{S_j})$. Recall from Chapter 2 that group sequential boundaries can be determined on a number of scales, including treatment effect, partial sum, and normalized Z statistic. Emerson et al. (2007) showed that there is a 1-to-1 mapping between these scales. An advantage of this relationship means that translating from one scale to another is not only possible, but may help instruct the scientific utility of proposed group sequential boundaries. For these reasons, it is often of interest to also examine proposed group sequential boundaries on the treatment effect estimate scale. This allows for a natural and direct way to interpret the clinical utility of such boundaries. In this chapter we will consider two scales: (1) the treatment effect scale (hazard ratio) to examine boundaries on the hazard ratio scale to assess clinical relevance; and (2) the normalized Z statistic scale to facilitate computation of GSD operating characteristics

across a range of scenarios without needing to calculate $\mathcal{I}_j$ directly in each scenario and analysis j.

With the score statistic S_j and the corresponding continuation set, $\mathcal{C}_{S_j}$, the trial stops if $S_j \in \mathcal{C}_{S_j}$; otherwise, the trial continues to the $j+1$ analysis (unless $j = J$). This implies that the trial stops at the first analysis when $S_j \notin \mathcal{C}_{S_j}$, denoted by $M = \min\left\{1 \leq j \leq J \ : \ S_j \notin \mathcal{C}_{S_j}\right\}$, and the corresponding S_j is denoted $S = S_M$ is the statistic at the Mth analysis. Together, M and S form the bivariate GSD test statistic ($M = m, S = s_m$). Assuming an *independent increments* structure holds (i.e., $S_{j+1} - S_j$ is independent of S_j), Armitage et al. (1969) defined the sequential sampling density for ($M = m, S = s_m$) recursively by

$$p(m, s_m; \theta) = \begin{cases} f(m, s_m; \theta) & \text{if } s_m \notin \mathcal{C}_{S_m} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

where

$$f(1, s_1; \theta) = \frac{1}{\sqrt{\mathcal{I}_1}} \, \phi\left(\frac{s_1 - \theta\mathcal{I}_1}{\sqrt{\mathcal{I}_1}}\right)$$

$$f(j, s_j; \theta) = \int_{\mathcal{C}_{S_{j-1}}} \frac{1}{\sqrt{v_j}} \, \phi\left(\frac{s_j - u - \theta v_j}{\sqrt{v_j}}\right) \, f(j-1, u; \theta) du$$

for $j = 2, \ldots, m$ where $v_j = \mathcal{I}_j - \mathcal{I}_{j-1}$ and $\phi(\cdot)$ is the standard normal density function. Integrating over this sequential sampling density allows computation of confidence intervals and p-values, and statistical operating characteristics to evaluate GSDs. Importantly, formulation of the group sequential sampling density (5.1) depends upon the variance of the score statistic (i.e., information, $\mathcal{I}_j$) for $j = 1, ..., m$. This means that if the observed $\mathcal{I}_j$ during the trial differs from what was assumed during the planning stage, the GSD operating characteristics will change. Therefore, it is imperative to understand how the information of a pre-specified statistic changes over the duration of the trial: the *information growth*.

Additionally, the timing of analyses and group sequential boundaries are a function of proportion of maximal information at time j, denoted by $\Pi_j = \mathcal{I}_j/\mathcal{I}_J$. In this chapter, we refer to information growth as how the set of $\{\Pi_j\}_{j=1}^J$ grow as data accumulates during the course of a trial compared to the proportion of maximal events, D_j/D_J, as is typically reported. Kittelson and Emerson (1999) coalesced many of the typical GSDs into a unified family. The unified family of GSDs parametrization allows for an infinite number of group sequential boundaries, including many traditional and commonly used boundary shapes (e.g., Pocock, 1977; O'Brien and Fleming, 1979; Whitehead and Stratton, 1983; Wang and Tsiatis, 1987). In practice, the timing of these analyses should balance scientifically relevant timing (the amount of follow-up for enrolled participants) and the level of statistical evidence desired to potentially stop a clinical trial before the maximal planned sample size. In time-to-event trials, 'sample size' typically refers to the number of events because events drive the information.

Because in practice, the timing of interim analyses may not occur at exactly the planned times, either operationally or because a data monitoring committee may request additional unplanned reviews of safety and possibly efficacy data, flexibility in modifying the group sequential boundaries can help to maintain trial design operating characteristics and therein the trial objectives. Approaches include those based on type I and/or type II error spending (Lan and DeMets, 1983; Pampallona et al., 1995; Chang et al., 1998). Another approach was proposed by Burington and Emerson (2003) called constrained boundaries. This latter approach allows for recalculating future group sequential boundaries while anchoring boundaries from already completed analyses all while maintaining trial design operating characteristics. This can lead to changes to the maximal sample size (or events), for example, when the observed information growth during the monitoring of a trial differs from that assumed at the design stage.

Often, the timing of analyses in a time-to-event GSD is based on the proportion of maximal

events. For a time-invariant treatment effect (e.g., proportional hazards), it has been shown that the statistical information of the partial likelihood score test (e.g., logrank test when there are two treatment arms) grows linearly with respect to the proportion of maximal events. Because group sequential boundaries are a function of the proportion of maximal information, in such a setting, timing of interim analyses can be determined by projecting the calendar time when the number of events will occur for each analysis. When there is a time-varying treatment effect (i.e., non-proportional hazards), however, information growth is no longer trivially proportional to the number of events observed (see, for example, Gillen and Emerson (2005)).

Non-proportional hazards with time-to-event primary endpoints has led to work on weighted statistics (in part to address what is of scientific importance: not all events have the same weight), included weighted logrank statistics. Previous research has investigated the impact of using weighted logrank statistics in the group sequential testing framework. Gillen and Emerson (2005) showed that weighted logrank statistics have non-linear information growth. Brummel and Gillen (2014) used a constrained boundaries approach by incorporating the non-linear information growth of weighted logrank statistics in the monitoring of time-to-event GSDs. An important takeaway from previous research for time-to-event GSDs is the role of the censoring distribution on (i) estimation under non-proportional hazards and (ii) and then because of the weighting function employed, on information growth even under proportional hazards. With this prior research as motivation, we extend the work of censoring-robust estimation in the fixed sample setting to the group sequential setting where targeting a marginal hazard ratio standardized to a common censoring distribution is of interest to enhance replicability and generalizability of inference from a trial.

Recently, adaptive randomization has become more common as a trial design feature. We showed in Chapter 4 how adaptive randomization can alter censoring patterns within a trial. We then proposed a censoring-robust estimator that incorporates the randomization scheme

via inverse probability of censoring weighting to standardize hazard ratio estimation to a marginal estimand that has a common censoring distribution (e.g., in the absence of intermittent censoring) yielding a more efficient estimator under adaptive randomization compared to the existing Boyd et al. (2012) censoring-robust estimator. Under fixed randomization, our adaptive randomization censoring-robust estimator reduces to that proposed by Boyd et al. (2012) that weights by the inverse of Kaplan-Meier estimates for treatment-specific censoring distributions.

In this chapter, we extend the work by Gillen and Emerson (2005) to censoring-robust estimators with adaptive randomization to characterize information growth and design GSDs with desired operating characteristics at the planning stage. The results presented also apply to cases of non-adaptive randomization. In Section 5.2 we show that the information growth of our adaptive randomization censoring robust estimator is non-linear in many practical settings owing to events occurring later receiving more weight when targeting a marginal hazard ratio estimand over the observed support. This can affect timing of analyses. In Section 5.2.3 we explore how to circumvent the censoring-robust estimator violating independent increments used in practice by Murray and Tsiatis (1999) that we use as a remedy in the adaptive randomization GSD setting. In Section 5.4 we illustrate how the frequentist operating characteristics of a GSD are not maintained when analysis times correspond to incorrect information fractions are performed (i.e., naively assuming information grows proportional to events). Further, we illustrate our proposed procedures from Sections 5.2 and 5.2.3 to modify the timing of analyses and account for non-independent increments to maintain the overall type I error rate for a GSD. Finally, we conclude this chapter with a discussion in Section 5.5.

5.2 Information growth of censoring-robust estimators

5.2.1 Censoring at interim analyses as a function of accrual

The time to an event in a clinical trial is right censored when the event does not occur by the last observed calendar time before a trial's database cutoff date for an analysis (whether interim or final). As previously discussed, censoring is either assumed to be independent of the true event times (i.e., independent censoring) or, more reasonably in practice, censoring is assumed to be independent of the true event conditional on the randomized treatment assignment (i.e., conditionally independent censoring). In these instances, censoring can then be parameterized according to trial accrual (entry times distribution), study dropout, and an administrative stopping of follow-up of participants (i.e., administrative censoring). Let E denote a random variable for the entry time of a participant during an accrual period from trial start until τ_A (with respect to calendar time since trial start). Let C_D denote a random variable for the dropout time of a participant (with respect to time since randomization). Let τ_j represent the administrative censoring time (with respect to calendar time since the start of the trial) for analysis j. Then, the administrative censoring time occurs at $\tau_j - E$. Since both accrual and dropout impact censoring, let $C \equiv (\tau_j - E) \wedge C_D \equiv \min(\tau_j - E, C_D)$. Thus, the cumulative distribution function of censoring is

$$F_C(t; \tau_j) \equiv \Pr[C \leq t; \tau_j] = \Pr[\min(\tau_j - E, C_D) \leq t].$$

Most often, study dropout is defined as withdrawal of consent or lost to follow-up. While it is possible to account for dropout through an assumed distribution, it is strongly advised to prevent, or at least minimize, the amount of dropout as much as possible during a trial (Council et al., 2010; Little et al., 2012; Fleming, 2011).

During study planning — when efforts to prevent study dropouts are being discussed —

it is of interest to at least evaluate candidate designs on the basis that there are no study dropouts. To this end, in this chapter we focus on scenarios in which there is no dropout (i.e., $C_D = \infty$) and censoring only occurs administratively at τ_j for analysis j. This implies that censoring depends on the administrative censoring time for analysis j and the accrual distribution of entry times, denoted by $C = \tau_j - E$. Furthermore, in the group sequential testing setting, analysis j may occur before trial accrual has completed (i.e., $\tau_j < \tau_A$).

We consider the weighting function used in our adaptive randomization censoring-robust estimator and then that for the Boyd et al. (2012) estimator. When assuming only administrative censoring, the treatment-subperiod-specific censoring distribution does not depend on the treatment arm and thus can be expressed as

$$\begin{aligned} S_C(t;\tau_j|Z=z,A=k) &= S_C(t;\tau_j|A=k) \\ &= \Pr[C>t|A=k,E<\tau_j] \\ &= \frac{F_E(\{\tau_j - t\} \wedge e_{k+1}) - F_E(e_k)}{F_E(\tau_j \wedge e_{k+1}) - F_E(e_k)} \end{aligned} \tag{5.2}$$

(see Appendix C.1 for derivation) where accrual in subperiod k corresponds to entry times (with respect to calendar time) between e_k and e_{k+1}. Note that for a randomization scheme in which the treatment arm allocation probability changes K times during accrual, define a K-partition of the accrual period $(0,\tau_A) = (e_0,e_1) \cup \{\cup_{k=2}^{K}[e_k,e_{k+1})\}$ where $e_1 = 0$ and $e_{K+1} = \tau_A$. The treatment-specific censoring distribution for our adaptive randomization censoring-robust estimator then takes the form

$$\begin{aligned} S_C^{\mathcal{AR}}(t;\tau_j|Z=z) &= \sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) S_C(t|Z=z,A=k) \\ &= \sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) S_C(t|A=k) \\ &= \sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) \left[\frac{F_E(\{\tau_j - t\} \wedge e_{k+1}) - F_E(e_k)}{F_E(\tau_j \wedge e_{k+1}) - F_E(e_k)}\right] \end{aligned} \tag{5.3}$$

with treatment-subperiod weight

$$w_k^{\mathcal{AR}}(z) \equiv \frac{\pi_{Z|A=k}(z)\pi_A(k)}{\sum_{l=1}^{K} \pi_{Z|A=l}(z)\pi_A(l)}$$

where $\pi_{Z|A=k}(z) \equiv \Pr[Z = z|A = k]$ is the conditional probability of being assigned to treatment arm z given study entry in accrual subperiod k and $\pi_A(k) = F_E(e_{k+1}) - F_E(e_k)$ is the proportion of accrual period in subperiod k. The marginal censoring distribution for our adaptive randomization censoring-robust estimator takes the form

$$\begin{aligned} S_C^{\mathcal{AR}}(t;\tau_j) &= \sum_{z=0}^{1} S_C^{\mathcal{AR}}(t;\tau_j|Z=z)\pi_Z(z) \\ &= \sum_{z=0}^{1}\sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) \left[\frac{F_E(\{\tau_j - t\} \wedge e_{k+1}) - F_E(e_k)}{F_E(\tau_j \wedge e_{k+1}) - F_E(e_k)}\right] \pi_Z(z) \\ &= \sum_{z=0}^{1}\sum_{k=1}^{K} \pi_{Z|A=k}(z)\pi_A(k) \left[\frac{F_E(\{\tau_j - t\} \wedge e_{k+1}) - F_E(e_k)}{F_E(\tau_j \wedge e_{k+1}) - F_E(e_k)}\right] \end{aligned} \tag{5.4}$$

where $\pi_Z(z) \equiv \Pr[Z = z] = \sum_{k=1}^{K} \pi_{Z|A=k}(z)\pi_A(k)$ is the marginal probability of being assigned to treatment arm z. Note that when there is a fixed randomization scheme (i.e., only one subperiod for the entire accrual period) and only administrative censoring at τ_j the treatment-subperiod-specific censoring distribution is equal to the marginal censoring distribution. Hence,

$$S_C(t;\tau_j) = \frac{F_E(\{\tau_j - t\} \wedge \tau_A)}{F_E(\tau_j \wedge \tau_A)} = \frac{F_E(\tau_j - t)}{F_E(\tau_j)} \tag{5.5}$$

with the latter equality since $F_E(\cdot)$ is a valid cumulative distribution function over the accrual period of $(0, \tau_A)$, where values greater than τ_A will return a value of one.

The inverse of $S_C(t;\tau_j)$ in (5.4) and (5.5) represent the weighting function used to define the statistical information at each analysis j of our adaptive randomization censoring-robust estimator and the Boyd et al. (2012) estimator, respectively. We next examine the infor-

mation growth. Then, in Section 5.3 we will address a consequence of the censoring-robust estimator weighting function depending on the timing at each analysis j, τ_j.

5.2.2 Statistical information at interim analyses

Tsiatis (1982) showed that a class of weighted statistics, including weighted logrank statistics and censoring-robust estimators, can be approximated by a sum of independent and identically distributed random variables. Furthermore, Tsiatis showed the form of the variance of such weighted score statistics (i.e., the statistical information) under the strong null (i.e., equality of survival distributions for event times) is

$$\sigma^2(\tau) \propto \int_0^{\tau} w^2(t) F_E(\tau - t)[1 - F_{C_D}(t)]\, dS_T(t). \tag{5.6}$$

where $w(t)$ is the weight function, $F_E(\cdot)$ is the cumulative distribution function of entry times, $F_{C_D}(\cdot)$ is the cumulative distribution function of dropout times, $S_T(\cdot)$ is the marginal survivor function of the true event times, and τ is the last observed event time for the analysis.

Since our adaptive randomization censoring-robust estimator, and that proposed by Boyd et al. (2012), are weighted statistics of the form in Tsiatis (1982), we can appeal to (5.6) to obtain the statistical information for censoring-robust estimators. Further, when we assume no dropout (i.e., $F_{C_D}(t) = 0$ for all $t < \infty$), the form of the statistical information of our adaptive randomization censoring-robust estimator when data is analyzed up to the last

observed event time t_j (with administrative censoring occurring at $\tau_j > t_j$) is

$$\begin{aligned}
\sigma_j^2(t_j) &\propto \int_0^{t_j} w^2(t) F_E(\tau_j - t)\, dS_T(t) \\
&= \int_0^{t_j} \left\{ \frac{1}{S_C^{\mathcal{AR}}(t;\tau_j)} \right\}^2 F_E(\tau_j - t)\, dS_T(t) \\
&= \int_0^{t_j} \frac{F_E(\tau_j - t)}{\left\{ \sum_{z=0}^{1} \sum_{k=1}^{K} \pi_{Z|A=k}(z) \pi_A(k) \left[\frac{F_E(\{\tau_j - t\} \wedge e_{k+1}) - F_E(e_k)}{F_E(\tau_j \wedge e_{k+1}) - F_E(e_k)} \right] \right\}^2} \, dS_T(t). \qquad (5.7)
\end{aligned}$$

Note that in this theoretical formulation (5.7), we make a distinction between the administrative censoring time τ_j and the last observed event time t_j such that $t_j < \tau_j$. This ensures that there are individuals still at risk by the last observed time (which is what would be expected in practice), and obviates issues with a weight approaching infinity when $F_E(0) = 0$ that would overly influence the statistical information. For a fixed randomization scheme, (5.7) reduces to

$$\begin{aligned}
\sigma_j^2(t_j) &\propto \int_0^{t_j} w^2(t) F_E(\tau_j - t)\, dS_T(t) \\
&= \int_0^{t_j} \left\{ \frac{1}{S_C(t;\tau_j)} \right\}^2 F_E(\tau_j - t)\, dS_T(t) \\
&= \int_0^{t_j} \left\{ \frac{1}{\frac{F_E(\tau_j - t)}{F_E(\tau_j)}} \right\}^2 F_E(\tau_j - t)\, dS_T(t) \\
&= \{F_E(\tau_j)\}^2 \int_0^{t_j} \frac{1}{F_E(\tau_j - t)}\, dS_T(t). \qquad (5.8)
\end{aligned}$$

5.2.3 Information growth of censoring-robust estimators

Information growth of an estimator is generally characterized as the proportion of maximal information (or information fraction) at analysis j, denoted by Π_j, as a function of the proportion of maximal events, D_j/D_J. Using the notation in (5.7), the proportion of maximal

information at analysis j is

$$\Pi_j \equiv \frac{\mathcal{I}_j}{\mathcal{I}_J} = \frac{\left(\frac{M_{j0}+M_{j1}}{M_{j0}M_{j1}}\right)\sigma_j^2}{\left(\frac{M_{J0}+M_{J1}}{M_{J0}M_{J1}}\right)\sigma_J^2} \tag{5.9}$$

where M_{jz} denotes the number initially at risk in treatment arm z for analysis $j = 1, ..., J$.

To characterize the information growth for censoring-robust estimators during the planning stage, we need to specify the hypothesized accrual patterns and survival distributions. To allow for a flexible range of accrual patterns, such as early, uniform, and late, we consider a powered uniform distribution for entry times. Specifically, let $E \sim PwrUnif(r, \tau_A)$ with cumulative distribution function

$$F_E(t) = \left(\frac{t}{\tau_A}\right)^r \cdot I(0 \leq t \leq \tau_A) \tag{5.10}$$

for accrual over the interval $(0, \tau_A)$. Additionally, the marginal survival distribution for a known randomization scheme (fixed or adaptive) can be expressed as

$$S_T(t) = \sum_{z=0}^{1} S_{T|Z=z}(t)\pi_Z(z) = \sum_{z=0}^{1} S_{T|Z=z}(t) \left\{ \sum_{k=1}^{K} \pi_{Z|A=k}(z)\pi_A(k) \right\} \tag{5.11}$$

where $\pi_Z(z) \equiv \Pr[Z = z]$ and $\pi_{Z|A=k}(z) \equiv \Pr[Z = z|A = k]$ are the marginal probability of being assigned to treatment arm z and conditional probability of being assigned to treatment arm z given study entry in accrual subperiod k, respectively (similar to Chapter 4 notation). Note that under the strong null, $S_T(t) = S_{T|Z=z}(t)$ for $z = 0, 1$ and all $t > 0$. In this chapter, we consider an adaptive randomization scheme over three years with three subperiods with allocation probability for treatment (tx1) is 0.50 (1:1 randomization) during the first year, 0.65 (∼2:1 randomization) during the second year, and 0.80 (4:1 randomization) during the final year of accrual; see Figure 5.1.

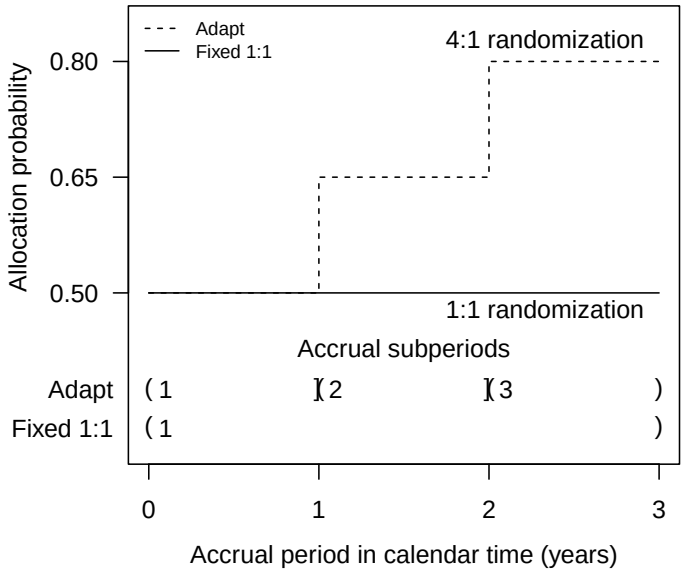


Figure 5.1: Fixed 1:1 vs. an adaptive randomization scheme over a three-year accrual period.

Calculating empirical information growth for censoring-robust estimators

We outline steps to estimate the empirical information growth for our adaptive randomization censoring-robust estimator as follows.

Step 1. Consider the maximal number of events to observe, D_{max}. Generate a single clinical trial data set (with simulation index by b), denoted here by `DATASET_b` consisting of $N > D_{max}$ participants to randomize according to a known randomization scheme and have variables for a subject identifier, entry time E (calendar time), treatment arm Z, true event time T (time since randomization), and true calendar event time T_{CAL} (calendar time). Specify an assumed accrual distribution of entry times, denoted by $F_E(t)$ (e.g., powered uniform as in (5.10)). Based on E and the randomization scheme, generate Z from a multinomial distribution with treatment arm allocation probabilities according the the randomization scheme. For a two-arm trial, without loss of generality, for each participant generate a Bernoulli random variable with probability

of 'success' equal to the allocation probability for the experimental arm; Z is then an indicator for the experimental arm. Also, specify assumed treatment-specific event times distributions, $S_T(t|Z)$, in accordance with the marginal hazard ratio of interest.

Step 2. Order `DATASET_b` according to T_{CAL} in ascending order: `ORD_DATASET_b`.

Step 3. Consider the minimum number of events to analyze the data, denoted D_{min} (e.g., 35 events), and the maximum number of events, D_{max}. Data will then be analyzed from `ORD_DATASET_b` after each event is observed, starting with D_{min} up to D_{max}. Let D_m be the mth ordered event in the set of events $\{D_{min}, ..., D_{max}\}$.

Step 4. Start with $m = 1$. Let the corresponding calendar time when D_m events have occurred be denoted by τ_m. Next, subset `ORD_DATASET_b` to only those participants entered the trial by τ_m (`ORD_DATASET_b_Dm`) and calculate the corresponding observed time $X = \min(\tau_j - E, T)$ and event indicator $\delta_1 = I(X = T)$. Fit the data using our adaptive randomization censoring-robust estimator. Store the estimated information by taking the inverse of the estimated variance, denoted here as $\hat{I}^{\mathcal{AR}}_{mb}$ where the b will correspond to the simulation index ($b = 1, ..., B$).

Step 5. Repeat Step 4 for $m = 2, ..., M$. For a given simulated clinical trial dataset `DATASET_b`, $\vec{\hat{I}}^{\mathcal{AR}}_b$ is a vector of M estimated informations with the mth entry corresponding to D_m events.

Step 6. Repeat Steps 1 – 5 ($B-1$) times (where B is the total number of simulated datasets, e.g., 100) in order to obtain a matrix of estimated information at each of $D_{min}, ..., D_{max}$ events (M columns) with B replications along the rows, denoted by $\hat{\mathbf{I}}^{\mathcal{AR}}_{B\times M}$.

Step 7. Fit a lowess smooth for estimated information (y-axis) against number of events (x-axis). From the resulting fitted values, map information to proportion of maximal information (y-axis) and events to proportion of maximal events (x-axis) to obtain the resulting information growth curve.

The above steps outline how to empirically calculate the information growth during the planning stage based upon assumptions for accrual pattern, the randomization scheme, survival times. The above can be modified to include the possibility of dropout with changes to Step 1 (adding generating of a true dropout time) and Step 4 (incorporating dropout as possibility in calculating the observed time and event indicator). After initiation of a GSD, the above steps can be modified to re-estimate the information growth during the trial. This would require an estimate of the accrual distribution, the known randomization scheme, and an estimate of the pooled survival distribution for event times.

In what follows, we focus on characterizing the information growth of our adaptive randomization censoring-robust estimator at the planning stage. We consider several GSDs with varying design features.

Characterizing information growth of censoring-robust estimators

Figure 5.2 shows the empirical information growth of our adaptive randomization censoring-robust estimator under the strong null with Exponential(rate=0.30) true event times for 600 randomized participants accrued uniformly over three years. The information growth curve is based on fitting a lowess smooth over the estimated information (inverse of the variance estimates) from 100 simulated datasets. For each simulated dataset, we estimated the variance of the adaptive randomization censoring-robust estimator, and then the information, once 35 events occurred until the maximal number of events ($D_J = 379$) occurred. In this setting, we found that the information growth is non-linear with the information growing slower than the number of events. As mentioned earlier in the chapter, group sequential boundaries are a function of the proportion of maximal information, Π_j. For a group sequential design with four equally spaced analyses with respect to information time, the analyses should occur once $\Pi_j \in \{0.25, 0.50, 0.75, 1.00\}$. If using a logrank statistic under proportional hazards, the timing of these analyses would correspond to proportion of

maximal events $D_j/D_J \in \{0.25, 0.50, 0.75, 1.00\}$. When using our adaptive randomization censoring-robust estimator, however, we see from Figure 5.2 that at 25% maximal events, there is only 21% of the maximal information. Instead, the first interim analysis should occur after approximately 30% events have occurred to have 25% of the maximal information. Similarly in this scenario, the second and third interim analyses would be projected to occur after approximately 58% and 80% events have occurred.

Figure 5.3 again displays the information growth for 600 randomized participants, but with early (left) or late (right) accrual over three years. For the early accrual scenario, as compared to the uniform accrual in Figure 5.2, the information growth attenuates towards a linear growth where there is less deviation from information growing proportional to the number of events. On the other hand, the late accrual scenario yields a slower information growth, suggesting that not all events equally contribute to the estimation of the hazard ratio based on adaptive randomization censoring-robust estimator. For example, the first interim analysis would occur after 37% of events occurred as compared to after 30% events occurred assuming uniform accrual or 27% events occurred assuming early accrual. These results illustrate the sensitivity of the information growth for our censoring-robust estimator to the accrual pattern. Hence, during monitoring of a trial, re-estimation of the information growth using available pooled data can facilitate more accurate projections of the timing of subsequent analyses.

Figure 5.4 displays the information growth for 400 (left) and 2000 (right) randomized participants uniformly accrued over three years, both with the same maximal number of events $D_J = 379$. With 400 randomized participants, the time to observe the maximal number of events will naturally take longer compared to with 2000 randomized participants. With the larger sample size, events tend to occur in a narrower interval in time resulting in events contributing nearly equally. With the smaller sample size, the information growth is slower since the later occurring events will have a larger weight in the re-weighted estimating equa-

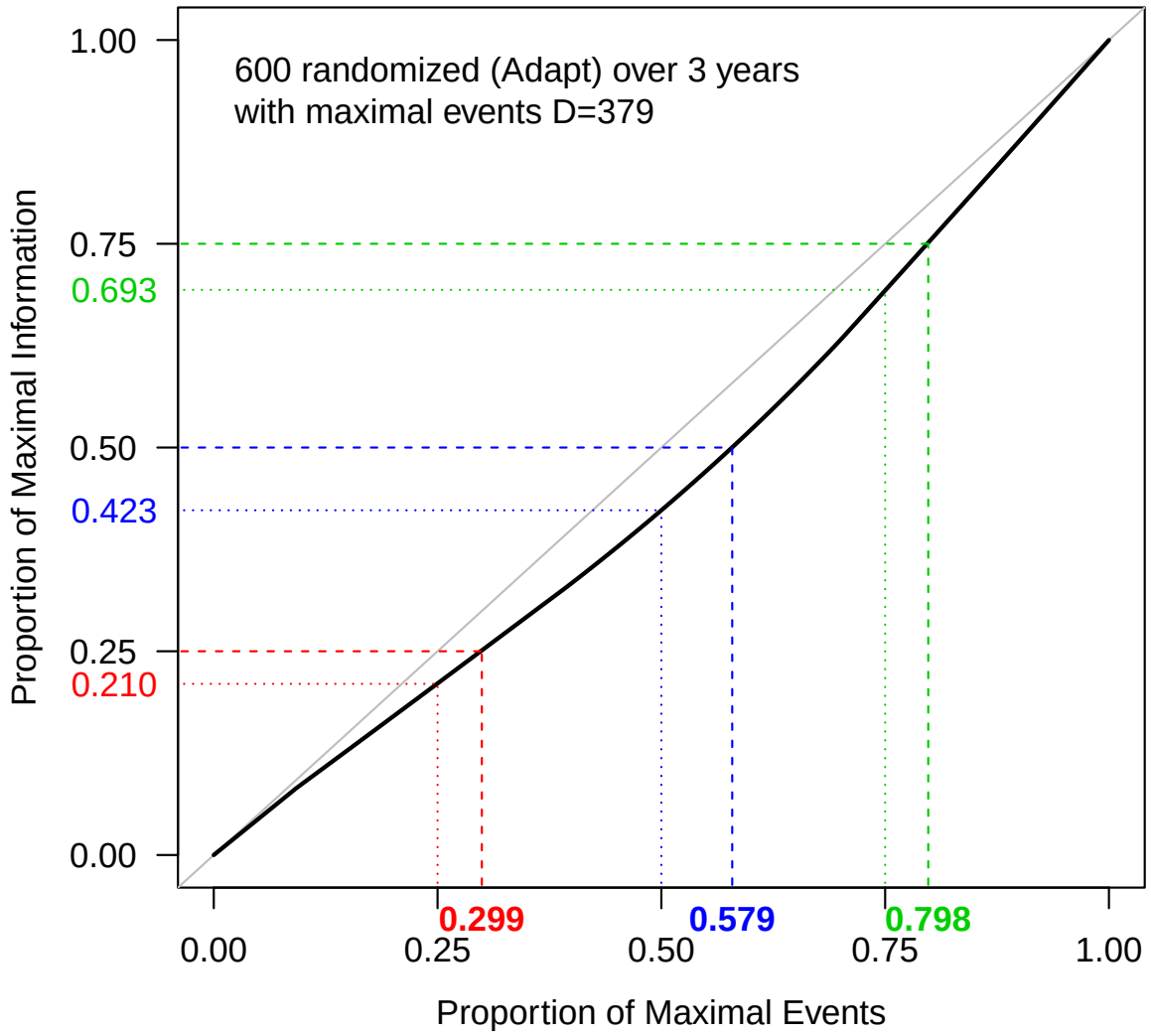


Figure 5.2: Empirical information growth for our adaptive randomization censoring-robust estimator (black curve) under the strong null with Exponential(rate=0.30) true event times for uniform accrual of 600 participants over 3 years. Gray solid line indicates proportion of maximal information equals proportion of maximal events, D_j/D_J. Dotted lines map equally-spaced proportion of maximal events (0.25, 0.50, 0.75) to the respective proportion of maximal information, Π_j. Dashed lines map equally-spaced Π_j to the respective D_j/D_J.

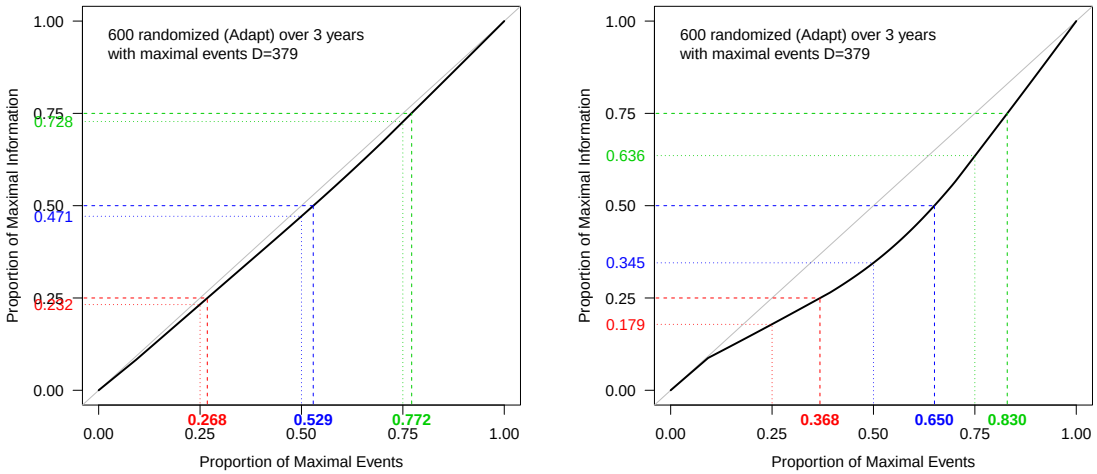


Figure 5.3: Empirical information growth for our adaptive randomization censoring-robust estimator (black curve) under the strong null with Exponential(rate=0.30) true event times for early ($r = 0.5$) and late ($r = 2.5$) accrual of 600 participants over 3 years, respectively. Gray solid line indicates proportion of maximal information equals proportion of maximal events. Dotted lines map equally-spaced proportion of maximal events (0.25, 0.50, 0.75) to the respective proportion of maximal information, Π_j. Dashed lines map equally-spaced Π_j (0.25, 0.50, 0.75), to the respective proportion of maximal events.

tion for the adaptive randomization censoring-robust estimator. In the scenarios examined, we found that randomizing a larger number of participants more quickly can yield a linear information growth for our censoring-robust estimator. The goal in characterizing the information growth, however, was not to to identify when there may be linear information growth. Instead, during the planning of a group sequential trial, characterizing the information growth under plausible scenarios that may arise will allow for collaborative discussions at the design stage, including but not limited to trial operations (e.g., activating sites) and timing of interim analyses (e.g., planning reviews by a data monitoring committee).

In this section, we focused on information growth and mapping proportion of maximal information to the respective proportion of maximal events so that timing of interim analyses matches with those for the group sequential boundaries. In the next section we will focus on the violation of independent increments that, unless accounted for, can affect the overall type I error rate for a GSD.

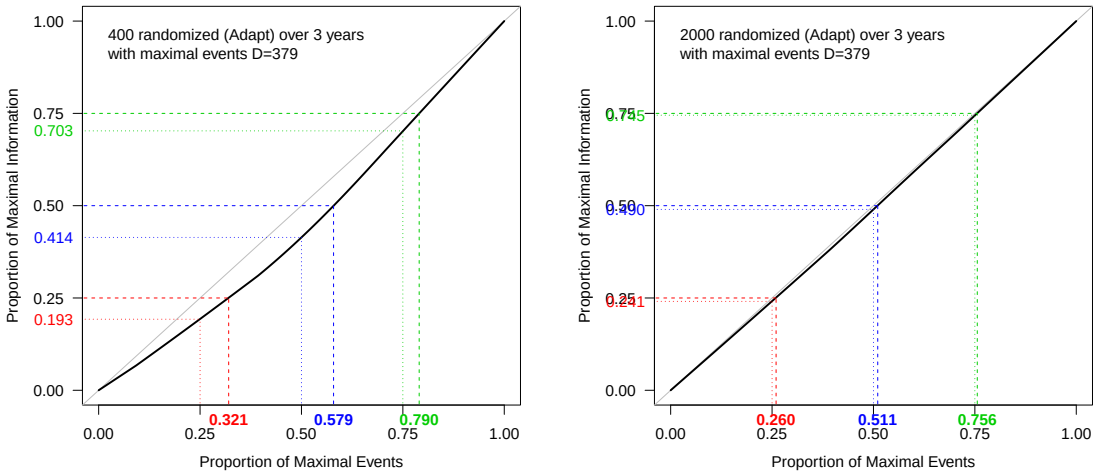


Figure 5.4: Empirical information growth for our adaptive randomization censoring-robust estimator (black curve) under the strong null with Exponential(rate=0.30) true event times for uniform accrual of 400 and 2000 participants, respectively, over 3 years. Gray solid line indicates proportion of maximal information equals proportion of maximal events. Dotted lines map equally-spaced proportion of maximal events (0.25, 0.50, 0.75) to the respective proportion of maximal information, Π_j. Dashed lines map equally-spaced Π_j (0.25, 0.50, 0.75), to the respective proportion of maximal events.

5.3 Violating independent increments

5.3.1 Censoring-robust estimators induce non-independent increments

Independent increments allows for tractable computation of the sequential sampling density via Armitage et al. (1969) used at the planning stage to calculate group sequential boundaries and design operating characteristics. However, not all statistics follow an independent increments structure. In particular, it has previously been shown that there are weighted statistics that violate the independent increments structure (e.g., the modified Wilcoxon statistic (Slud and Wei, 1982)). Tsiatis (1982) noted that, for weighted score statistics, when the weighting function converges to a quantity that depends on the analysis time, asymptotic independent increments does not hold.

Recall, the weighting function of our adaptive randomization censoring-robust estimator in a group sequential trial takes the form

$$\hat{W}^{\mathcal{AR}}(t;\tau_j) = \frac{1}{\hat{S}_C^{\mathcal{AR}}(t;\tau_j|Z)}$$

and $\hat{W}(t;\tau_j) = 1/\hat{S}_C^{KM}(t;\tau_j|Z)$ for the Boyd et al. (2012) censoring-robust estimator. For both censoring-robust estimators, the estimator for the treatment-specific survivor function of censoring converges in probability to $S_C(t;\tau_j|Z)$ that depends on the timing of the analysis at calendar time τ_j. This implies that the targeted weighting function differs from analyses before and after accrual ends. In particular, (Boyd, 2009) noted that after accrual has ended the weighting function of the Boyd et al. (2012) censoring-robust estimator at analysis time τ_j dominates the corresponding weighting function at a later analysis time $\tau_{j\prime}$. Since the weighting function of our adaptive randomization censoring-robust estimator consistently estimates the same $W(t;\tau_j)$ at analysis time τ_j, after accrual ends, $\hat{W}^{\mathcal{AR}}(t;\tau_j)$ dominates $\hat{W}^{\mathcal{AR}}(t;\tau_{j\prime})$. Hence, from Tsiatis (1982), asymptotic independent increments does not hold for our adaptive randomization censoring-robust estimator.

Violating the independent increments structure implies that direct application of Armitage et al. (1969) can yield an erroneous sequential sampling density and operating characteristics. In practice, there are typically two approaches to remedy non-independent increments: (1) computing the sequential sampling density via a multivariate integration (e.g., MULNOR software (Schervish, 1984)); or (2) assume independent increments holds for the first $J-1$ analyses and then find the final (efficacy) boundary that maintains the overall type I error at the nominal level α (Murray and Tsiatis, 1999). We proceed with outlining an algorithm based on the approach by Murray and Tsiatis (1999) to maintain overall type I error rate for a group sequential time-to-event trial where the primary analysis uses our adaptive randomization censoring-robust estimator.

5.3.2 A remedy for non-independent increments

To account for the non-independent increments of our adaptive randomization censoring-robust estimator, we propose the following algorithm to use in practice. Steps 1-3 describe taking bootstrap samples from observed data at analysis J. Steps 4-8 apply the approach of Murray and Tsiatis (1999) in our context to find the final efficacy boundary at analysis J that maintains the overall type I error rate at the nominal level α.

To avoid needing to estimate the information at each analysis j, we compute the normalized Z statistic at each analysis and use the corresponding sequential sampling density based on the normalized Z statistic. We then use the sequential boundaries on the Z scale to calculate stopping probabilities.

Step 1. At analysis J, take a single bootstrap sample of length N (the number randomized by analysis J) with replacement from the observed $\mathcal{O} = \{(x_i, \delta_i, z_i, e_i)\}_{i=1}^{N}$ where x_i is the observed time, δ_i is the event indicator, z_i is the treatment arm assignment, and e_i is the entry time for participant i. Denote this bootstrapped sample by $\mathcal{B}_b^* = \{(x_i^*, \delta_i^*, z_i^*, e_i^*)\}_{i=1}^{N}$ with sample index b.

Step 2. For each analysis $j = 1, ..., J$, compute the corresponding Z statistic for the adaptive randomization censoring-robust estimator, $Z_{jb}^{*\mathcal{AR}}$, when administratively censoring at τ_j. For $\mathcal{B}_b^*$, there will be J Z-statistics $\{Z_{jb}^{*\mathcal{AR}}\}_{j=1}^{J}$.

Step 3. Repeat Steps 1 and 2 B times to obtain B Z-statistics at each analysis j. Store as a $B \times J$ matrix of $Z^{*\mathcal{AR}}$-statistics, denoted by $\mathbf{Z}_{B\times J}^{*\mathcal{AR}}$.

Step 4. Calculate the $J \times J$ variance-covariance matrix of $\mathbf{Z}_{B\times J}^{*\mathcal{AR}}$, denoted by $\mathbf{\Sigma}_{\mathbf{Z}^{*\mathcal{AR}}}$.

Step 5. Generate 10,000 $J \times 1$ random vectors from a J-dimensional multivariate normal distribution with mean vector $\vec{0}$ and variance-covariance matrix $\mathbf{\Sigma}_{\mathbf{Z}^{*\mathcal{AR}}}$. Store as a 10,000 $\times$ J matrix of $Z_{sim}^{*\mathcal{AR}}$-statistics, denoted by $\mathbf{Z}_{sim}^{*\mathcal{AR}}$.

Step 6. Use the original group sequential design's first $J-1$ boundaries to calculate the stopping probabilities for efficacy and futility at analyses $j = 1, ..., J-1$. Denote stopping probabilities for efficacy at these analyses ($j = 1, ..., J-1$) by $\{\alpha_j\}_{j=1}^{J-1}$. Note that the stopping probability at analysis J is denoted similarly by α_J.

Step 7. For a one-sided level α hypothesis test for efficacy, calculate the value of the stopping probability at analysis J such that the overall type I error rate is α. That is, $\alpha_J^* = \alpha - \sum_{j=1}^{J-1} \alpha_j$.

Step 8. Obtain the new final Z statistic-based efficacy boundary (which will be the same as the futility boundary in our group sequential setting) by calculating the empirical lower $\alpha_J^* \times 100$-percentile of the vector of simulated $Z_{sim}^{*\mathcal{AR}}$-statistics at analysis J (from Step 5), denoted by $a_{J,Z^*} = \mathbb{F}^{-1}_{\vec{Z}^{*\mathcal{AR}}_{J,sim}}(\alpha_J^*)$.

The above steps outline how to find the final efficacy boundary in practice. In the next section, we illustrate this process based on a hypothetical setting via simulation. For the hypothetical setting, Steps 1-3 are replaced by having simulated data for B group sequential clinical trials in which we have all corresponding Z statistics at each analysis $j = 1, ..., J$ prior to apply the group sequential stopping rule. Therefore, we can start with Step 4 and proceed to Step 8, as described above.

5.4 Timing of analyses and maintaining design operating characteristics

Consider a time-to-event randomized clinical trial (RCT) whose primary aim is to determine whether the hazard of an event (e.g., death) in the experiment arm is 30% lower than that for the control arm (i.e., hazard ratio, $\theta = 0.70$). As a hypothesis test, we are interested in

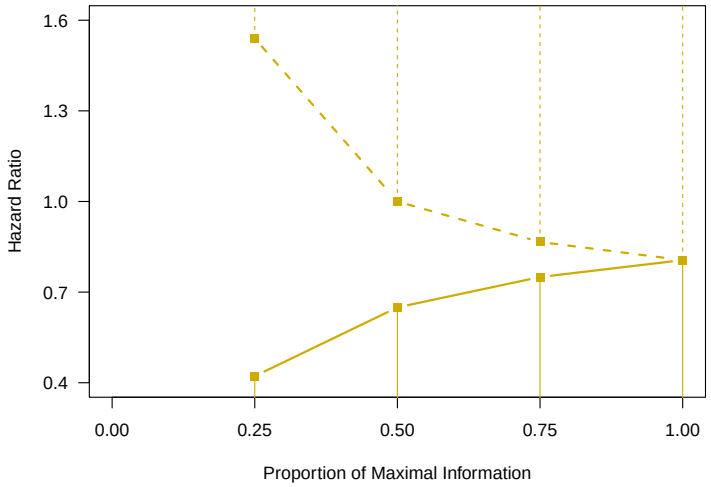


Figure 5.5: Group sequential design with symmetric O'Brien-Fleming boundaries assuming equally spaced analyses in information time. Solid lines connecting filled squares correspond to efficacy boundaries and dashed lines to futility boundaries.

a one-sided level $\alpha = 0.025$ test of the null $H_0 : \theta = 1$ versus the alternative $H_A : \theta < 1$ in which we have approximately 90% statistical power to reject the null when $\theta = 0.70$. For a fixed sample RCT, assuming a fixed 1:1 randomization scheme, 331 events would need to be observed during the trial to have 90% power to test the above hypothesis.

Alternatively, consider a group sequential design (GSD) with up to 4 analyses equally spaced in information time. That is, interim analysis 1 occurs when $\Pi_1 = 0.25$, interim analysis 2 when $\Pi_2 = 0.50$, interim analysis 3 when $\Pi_3 = 0.75$, and the final planned analysis ($J = 4$) when $\Pi_4 = 1$). Furthermore, suppose for this hypothetical setting, it is of interest to use a symmetric O'Brien-Fleming design for efficacy and futility. Then, this GSD (see Figure 5.5) would have efficacy boundaries corresponding to a hazard ratio of 0.4218, 0.6494, 0.7499, and 0.8059, at analysis 1, 2, 3, and 4, respectively. The corresponding futility boundaries would be for a hazard ratio of 1.5398, 1.0000, 0.8660, and 0.8059, respectively.

If we assume proportional hazards and choose to analyze data using the Cox proportional hazards model, information grows proportional to the the number of events. Suppose that

we are unsure if proportional hazards will hold during the course of the trial, and we want to implement the adaptive randomization scheme. Hence, we consider *a priori* using our adaptive randomization censoring-robust estimator.

While pre-specifying a censoring-robust estimator to guard against censoring-dependent estimation in the presence of non-proportional hazards, it may turn out once data has been collected that proportional hazards reasonably holds. Even in such an instance, we would not deviate from the pre-specified primary analysis plan to avoid data-driven decision making, as would be the stance mandated by regulatory agencies. While acknowledging this, we proceed with an investigation assuming proportional hazards (for the strong null, the design alternative, and an intermediate alternative). This will ensure that our target estimand remains the same at each interim analysis and also allow us to compare to the Cox proportional hazards model.

For illustrative purposes in the remainder of this section, we assume proportional hazards holds. This allows us to consider the group sequential boundaries from the original design that assumed proportional hazards to be equivalent when using a censoring-robust estimator. As a remark for use in practice, however, we would want to determine what boundaries would make sense in the presence of time-varying treatment effects. Additionally, here we are considering a GSD with symmetric O'Brien-Fleming boundaries for efficacy and futility. This is not customarily chosen in practice. It may be preferred to select an O'Brien-Fleming efficacy boundary and a Pocock futility boundary instead. One explanation for this is that since the O'Brien-Fleming boundary is conservative early, crossing the boundary earlier with respect to futility can result in establishing harm, which is much too late to stop. Efficacy, on the other hand, may not be conservative enough in certain time-varying treatment effect settings — part of what should determine the boundaries is discussion among stakeholders about the clinical relevance of boundaries, often corresponding to the treatment effect scale.

In what follows, we first consider the strong null hypothesis where the true event times for

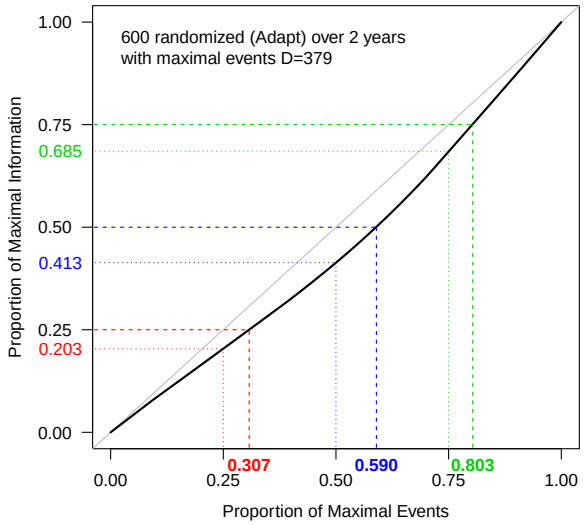


Figure 5.6: Empirical information growth for our adaptive randomization censoring-robust estimator (black curve) under the strong null with Exponential(rate=0.30) true event times for uniform accrual of 600 participants over 2 years with an adaptive randomization scheme (1:1 over first third, 0.65:0.35 over second third, and 4:1 for last third of accrual). Gray solid line indicates proportion of maximal information equals proportion of maximal events, D_j/D_J. Dotted lines map equally-spaced proportion of maximal events (0.25, 0.50, 0.75) to the respective proportion of maximal information, Π_j. Dashed lines map equally-spaced Π_j to the respective D_j/D_J.

experimental treatment and control arms both follow an Exponential(rate = 0.30) distribution. Further, we assume uniform accrual of 600 participants over two years and an adaptive randomization scheme that starts with 1:1 randomization over the first year followed by ∼2:1 (0.65:0.35) over the second year. There is a planned maximal number of events $D_J = 379$. Then, the information growth for our adaptive randomization censoring-robust estimator is non-linear (see Figure 5.6). This implies the timing of interim analyses will not align with the information fraction Π_j for which the group sequential boundaries are based.

Table 5.1 summarizes the operating characteristics for variations of the original GSD described above. First, the original GSD boundaries are displayed on the hazard ratio and the standardized Z statistic scales. The corrected final boundary J is also included (to be discussed below). The proportion of maximal events at each analysis are displayed based on (i) naively assuming information is proportion to events and (ii) mapping from the adap-

tive randomization censoring-robust estimator information growth. The remainder of the table summaries the stopping probabilities at each analysis for futility and efficacy, along with the corresponding power (cumulative stopping probability). The average number and 75th percentile of events observed by the time of stopping is summarized. These operating characteristics are provided for the Cox PH estimator, AR-CRE (naive), AR-CRE only correcting for AR-CRE information growth, and AR-CRE with correct information growth and corrected final bound (based on the procedure outlined in Section 5.3.2).

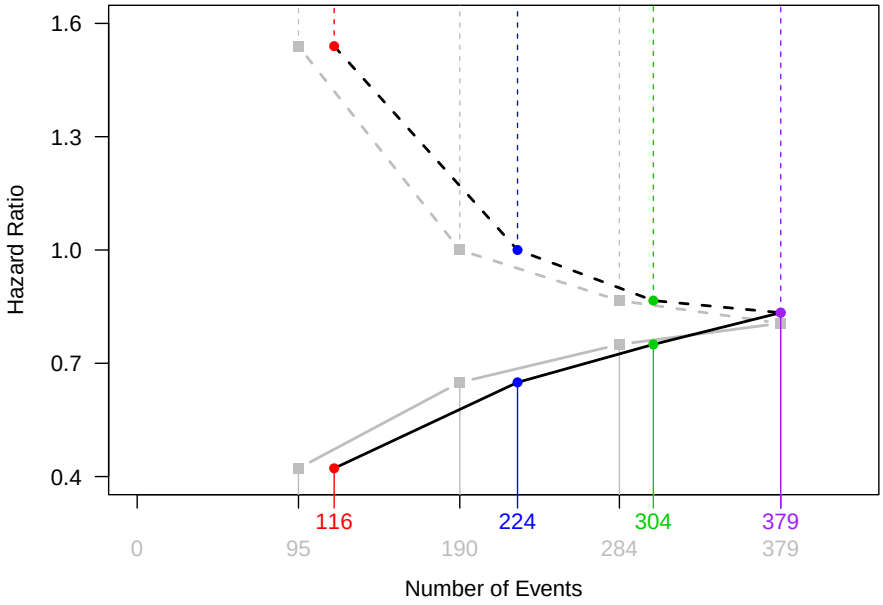
In Table 5.1 we show that when naively using the incorrect mapping of information growth to proportion of maximal events for our adaptive randomization censoring-robust estimator results in a type I error rate of 0.016, below the nominal level. Even after using the correct mapping from information growth to proportion of events, the type I error rate did not achieve the nominal level (in fact, remained the same in this scenario) because as we discussed in 5.3, censoring-robust estimators violate independent increments. To this end, we use the correct information growth and corrected final bound (displayed on the hazard ratio and Z scales in the table) to achieve the nominal level when using our adaptive randomization censoring-robust estimator. These group sequential boundaries are also visually depicted in the top plot of Figure 5.7. (For a scenario with a fixed randomization GSD, see Appendix Figure C.1 and Appendix Table C.1.)

While it is important to have a GSD with the appropriate type I error rate according to the level of the hypothesis test, examining the power under alternatives is important during the planning stage. Recall that the original GSD was designed to detect a hazard ratio of 0.70 with 90% power under proportional hazards when using the Cox PH estimator. As discussed in the fixed sample setting in Chapter 4, censoring-robust estimators will not be as efficient as the Cox estimator under proportional hazards. As such, we would expect to have lower power compared to the Cox PH estimator in this setting (which was $\sim$85.6%, though not a substantial loss of power for this scenario) even after the corrections made to timing

Table 5.1: Operating characteristics based on 2000 simulations for the group sequential design with symmetric O'Brien-Fleming boundaries with uniform accrual of 600 participants over 2 years and an adaptive randomization scheme (1:1 over the first year and 0.65:0.35 over the second year of accrual). Boundaries are displayed on the hazard ratio (HR) and Z statistic scales, along with the corrected final boundary. Proportion of maximal events at each analysis is summarized according to naively assuming information growing proportional to events and the empirical AR-CRE information growth from Figure 5.6. Operating characteristics include stopping probabilities at each analysis and power (cumulative probability of stopping) for futility and efficacy, along with the average and 75th percentile of events and maximum follow-up (years) at stopping. Power for efficacy in this setting is the overall type I error rate where nominal level is $\alpha = 0.025$, with Monte Carlo error (0.018, 0.032).

Analysis j	1	2	3	$4 = J$		
Boundaries at equally spaced Π_j on the hazard ratio scale at j						
					(J corrected)	
Futility	1.5398	1.0000	0.8660	0.8059	(0.8339)	
Efficacy	0.4218	0.6494	0.7499	0.8059	(0.8339)	
Boundaries at equally spaced Π_j on Z scale at j						
					(J corrected)	
Futility	-4.006	-2.833	-2.313	-2.003	(-1.687)	
Efficacy	2.003	0.000	-1.157	-2.003	(-1.687)	
Proportion of maximal events D_j/D_J at j with $D_J = 379$						
Naive: assume $\Pi_j \propto D_j$	0.250	0.500	0.750	1.000		
Map from AR-CRE Π_j	0.307	0.590	0.803	1.000		
	Stopping Probability at j				**Power**	**Events [Max F-U]**
Cox PH						
Futility	0.024	0.488	0.379	0.090	0.980	243 (284)
Efficacy	0.000	0.001	0.007	0.012	0.020	[2.8 (3.2)]
AR-CRE (naive)						
Futility	0.025	0.492	0.392	0.074	0.984	240 (284)
Efficacy	0.000	0.002	0.006	0.008	0.016	[2.8 (3.2)]
AR-CRE (only correct Π_j)						
Futility	0.024	0.474	0.411	0.076	0.984	268 (304)
Efficacy	0.000	0.001	0.004	0.011	0.016	[3.1 (3.4)]
AR-CRE (correct Π_j + corrected final bound)						
Futility	0.024	0.474	0.411	0.068	0.975	268 (304)
Efficacy	0.000	0.001	0.004	0.019	0.024	[3.1 (3.4)]

Events and maximum follow-up (F-U, in years) summarized with mean (75th percentile);
AR-CRE = adaptive randomization censoring-robust estimator



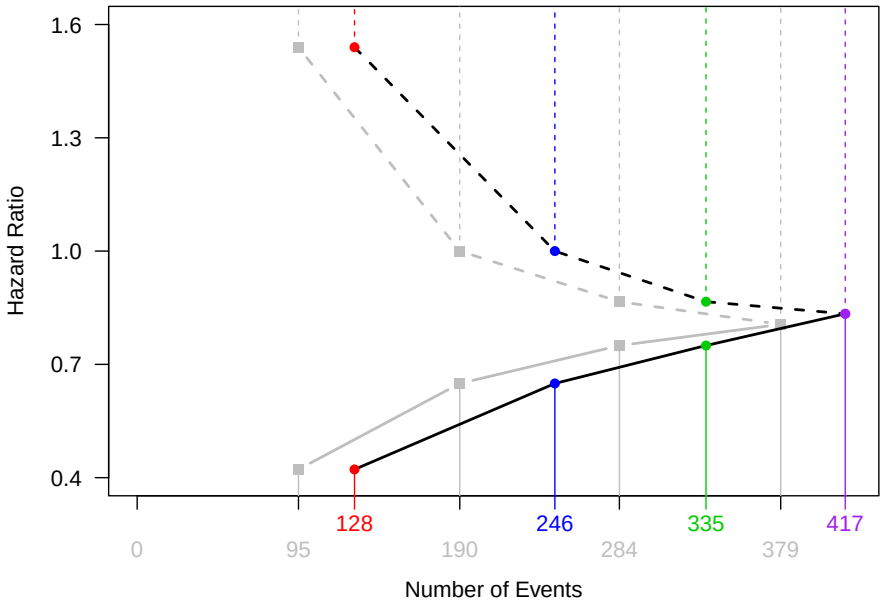


Figure 5.7: Group sequential design with symmetric O'Brien-Fleming boundaries assuming equally spaced analyses in information time with $D = 379$ maximal events (top) and $D_J^* = 417$ maximal events (bottom). Solid lines connecting filled squares correspond to efficacy boundaries and dashed lines to futility boundaries. Gray lines correspond original design for the Cox PH estimator assuming proportional hazards.

of analyses and overall type I error rate for the sample maximal number of events $D_J = 379$. Because the power is still below the desired level for the censoring-robust estimators, simply using the AR-CRE information growth with the same D_J is not adequate for the selected group sequential design using a censoring-robust estimator to maintain maximal information (statistical power). As we saw in Chapter 4, the larger variance of a censoring-robust estimator is a tradeoff when proportional hazards holds in order to *a priori* target the desired estimand while allowing for the possibility of non-proportional hazards. Therefore, in practice when designing a GSD, we would want to specify the D_J^* that maintains the desired power for the design alternative.

In order to have a maximal information GSD in this setting, we need to increase the maximal number of events to observe to $D_J^* = 417$ events in this setting to have 90% power to detect a hazard ratio of 0.70 when using our adaptive randomization censoring-robust estimator as the primary analysis method at each interim analysis. Table 5.2 summarizes the same operating characteristics now for the $D_J^* = 417$ maximal information GSD that maintains power for the design alternative. Further, note with the increase in the maximal number of events, average number of events at stopping is higher and the maximum follow-up at stopping is higher, compared to the $D_J = 379$ GSD. The corresponding group sequential boundaries are visualized in the bottom plot of Figure 5.7.

In the scenarios examined, we assumed proportional hazards. During the planning of a time-to-event GSD the potential for time-varying treatment effects may, however, need to be considered. In such cases, censoring-robust estimation is warranted to target a marginal estimand (standardized to a common censoring distribution) over the observed support. Attention towards maximal follow-up at stopping is critical to consider what is the target estimand and whether the GSD, including the sequential boundaries, adequately address the scientific, ethical, and logistical constraints to answer the trial's primary aim.

Table 5.2: Maximal information ($D_J^* = 417$) group sequential design with symmetric O'Brien-Fleming boundaries with uniform accrual of 600 participants over 2 years and an adaptive randomization scheme (1:1 over the first year and 0.65:0.35 over the second year of accrual). Boundaries are displayed on the hazard ratio and Z statistic scales, along with the corrected final boundary. Proportion of maximal events at each analysis is summarized according to naively assuming information growing proportional to events and the empirical AR-CRE information growth from Figure 5.6. Operating characteristics include stopping probabilities at each analysis and power (cumulative probability of stopping) for futility and efficacy, along with the average and 75th percentile of events and maximum follow-up (years) at stopping. Power for efficacy under null is the overall type I error rate where nominal level is $\alpha = 0.025$, with Monte Carlo error (0.018, 0.032).

Analysis j	1	2	3	$4 = J$		
Boundaries at equally spaced Π_j on the hazard ratio scale at j						
				(corrected)		
Futility	1.5398	1.0000	0.8660	0.8339		
Efficacy	0.4218	0.6494	0.7499	0.8339		
Proportion of maximal events D_j/D_J at j with $D_J^* = 417$						
Map from AR-CRE Π_j	0.307	0.590	0.803	1.000		
	Stopping Probability at j				**Power**	**Events [Max F-U]**
AR-CRE (correct Π_j + corrected final bound)						
HR = 1.00 (Strong null)						
Futility	0.024	0.472	0.404	0.070	0.971	296 (335)
Efficacy	0.000	0.002	0.005	0.022	0.029	[3.4 (3.8)]
HR = 0.85 (Intermediate PH alternative)						
Futility	0.002	0.154	0.342	0.157	0.656	347 (417)
Efficacy	0.001	0.030	0.122	0.192	0.345	[4.3 (5.3)]
HR = 0.70 (Design PH alternative)						
Futility	0.000	0.014	0.059	0.028	0.100	325 (335)
Efficacy	0.009	0.280	0.417	0.194	0.900	[4.3 (4.6)]

Events and maximum follow-up (F-U, in years) summarized with mean (75th percentile);
AR-CRE = adaptive randomization censoring-robust estimator

5.5 Discussion

In this chapter, we extended our work on censoring-robust estimation and adaptive randomization in the fixed sample time-to-event trial setting in Chapter 4 to one with a group sequential design (GSD). GSDs allow for early stopping of a trial for efficacy, futility, or harm. Decisions of whether to stop for efficacy or futility are guided by stopping rules (or, monitoring guidelines) based upon group sequential boundaries. Group sequential boundaries are a function of the proportion of maximal information at the time of an interim analysis. When information does not grow proportionally to the number of events (e.g., non-linear information growth), the timing of interim analyses will not correspond to the chosen group sequential boundaries for which the design is based too maintain operating characteristics. Thus, group sequential boundaries require computing the sequential sampling density and the information growth of the statistic.

To adequately evaluate group sequential designs for our adaptive randomization censoring-robust estimator, we addressed two issues in this chapter: (1) characterizing the information growth of our estimator; and (2) account for non-independent increments to maintain design operating characteristics.

We demonstrated the non-linearity of our adaptive randomization censoring-robust estimator's information growth in a number of practical settings and how such non-linearity impacts the timing of interim analyses. We illustrated a procedure to calculate the information growth for a specified design, and how it can be done during the monitoring of a trial. We also showed that simply mapping the correct information growth to timing of events does not guarantee maintaining group sequential operating characteristics for a maximal information design. Then, we discussed how censoring-robust estimators, including our AR-CRE, violate the independent increments structure in a sequential testing setting. This implies directly applying Armitage et al. (1969) to obtain the sequential sampling density is not appropriate

without accounting for the non-independent increments. We outlined the approach by Murray and Tsiatis (1999) in the context of an AR-CRE to maintain overall type I error rate. Furthermore, for maximal information trials (i.e., maintaining power), we illustrated finding the maximal events D_J^* that attains the power to detect the design alternative.

Our investigation restricted to planning at the design stage where we assumed knowledge of the true information growth assuming no dropout. In practice, however, the assumed information growth — requiring assumptions of the accrual distribution, event times distribution, dropout times distribution, and administrative censoring time — will likely differ from the observed information growth, the degree to which they differ will be scenario dependent. Invoking a constrained boundaries approach (Burington and Emerson, 2003) is one way to flexibly account for revised estimates of information growth and timing of analyses during the monitoring of a GSD.

A strength of our adaptive randomization censoring-robust estimator in the fixed sample setting is that it consistently estimates a marginal hazard ratio randomized clinical trial estimand (RCT-E), standardized to a common censoring distribution, and does so more efficiently than existing censoring-robust estimators. A limitation under sequential sampling, however, is that the duration of follow-up grows with each subsequent interim analysis. Under non-proportional hazards, censoring-robust estimation at each interim analysis targets, at best, a surrogate of the RCT-E based on support at the planned final analysis. Future work to address this limitation includes reweighting the statistic based on the amount of follow-up at an interim analysis relative to the duration desired for the target RCT-E, à la Gillen (2003) for estimation at interim analyses under non-proportional hazards. Additionally, if shorter duration of follow-up from an efficacy standpoint is not adequate from a scientific perspective, the number of interim analyses for efficacy, if any, should be deliberated in accordance to a range of hypothesized treatment effects that may be observed once the trial starts.

As efforts to accelerate drug development continue, the need for careful evaluation of complex innovative designs is imperative. Our research from this chapter equips investigators and trialists with necessary tools to make informed decisions about the design and monitoring of time-to-event group sequential designs for which censoring-robust estimation is warranted.

Chapter 6

Conclusion

6.1 Summary

Randomized clinical trials (RCT) represent the gold standard in determining a causal relationship between an intervention and an outcome; however, they are not infallible. As more layers of complexity get added to trial designs, challenges arise in trying to obtain meaningful interpretations of scientific relevance robust to a violation of certain statistical assumptions. This dissertation aimed to elucidate the impact of these features in settings to see their utility when certain statistical assumptions may be violated, and provide potential remedies to obtain valid inference for the RCT estimand (RCT-E) or a real-world estimand (RW-E), as a step towards supporting efforts for evaluating benefit-risk of candidate interventions. To this end, we explored two types of clinical trial design features that have potential utility to support efforts of expediting drug development and lead to the adoption of approved indications to improve public health: enrichment and adaptive randomization.

First, we examined estimating a RW-E from fixed sample pre-post RCTs with a continuous primary outcome and a fixed pre-randomization enrichment strategy. We quantified the bias

induced by enrichment, by way of regression to the mean, and developed a bias-adjusted estimator assuming normality. An alternative remedy we found, if feasible, would be to have an enrichment period consisting of more than one pre-randomization assessment at a single time point to remove the chances of random high or low bias from regression to the mean. From our empirical studies, enrichment based on an individuals 'true' mean pre-randomization assessment score would allow for estimation of a RW-E provided there is a homogeneous pre-post intervention effect in the broader population. If there is belief of effect modification between those who would meet the enrichment criterion and those who would not, careful consideration is needed to decide whether only randomizing those meeting the enrichment criterion should be randomized. One approach used in practice has been to define the full analysis set population for the primary analysis on the enriched sample, but have some proportion of randomized participants not meet the enrichment criterion for secondary or exploratory objectives. This would allow estimation of both the RCT-E and RW-E.

We additionally examined estimating the RCT-E in fixed sample time-to-event RCTs with adaptive randomization, with emphasis on settings in the presence of time-varying treatment effects. We showed that adaptive randomization alters treatment arm specific censoring patterns differentially, even when censoring is only administrative. We extended the censoring-robust framework by accounting for a known adaptive randomization scheme in the inverse probability of censoring weighting for the reweighted partial likelihood estimating equation. This in turn allows us to target a standardized marginal hazard ratio (e.g., in the absence of intermittent censoring). Furthermore, by incorporating the known randomization scheme in our proposed adaptive randomization censoring-robust estimator, for time-to-event trials with adaptive randomization we have a a more efficient estimator compared to that by Boyd and colleagues (2012), and equivalent under fixed randomization. If proportional hazards truly holds, then unsurprisingly censoring-robust estimators are less efficient compared to the Cox proportional hazards estimator. Because data-driven modeling choices can invalidate trial results, waiting until the time of analysis to to assess whether proportional hazards

holds to decide which analysis method to use is not appropriate. Pre-specifying the analysis plan is therefore critical to ensure reliability of trial results. As such, we recommend pre-specifying a censoring-robust estimator when designing a time-to-event clinical trial to guard against the potential for time-varying treatment effects while maintaining an interpretable and replicable estimate of the RCT-E.

Lastly, we examined planning group sequential time-to-event RCTs with adaptive randomization when targeting the RCT-E via censoring-robust estimation. Because group sequential boundaries are based on the information fraction (proportion of maximal information), the timing of interim analyses (and the final analysis) should be based on information time. Under non-proportional hazards, however, it has previously been shown that weighted statistics exhibit non-linear information growth, suggesting that the proportion of maximal events is not equal to the information fraction. We demonstrated this phenomenon by characterizing the non-linearity of our adaptive randomization censoring-robust estimator's information growth. We proposed a three-pronged approach to maintain group sequential design operating characteristics: (i) mapping the information fraction of our adaptive randomization censoring-robust estimator to the proportion of maximal events; (ii) accounting for non-independent increments of censoring-robust estimators by outlining a procedure to find the final boundary (Murray and Tsiatis, 1999) that maintains the overall type I error rate; and (iii) modifying the original (naively assumed proportional hazards) design maximal events D to D^* to maintain power to detect the design alternative of scientific interest. While our empirical evaluation assumed proportional hazards, in practice, our proposed corrections can be applied time-to-event group sequential designs under non-proportional hazards (including for a weak null with a marginal hazard ratio of one and non-proportional hazards alternatives). This allows for greater flexibility in the planning stage when targeting a RCT-E.

Overall, valid statistical inference is a key component in making a reliable benefit-risk assessment for a candidate intervention in a RCT. This is essential to improve precision medicine

for the effective treatment and prevention of diseases. It is furthermore imperative that we obtain reliable answers to scientific questions in a timely manner while minimizing the number of patients randomly assigned to an intervention that may not be beneficial. To address these critical aspects of the drug discovery and regulatory process, there are different designs a sponsor study team can consider at the planning stage of a RCT. Understanding the impact of design choices and the statistical operating characteristics before selecting a design is essential to maintaining high standards for individual- and group-level ethics in clinical research. Because enrichment strategies and adaptive randomization are increasingly used, the potential impact of this dissertation research cannot be underestimated. Our research provides a framework and tools for trialists to be well-informed when designing enriched and adaptively randomized clinical trials for any disease.

6.2 Future Research Directions

6.2.1 Estimating the RCT-E and RW-E in time-to-event GSDs when response-adaptive randomization breaks independent increments

Response-adaptive randomization changes the randomization ratios for enrolling patients to treatment or control in a time-to-event setting where the randomization ratio allocations are modified according to interim estimates of the hazard ratio. We demonstrated how a known adaptive randomization scheme changes the treatment-specific censoring distributions by altering the number of subjects at risk within each treatment arm during the course of the trial, impacting estimation of the underlying RCT-E marginal hazard ratio.

Estimating a RW-E marginal hazard ratio in the fixed sample setting requires mapping

the RCT-E estimate back after accounting for the probability of sampling on the risk sets that are more representative of a broader population of interest. Obtaining the asymptotic sampling distribution of the RW-E estimator can be obtained appealing to the Rebolledo martingale central limit theorem and an empirical sandwich variance estimator via a Taylor series expansion around the weights.

In the sequential sampling setting, deviations from the independent increments structure can arise because the sampling weights may depend upon survival. Additionally, independent increments breaks down with response-adaptive randomization since those who are randomized in a particular allocation ratio next are determined by comparative outcome data from the current trial. In this dissertation, we assumed a known randomization scheme that did not depend upon survival. Next steps include considering independent increments structures where the weights depend upon: (i) only the history (current analysis time and everything before); and (ii) future treatment effects. Simulation-based techniques may be used to estimate sequential stopping boundaries. As was done in Chapter 3, an approach to maintain the overall family-wise type I error rate is to assume independent increments holds until the penultimate analysis time, at which time we adjust for non-independent increments via bootstrapping off the data with the observed correlation structure (Murray and Tsiatis, 1999).

6.2.2 Estimating the RCT-E and RW-E in time-to-event clinical trials with adaptive enrichment

In time-to-event clinical trials with adaptive randomization, we have shown that adaptive randomization can induce covariate-dependent censoring across treatment arms. We extended the censoring-robust estimation framework of Boyd et al. (2012) to fixed sample and group sequential time-to-event trials for a known adaptive randomization scheme. An exten-

sion of this work is to also consider the role of adaptive enrichment on estimation for both fixed sample and group sequential designs.

Adaptive enrichment creates sub-cohorts of subjects who are randomized and followed in a RCT. These cohorts, coupled with the randomized treatment arm to which each are assigned, can formulate mutually exclusive groups and then appeal to our and existing censoring-robust estimation approaches. As an example, consider an Alzheimer's disease RCT with the primary outcome being time to dementia, an adaptive enrichment strategy may be employed with apolipoprotein E (APOE) $\varepsilon 4$ carriers and non-carriers. Suppose there are two treatment arms. Analogous to our adaptive randomization censoring-robust estimator, here four censoring distributions require estimation, for: (i) APOE $\varepsilon 4$ carriers assigned to treatment; (ii) APOE $\varepsilon 4$ carriers assigned to control; (iii) APOE $\varepsilon 4$ non-carriers assigned to treatment; and (iv) APOE $\varepsilon 4$ non-carriers assigned to control. An estimate of the RCT-E $\exp(\beta_{RCT})$ is then obtained by solving a corresponding reweighted estimating equation for β and exponentiating the result, $\exp(\hat{\beta}_{RCT})$. Furthermore, accounting for the sampling bias induced by the adaptive enrichment with the estimate of the RCT-E can map back the RW-E, $\exp(\beta_{RW})$.

6.2.3 Estimating the RCT-E and RW-E in enriched clinical trials with a longitudinal continuous primary endpoint

Overall, in this dissertation research, we have demonstrated analytically and via simulation studies that for a fixed enrichment pre-post RCT: (i) treatment effect estimates can be biased despite using existing robust statistical methods (e.g., White, 1980); (ii) our novel method can reduce the bias; and (iii) response-adaptive randomization induces bias of the marginal hazard ratio even in the setting of no enrichment, which we conjecture will be exacerbated when combined with enrichment. We thus conjecture that the bias induced in a pre-post

RCT will also arise in RCTs with repeated measures of the outcome over time (longitudinal, e.g., changes in activities of daily living scores).

Adaptive enrichment in RCTs with a longitudinal primary outcome complicates the estimation of a standardized contrast of interest in the presence of effect modification. In such a setting, two estimands may be of interest, the RCT-E marginal change in response trajectories, θ_{RCT}, and the RW-E marginal change in response trajectories, θ_{RW}, in adaptive RCTs with a longitudinal continuous primary outcome and response-adaptive enrichment. Investigations can consider both the semiparametric (generalized estimating equation, GEE: Liang and Zeger, 1986) and parametric (linear mixed-effects model, LME: Laird and Ware, 1982) frameworks. Our next step is to extend the misspecification in the pre-post design setting to that of a slope model over time where the length of follow-up times vary.

Reweighting by the probability of sampling subpopulations based on the enrichment strategy can guard against effect modification. Two potential approaches include: (i) reweighting at the time of forming contrasts at each time point; and (ii) obtaining the estimate of θ (averaged across all times) and then reweight. In such a setting, a target contrast θ_{RCT} that we consistently test for at each interim analysis may serve as an appropriate RCT-E. In the presence of model misspecification (e.g., effect modification or time-varying treatment effects), one can appeal to Kittelson et al. (2005) by standardizing to a common θ to obtain an estimate of the marginal change in response trajectories conditional upon the sequential sampling scheme. This approach protects against non-linearities by only accounting for observed follow-up times (i.e., the support is up to the maximum observed time) at the time of the interim analysis. That is, each $\hat{\theta}_j$ (where j indexes an interim analysis time) has to be consistent for (the standardized) θ. Now with enrichment, θ can also be different in the presence of effect modification. Similar considerations for departures from independent increments under sequential sampling should be examined in this longitudinal setting as described in Section 6.2.1. Furthermore, using sampling weights is one approach to then

map the RCT-E estimate to an estimate of the RW-E of interest.

Bibliography

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., et al. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, 330(10):657–662.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

AlzForum (n.d.). Lecanemab. `https://www.alzforum.org/therapeutics/lecanemab`. Accessed on Aug 09, 2021.

Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, 132(2):235–244.

Boyd, A. P. (2009). *Censoring-robust treatment effect estimation in clinical trials with time-to-event outcomes*. University of Colorado Health Sciences Center.

Boyd, A. P., Kittelson, J. M., and Gillen, D. L. (2012). Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Statistics in medicine*, 31(28):3504–3515.

Brummel, S. S. and Gillen, D. L. (2014). Flexibly monitoring group sequential survival trials when testing is based upon a weighted log-rank statistic. *Sequential analysis*, 33(1):39–59.

Burington, B. E. and Emerson, S. S. (2003). Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics*, 59(4):770–777.

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics*, pages 247–254.

Chang, M. N., Hwang, I. K., and Shin, W. J. (1998). Group sequential designs using both type i and type ii error probability spending functions. *Communications in Statistics-Theory and Methods*, 27(6):1323–1339.

Council, N. R., of Behavioral, D., Sciences, S., on National Statistics, C., and on Handling Missing Data in Clinical Trials, P. (2010). *The prevention and treatment of missing data in clinical trials*. National Academies Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dabrowska, A. and Thaul, S. (2018). *Prescription Drug User Fee Act (PDUFA): 2017 Reauthorization as PDUFA VI*. Congressional Research Service.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):875–892.

Emerson, S. S., Kittelson, J. M., and Gillen, D. L. (2007). Frequentist evaluation of group sequential clinical trial designs. *Statistics in medicine*, 26(28):5047–5080.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 23(4):335–353.

Fleming, T. R. (2011). Addressing missing data in clinical trials. *Annals of internal medicine*, 154(2):113–117.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*, volume 256. John Wiley & Sons.

Gillen, D. L. (2003). *The use of weighted logrank statistics in group sequential testing and nonproportional hazards*. PhD thesis, University of Washington, Seattle, WA.

Gillen, D. L. and Emerson, S. S. (2005). Information growth in a family of weighted logrank statistics under repeated analyses. *Sequential Analysis*, 24(1):1–22.

Grill, J. D. (2021). Nicotinamide as an early alzheimer's disease treatment (neat). `https://clinicaltrials.gov/study/NCT03061474`. ClinicalTrials.gov identifier NCT03061474.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

Hu, F. and Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons.

Inc., E. (2024). A study to evaluate safety, tolerability, and efficacy of lecanemab in subjects with early alzheimer's disease. `https://clinicaltrials.gov/study/NCT01767311`. ClinicalTrials.gov identifier NCT01767311.

Jennison, C. and Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kittelson, J. M. and Emerson, S. S. (1999). A unifying family of group sequential test designs. *Biometrics*, 55(3):874–882.

Kittelson, J. M., Sharples, K., and Emerson, S. S. (2005). Group sequential clinical trials for longitudinal data with analyses using summary statistics. *Statistics in Medicine*, 24(16):2457–2475.

Korn, E. L. and Freidlin, B. (2011). Outcome-adaptive randomization: is it useful? *Journal of Clinical Oncology*, 29(6):771–776.

Korn, E. L. and Freidlin, B. (2017). Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. *JNCI: Journal of the National Cancer Institute*, 109(6):djx013.

Korn, E. L. and Freidlin, B. (2022). Time trends with response-adaptive randomization: the inevitability of inefficiency. *Clinical Trials*, 19(2):158–161.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

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

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liang, K.-Y. and Zeger, S. L. (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 134–148.

Little, R. J., D'agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360.

Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., Sano, M., Bieliauskas, L., Geldmacher, D., Clark, C., et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the alzheimer's disease assessment scale that broaden its scope. *Alzheimer Disease & Associated Disorders*, 11:13–21.

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

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

O'Connell, N. S., Dai, L., Jiang, Y., Speiser, J. L., Ward, R., Wei, W., Carroll, R., and Gebregziabher, M. (2017). Methods for analysis of pre-post data in clinical research: a comparison of five common methods. *Journal of biometrics & biostatistics*, 8(1):1.

of Health, U. D. and Services, H. (2020). National plan to address alzheimer's disease: 2020 update. https://aspe.hhs.gov/sites/default/files/migrated_legacy_files//197726/NatlPlan2020.pdf.

Pampallona, S., Tsiatis, A., and Kim, K. (1995). Spending functions for the type i and type ii error probabilities of group sequential tests. *J. Stat. Plan. Inference*, 42:1994–35.

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

Proschan, M. and Evans, S. (2020). Resist the temptation of response-adaptive randomization. *Clinical Infectious Diseases*, 71(11):3002–3004.

Proschan, M. and Evans, S. (2021). Reply to villar et al. *Clinical infectious diseases*, 73(3):e842–e843.

Rosen, W. G., Mohs, R. C., and Davis, K. L. (1984). A new rating scale for alzheimer's disease. *The American journal of psychiatry*, 141(11):1356–1364.

Rosenberger, W. F., Sverdlov, O., and Hu, F. (2012). Adaptive randomization for clinical trials. *Journal of biopharmaceutical statistics*, 22(4):719–736.

Schervish, M. J. (1984). Algorithm as 195: Multivariate normal probabilities with error bound. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(1):81–94.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in medicine*, 25(24):4334–4344.

Simon, N. and Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625.

Slud, E. and Wei, L. (1982). Two-sample repeated significance tests based on the modified wilcoxon statistic. *Journal of the American Statistical Association*, 77(380):862–868.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2):363–369.

Thall, P. F. (2021). Adaptive enrichment designs in clinical trials. *Annual review of statistics and its application*, 8(1):393–411.

Tsiatis, A. A. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, 9(1):93–108.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*, 77(380):855–861.

U.S. Food and Drug Administration (2016). Pdufa reauthorization performance goals and procedures fiscal years 2018 through 2022. https://www.fda.gov/media/99140/download.

U.S. Food and Drug Administration (2019a). Adaptive designs for clinical trials of drugs and biologics - guidance for industry. https://www.fda.gov/media/78495/download.

U.S. Food and Drug Administration (2019b). Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products - guidance for industry. https://www.fda.gov/media/121320/download.

U.S. Food and Drug Administration (2021). E9(r1) statistical principles for clinical trials: Addendum: Estimands and sensitivity analysis in clinical trials - guidance for industry. https://www.fda.gov/media/148473/download.

U.S. Food and Drug Administration (2023). Benefit-risk assessment for new drug and biological products - guidance for industry. https://www.fda.gov/media/152544/download.

van Houwelingen, H. C., van de Velde, C. J., and Stijnen, T. (2005). Interim analysis on survival data: its potential bias and how to repair it. *Statistics in medicine*, 24(18):2823–2835.

Wan, F. (2021). Statistical analysis of two arm randomized pre-post designs with one post-treatment measurement. *BMC medical research methodology*, 21:1–16.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, pages 193–199.

Wei, L. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, pages 227–236.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439.

Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321.

Ye, T. and Shao, J. (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5):1301–1323.

Zhang, L. and Rosenberger, W. F. (2007). Response-adaptive randomization for survival trials: the parametric approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 56(2):153–165.

Appendix A

For Chapter 3

A.1 Derivation of the analytic form of the bias of the RCT-E estimator with respect to the RW-E in enriched pre-post RCTs

Here, we derive the analytic form of the bias induced from enrichment when estimating the RW-E β_1 with the RCT-E estimator $\hat{\theta}$. First, we want to find the distribution function of a post measurement Y_2 given the pre measurement Y_1 meeting some enrichment criterion/threshold c. Without loss of generality, assume the enrichment criterion is $Y_1 > c$. Further, assume that the joint distribution function of (Y_1, Y_2) is denoted by $F_{\vec{Y}} \equiv F_{(Y_1,Y_2)}(y_1, y_2)$.

Then, the distribution function is

$$\begin{aligned}
F_{Y_2|Y_1>c}(y_2) &= \Pr[Y_2 \leq y_2 | Y_1 > c] \\
&= \frac{\Pr[Y_2 \leq y_2, Y_1 > c]}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)\times(-\infty,y_2]} dF_{\vec{Y}}(u_1, u_2)}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} \int_{(-\infty,y_2]} dF_{Y_2|Y_1=u_1}(u_2) dF_{Y_1}(u_1)}{\Pr[Y_1 > c]}
\end{aligned}$$

where the last equality follows after appealing to Fubini-Tonelli since $F_{\vec{Y}}$ is integrable on $\mathbf{R} \times \mathbf{R}$. Then, for an absolutely continuous random vector $\vec{Y}$, the density function is

$$\begin{aligned}
f_{Y_2|Y_1>c}(y_2) &= \frac{d}{dy_2} F_{Y_2|Y_1>c}(y_2) \\
&= \frac{\frac{d}{dy_2} \int_{(c,\infty)} \int_{(-\infty,y_2]} dF_{Y_2|Y_1=u_1}(u_2) dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} \frac{d}{dy_2} \int_{(-\infty,y_2]} dF_{Y_2|Y_1=u_1}(u_2) dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} f_{Y_2|Y_1=u_1}(y_2) dF_{Y_1}(u_1)}{\Pr[Y_1 > c]}
\end{aligned}$$

where the second to last equality follows after appealing to the dominated convergence theorem with bounding function of 1. Then, using $E[Y_2|Y_1 = u_1] = \mu_2 + \rho\frac{\sqrt{v_2}}{\sqrt{v_1}}(u_1 - \mu_1)$, the

post-randomization mean given the enrichment criterion is

$$
\begin{aligned}
E[Y_2|Y_1 > c] &= \int_{(-\infty,\infty)} y_2 dF_{Y_2|Y_1>c}(y_2) \\
&= \int_{(-\infty,\infty)} y_2 f_{Y_2|Y_1>c}(y_2) dy_2 \\
&= \frac{\int_{(-\infty,\infty)} y_2 \int_{(c,\infty)} f_{Y_2|Y_1=u_1}(y_2) dF_{Y_1}(u_1) dy_2}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(-\infty,\infty)} \int_{(c,\infty)} y_2 f_{Y_2|Y_1=u_1}(y_2) dF_{Y_1}(u_1) dy_2}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} \int_{(-\infty,\infty)} y_2 f_{Y_2|Y_1=u_1}(y_2) dy_2 dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} \int_{(-\infty,\infty)} y_2 dF_{Y_2|Y_1=u_1}(y_2) dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} E[Y_2|Y_1 = u_1] dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \frac{\int_{(c,\infty)} \mu_2 + \rho\frac{\sqrt{v_2}}{\sqrt{v_1}}(u_1 - \mu_1) dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \mu_2 + \rho\sqrt{v_2}\frac{\int_{(c,\infty)} \frac{(u_1-\mu_1)}{\sqrt{v_1}} dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \mu_2 - \rho\sqrt{v_2}\frac{\int_{(-\infty,c)} \frac{(u_1-\mu_1)}{\sqrt{v_1}} dF_{Y_1}(u_1)}{\Pr[Y_1 > c]} \\
&= \mu_2 - \rho\sqrt{v_2}\frac{\int_{(-\infty,c^*)} z_1 \cdot d\Phi(z_1)}{1 - \Phi(c^*)}
\end{aligned}
$$

where $z_1 \equiv \frac{(u_1-\mu_1)}{\sqrt{v_1}}$.

Appendix B

For Chapter 4

B.1 Deriving a decomposition of a treatment-specific censoring distribution

For an adaptive randomization rule denoted by $\mathcal{AR} = \{\pi_{Z|A=k}(z) \text{ for all } z \text{ and } k = 1, ..., K\}$ the treatment-specific censoring distribution

$$
\begin{aligned}
S_C(t|Z=z) &\equiv \Pr(C > t|Z=z) \\
&= \frac{\Pr(C > t, Z=z)}{\Pr(Z=z)} \\
&= \frac{\sum_{k=1}^{K} \Pr(C > t, Z=z|A=k) \Pr(A=k)}{\Pr(Z=z)} \\
&= \frac{\sum_{k=1}^{K} \Pr(C > t|Z=z, A=k) \Pr(Z=z|A=k) \Pr(A=k)}{\Pr(Z=z)} \\
&= \frac{\sum_{k=1}^{K} \Pr(C > t|Z=z, A=k) \Pr(Z=z|A=k) \Pr(A=k)}{\sum_{k=1}^{K} \Pr(Z=z|A=k) \Pr(A=k)} \\
&\equiv \frac{\sum_{k=1}^{K} S_C(t|Z=z, A=k) \pi_{Z|A=k}(z) \pi_A(k)}{\sum_{l=1}^{K} \pi_{Z|A=l}(z) \pi_A(l)}
\end{aligned}
$$

where the treatment-arm specific adaptive randomization accrual subperiod weight is

$$w_k^{\mathcal{AR}}(z) \equiv \frac{\pi_{Z|A=k}(z)\pi_A(k)}{\sum_{l=1}^{K}\pi_{Z|A=l}(z)\pi_A(l)}.$$

B.2 Proof of Proposition 4.1

For trial participant i, let T_i denote the event time, C_i the censoring time, $X_i = \min\{T_i, C_i\}$ the observed time, $\Delta_i = I(T_i \leq C_i)$ the event indicator, $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ the indicator of an event at time t, $Y_i(t) = I(X_i \geq t)$ the indicator of being at risk for an event at time t, and $\boldsymbol{Z}_i$ the covariate vector. Define two filtrations (history processes)

$$\mathcal{F}_{it}^{-} = \sigma\{\, N_i(s), W_i^{\mathcal{AR}}(s^+), Y_i(s^+), \boldsymbol{Z}_i\,;\ s \in [0,t)\,\}$$
$$\mathcal{F}_{t}^{-} = \sigma\{\, N_i(s), W_i^{\mathcal{AR}}(s^+), Y_i(s^+), \boldsymbol{Z}_i\,;\ s \in [0,t)\,;\ i = 1,...,n\,\}$$

where $W_i^{\mathcal{AR}}(X_i) = 1/S_C(X_i|Z_i)$ for $S_C(X_i|Z_i)$ as specified in (4.5). Earlier in Appendix B.1, we showed that $S_C(t|Z=z) = \sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) S_C(t|Z=z, A=k)$. Since the right hand side contains known weights $w_k^{\mathcal{AR}}(z)$ and we estimate $S_C(t|Z=z, A=k)$ with the Kaplan-Meier estimator (1958), our proposed $W_i^{\mathcal{AR}}(t)$ is a left-continuous and adapted to $\mathcal{F}_t$; hence, it is an $\mathcal{F}_t$-predictable process, as are the weights for the Boyd-Kittelson-Gillen censoring-robust estimator (2012). Furthermore, define

$$\boldsymbol{S}_{\mathcal{AR}}^{(r)}(t;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} W_i^{\mathcal{AR}}(t) Y_i(t) \boldsymbol{Z}_i^r \exp(\boldsymbol{\beta}\boldsymbol{Z}_i)$$
$$\boldsymbol{s}_{\mathcal{AR}}^{(r)}(t;\boldsymbol{\beta}) = E\left[\boldsymbol{S}_{\mathcal{AR}}^{(r)}(t;\boldsymbol{\beta})\right]$$

for $r = 0, 1, 2$ that correspond to a scalar, vector, and matrix, respectively,

$$\boldsymbol{V}_W(t;\boldsymbol{\beta}) = \frac{\boldsymbol{S}^{(2)}_{\mathcal{AR}}(t;\beta)}{S^{(0)}_{\mathcal{AR}}(t;\beta)} - \overline{\boldsymbol{Z}}_W(t;\boldsymbol{\beta})\overline{\boldsymbol{Z}}'_W(t;\boldsymbol{\beta})$$
$$\boldsymbol{v}_W(t;\boldsymbol{\beta}) = \frac{\boldsymbol{s}^{(2)}_{\mathcal{AR}}(t;\beta)}{s^{(0)}_{\mathcal{AR}}(t;\beta)} - \overline{\boldsymbol{z}}_W(t;\boldsymbol{\beta})\overline{\boldsymbol{z}}'_W(t;\boldsymbol{\beta}),$$

and intensity process

$$A^W_i(t;\boldsymbol{\beta}) = \int_0^t W_i^{\mathcal{AR}}(s)Y_i(s)d\Lambda_i(s;\boldsymbol{\beta})$$
$$dA^W_i(t;\boldsymbol{\beta}) = W_i^{\mathcal{AR}}(s)Y_i(s)\exp(\beta Z_i)d\Lambda_0(s)$$

where $\Lambda_0(s)$ is the baseline cumulative hazard.

The observed data consists n independent and identically distributed $(X_i, \Delta_i, \boldsymbol{Z}_i)$ for $i = 1, ..., n$. Here, we denote the treatment arm variable by Z_i (while still allowing $\boldsymbol{Z}_i$ to include other covariates in addition to treatment arm). Assume conditionally independent censoring, $T_i \perp\!\!\!\perp C_i | Z_i$. Consider a reweighted partial likelihood of the form

$$PL_W(\boldsymbol{\beta}) = \prod_{i=1}^n \left\{ \frac{\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)}{\sum_{j=1}^n W_j^{\mathcal{AR}}(X_i)Y_j(X_i)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_j)} \right\}^{W_i^{\mathcal{AR}}(X_i)\Delta_i}.$$

Then, the log reweighted partial likelihood is

$$\begin{aligned}\mathcal{L}_W(\boldsymbol{\beta}) &\equiv \log PL_W(\boldsymbol{\beta}) \\ &= \sum_{i=1}^n W_i^{\mathcal{AR}}(X_i)\Delta_i \left[\boldsymbol{\beta}'\boldsymbol{Z}_i - \log\left\{ \sum_{j=1}^n W_j^{\mathcal{AR}}(X_i)Y_j(X_i)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_j) \right\} \right]\end{aligned}$$

and the reweighted partial likelihood score, in counting process notation, is

$$\mathcal{U}_W(\boldsymbol{\beta}, t) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log PL_W(\boldsymbol{\beta}) \tag{B.1}$$

$$= \sum_{i=1}^{n} \int_0^t W_i^{\mathcal{AR}}(s) \left\{ \boldsymbol{Z}_i - \frac{\sum_{j=1}^n \boldsymbol{Z}_j W_j^{\mathcal{AR}}(s) Y_j(s) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)}{\sum_{j=1}^n W_j^{\mathcal{AR}}(s) Y_j(s) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)} \right\} dN_i(s) \tag{B.2}$$

where

$$\overline{\boldsymbol{Z}}_W(t; \boldsymbol{\beta}) = \frac{\sum_{j=1}^n \boldsymbol{Z}_j W_j^{\mathcal{AR}}(t) Y_j(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)}{\sum_{j=1}^n W_j^{\mathcal{AR}}(t) Y_j(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)} = \frac{\boldsymbol{S}_{\mathcal{AR}}^{(1)}(t; \boldsymbol{\beta})}{S_{\mathcal{AR}}^{(0)}(t; \boldsymbol{\beta})}$$

$$\overline{\boldsymbol{z}}_W(t; \boldsymbol{\beta}) = \frac{\boldsymbol{s}_{\mathcal{AR}}^{(1)}(t; \boldsymbol{\beta})}{s_{\mathcal{AR}}^{(0)}(t; \boldsymbol{\beta})}.$$

We assume the following regularity conditions (Fleming and Harrington, 1991, p.289-290):

(RC.1) The time τ is such that $\int_0^\tau \lambda_0(x)dx < \infty$

(RC.2) For $\boldsymbol{S}_{\mathcal{AR}}^{(j)}$, $j = 0, 1$, and 2, there exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ and, respectively, scalar, vector, and matrix functions $s_{\mathcal{AR}}^{(0)}$, $\boldsymbol{s}_{\mathcal{AR}}^{(1)}$, and $\boldsymbol{s}_{\mathcal{AR}}^{(2)}$ defined on $\mathcal{B} \times [0, \tau]$ such that $j = 0, 1, 2$,

$$\sup_{x \in [0,\tau], \boldsymbol{\beta} \in \mathcal{B}} ||\boldsymbol{S}_{\boldsymbol{\mathcal{AR}}}^{(j)}(\boldsymbol{\beta}, x) - \boldsymbol{s}_{\mathcal{AR}}^{(j)}(\boldsymbol{\beta}, x)|| \to_p 0 \quad \text{as } n \to \infty.$$

(RC.3) (modified since we only consider fixed covariates) There exists a $\delta > 0$ such that

$$n^{-1/2} \sup_{1 \le i \le n,\, 0 \le x \le r} |\boldsymbol{Z}_i| Y_i(x) I\{\boldsymbol{\beta}^{*\prime} \boldsymbol{Z}_i > \delta |\boldsymbol{Z}_i|\} \to_p 0 \quad \text{as } n \to \infty.$$

(RC.4) Let $\mathcal{B}$ and $\boldsymbol{s}_{\mathcal{AR}}^{(j)}$, $j = 0, 1, 2$, be defined as in the previous condition, and let

$$\boldsymbol{v}_W(x; \boldsymbol{\beta}) = \frac{\boldsymbol{s}_{\mathcal{AR}}^{(2)}(x; \boldsymbol{\beta})}{s_{\mathcal{AR}}^{(0)}(x; \boldsymbol{\beta})} - \overline{\boldsymbol{z}}_W(x; \boldsymbol{\beta}) \overline{\boldsymbol{z}}_W'(x; \boldsymbol{\beta}).$$

Then, for all $\boldsymbol{\beta} \in \mathcal{B}$ and $0 \leq x \leq \tau$, $\frac{\partial}{\partial \boldsymbol{\beta}} s^{(0)}_{\mathcal{AR}}(x;\beta) = \boldsymbol{s}^{(1)}_{\mathcal{AR}}(x;\boldsymbol{\beta})$ and $\frac{\partial^2}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'} s^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}) = \boldsymbol{s}^{(2)}_{\mathcal{AR}}(x;\boldsymbol{\beta})$.

(RC.5) The functions $\boldsymbol{s}^{(j)}_{\mathcal{AR}}$ are bounded and $s^{(0)}_{\mathcal{AR}}$ is bounded away from 0 on $\mathcal{B} \times [0,\tau]$; for $j = 0,1,2$, the family of functions $\boldsymbol{s}^{(j)}_{\mathcal{AR}}(x;\cdot)$, $0 \leq x \leq \tau$, is an equicontinuous family at $\boldsymbol{\beta}^*$.

(RC.6) The matrix $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*,\tau) = \int_0^\tau \boldsymbol{v}_W(x;\boldsymbol{\beta}^*) s^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)\lambda_0(x)dx$ is positive definite.

Consistency of of the proposed adaptive randomization censoring-robust estimator $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE}$

To show that $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE} \to_p \boldsymbol{\beta}^*$, under regularity conditions, we first show that the log reweighted partial likelihood converges to a function maximized by $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. Then to get the desired result, we appeal to a result of concave functions (Lemma 8.3.1 in Fleming and Harrington, 1991, p.297).

We start with the difference in log reweighted partial likelihoods over $[0,t]$ for an arbitrary $\boldsymbol{\beta}$ and true value $\boldsymbol{\beta}^*$. Scaling this process by n^{-1} corresponds to

$$
\begin{aligned}
&X_n(t;\boldsymbol{\beta}) = n^{-1}\left\{\mathcal{L}_W(t;\boldsymbol{\beta}) - \mathcal{L}_W(t;\boldsymbol{\beta}^*)\right\} \\
&= n^{-1}\left[\sum_{i=1}^n \int_0^t (\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{Z}_i dN_i(x) \right. \\
&\qquad\qquad \left. - \int_0^t \log\left\{\frac{\sum_{i=1}^n W_i^{\mathcal{AR}}(x) Y_i(x)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)}{\sum_{i=1}^n W_i^{\mathcal{AR}}(x) Y_i(x)\exp(\boldsymbol{\beta}^{*\prime}\boldsymbol{Z}_i)}\right\} d\overline{N}(x)\right].
\end{aligned}
$$

If the compensator of $X_n(t;\boldsymbol{\beta})$ is

$$A_n(t;\boldsymbol{\beta}) = n^{-1}\left[\sum_{i=1}^{n}\int_0^t (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{Z}_i W_i^{\mathcal{AR}}(x)Y_i(x)\exp\{\boldsymbol{\beta}^{*\prime}\boldsymbol{Z}_i\}\lambda_0(x)dx \right.$$
$$\left. - \int_0^t \log\left\{\frac{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})}{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)}\right\} W_i^{\mathcal{AR}}(x)Y_i(x)\exp\{\boldsymbol{\beta}^{*\prime}\boldsymbol{Z}_i\}\lambda_0(x)dx\right],$$

then

$$X_n(t;\boldsymbol{\beta}) - A_n(t;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^t \left\{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{Z}_i - \log\frac{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})}{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)}\right\} dM_i(x).$$

is a martingale. In particular, since the integrand in the above equation is locally bounded (because $\boldsymbol{Z}_i$ are bounded with probability 1) and predictable, the process $X_n(\cdot;\boldsymbol{\beta}) - A_n(\cdot;\boldsymbol{\beta})$ is a locally square integrable martingale. Furthermore, $X_n(\cdot;\boldsymbol{\beta}) - A_n(\cdot;\boldsymbol{\beta})$ has a predictable variation process at t

$$\begin{aligned}
&\langle X_n(\cdot;\boldsymbol{\beta}) - A_n(\cdot;\boldsymbol{\beta}), X_n(\cdot;\boldsymbol{\beta}) - A_n(\cdot;\boldsymbol{\beta})\rangle(t)\\
&= n^{-2}\sum_{i=1}^{n}\int_0^t \left\{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{Z}_i - \log\frac{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})}{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)}\right\}^2 Y_i(x)\exp(\boldsymbol{\beta}^{*\prime}\boldsymbol{Z}_i)\lambda_0(x)dx\\
&= n^{-1}\int_0^t \left[(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{S}^{(2)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - 2(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{S}^{(1)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)\log\left\{\frac{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})}{S^{(0)}_{\mathcal{AR}}(t;\boldsymbol{\beta}^*)}\right\}\right.\\
&\quad \left. + 2\log\left\{\frac{S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})}{S^{(0)}_{\mathcal{AR}}(t;\boldsymbol{\beta}^*)}\right\} S^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)\right]\lambda_0(x)dx.
\end{aligned}$$

Regularity conditions (RC.1), (RC.2), and (RC.5) imply that the predictable variation process at τ converges to a finite limit. Then, appealing to Lenglart's inequality yields $X_n(\tau;\beta) - A_n(\tau;\beta) \to_p 0$. It can be shown that the compensator $A_n(\tau;\boldsymbol{\beta})$ converges in probability to

$$A_{\mathcal{AR}}(\tau;\beta) = \int_0^\tau \left[(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\mathbf{s}^{(1)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*) - \log\left\{\frac{s^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})}{s^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta}^*)}\right\} s^{(0)}_{\mathcal{AR}}(x;\boldsymbol{\beta})\right]\lambda_0(x)dx,$$

which is concave with a unique maximum at $\boldsymbol{\beta}^*$ and then $X_n(\tau;\boldsymbol{\beta}) \to_p A_{\mathcal{AR}}(\tau;\beta)$ when $\boldsymbol{\beta} \in \mathcal{B}$. It can also be shown that, with regularity conditions (RC.4) and (RC.5), $X_n(\tau;\boldsymbol{\beta})$ is concave with a unique maximum at $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE}$. Then, applying Lemma 8.3.1 (3) (Fleming and Harrington, 1991, p.297) implies $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE} \to_p \boldsymbol{\beta}^*$.

Asymptotic normality of the reweighted partial likelihood score $\mathcal{U}_W(\boldsymbol{\beta})$

Since $\boldsymbol{Z}_i$ is bounded with probability 1, $\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}\}$ is $\mathcal{F}_t$ predictable. Then, the reweighted partial likelihood score (B.1) has counting process $W_i^{\mathcal{AR}}(t)N_i(t)$ with compensator $A_i^W(t) = \int_0^\infty W_i^{\mathcal{AR}}(x)Y_i(x)\lambda_0(x)dx$. Then, $M_i^W(t;\boldsymbol{\beta}) = W_i^{\mathcal{AR}}(t)N_i(t) - A_i^W(t)$ is an $\mathcal{F}_t$-martingale with mean zero. Because the sum of stochastic intergrals with respect to the compensator is zero, we can now express the reweighted partial likelihood score process as a martingale transform

$$\mathcal{U}_W(\boldsymbol{\beta},t) = \sum_{i=1}^n \int_0^t \left\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_W(s;\boldsymbol{\beta})\right\} dM_i^W(s;\boldsymbol{\beta})$$

and $n^{-1/2}\mathcal{U}_W(\boldsymbol{\beta}^*,t)$ is a martingale process with variation process

$$\begin{aligned}
&\langle n^{-1/2}\mathcal{U}_W(\boldsymbol{\beta}^*), n^{-1/2}\mathcal{U}_W(\boldsymbol{\beta}^*)\rangle(t) \\
&\qquad = n^{-1}\sum_{i=1}^n \int_0^t \left\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_W(s;\boldsymbol{\beta})\right\}^{\otimes 2} W_i^{\mathcal{AR}}(s)Y_i(s)\exp(\boldsymbol{\beta}^{*\prime})\lambda_0(s)ds \\
&\qquad = \int_0^t \boldsymbol{V}_W(s;\boldsymbol{\beta}^*)\boldsymbol{S}^{(0)}_{\mathcal{AR}}(s;\boldsymbol{\beta}^*)\lambda_0(s)ds \\
&\qquad \to_p \int_0^t \boldsymbol{v}_W(s;\boldsymbol{\beta}^*)\boldsymbol{s}^{(0)}_{\mathcal{AR}}(s;\boldsymbol{\beta}^*)\lambda_0(s)ds \equiv \boldsymbol{\Sigma}_W(t).
\end{aligned}$$

appealing to regularity conditions (RC.1), (RC.2), and (RC.5). It can then be shown that the Lindeberg condition is met by first defining, for $\ell = 1, ..., p$,

$$n^{-1/2}\mathcal{U}_{W,\ell\epsilon}(\boldsymbol{\beta}^*, t) = n^{-1/2}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_{i\ell} - \overline{\boldsymbol{Z}}_{W,\ell}(s;\boldsymbol{\beta}^*)\right\} \cdot I\left\{n^{-1/2}|\boldsymbol{Z}_{i\ell} - \overline{\boldsymbol{Z}}_{W,\ell}(s;\boldsymbol{\beta}^*)| > \epsilon\right\} dM_i^W(s;\boldsymbol{\beta}^*)$$

and then showing that $\langle n^{-1/2}\mathcal{U}_{W,\ell\epsilon}(\boldsymbol{\beta}^*, t), n^{-1/2}\mathcal{U}_{W,\ell\epsilon}(\boldsymbol{\beta}^*, t)\rangle(t)$

$$\begin{aligned}
&= n^{-1}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_{i\ell} - \overline{\boldsymbol{Z}}_{W,\ell}(s;\boldsymbol{\beta}^*)\right\}^2 \cdot I\left\{n^{-1/2}|\boldsymbol{Z}_{i\ell} - \overline{\boldsymbol{Z}}_{W,\ell}(s;\boldsymbol{\beta}^*)| > \epsilon\right\} W_i^{\mathcal{AR}}(s)Y_i(s)\exp(\boldsymbol{\beta}^{*\prime})\lambda_0(s)ds \\
&\leq \int_0^t n^{-1}\sum_{i=1}^{n} W_i^{\mathcal{AR}}(s)Y_i(s)\exp(\boldsymbol{\beta}^{*\prime})K_W^2 \cdot I\left\{n^{-1/2}|\boldsymbol{Z}_{i\ell} - \overline{\boldsymbol{Z}}_{W,\ell}(s;\boldsymbol{\beta}^*)| > \epsilon\right\}\lambda_0(s)ds \\
&= \int_0^t O_p(1)\cdot o_p(1)\lambda_0(s)ds \\
&\to_p 0
\end{aligned}$$

where $K_W = \sup_{i,s}|\boldsymbol{Z}_{i\ell} - \overline{\boldsymbol{Z}}_{W,\ell}(s;\boldsymbol{\beta}^*)| < \infty$ and using regularity conditions (RC.2), (RC.3), and (RC.5). Then, by the Rebolledo central limit theorem, $n^{-1/2}\mathcal{U}_W(\boldsymbol{\beta}^*, t) \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_W(t))$ (where $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\boldsymbol{\tau})$).

Asymptotic normality of the proposed adaptive randomization censoring-robust estimator $\hat{\boldsymbol{\beta}}_{CRE}^{\mathcal{AR}}$

From the previous part, the reweighted score equation can be expressed as

$$\mathcal{U}_W(\hat{\boldsymbol{\beta}}_{CRE}^{\mathcal{AR}}) = \sum_{i=1}^{n}\int_0^\infty \left\{\boldsymbol{Z}_i - \frac{\boldsymbol{S}_{\mathcal{AR}}^{(1)}(s;\boldsymbol{\beta})}{S_{\mathcal{AR}}^{(0)}(s;\boldsymbol{\beta}^*)}\right\} dM_i^W(s;\boldsymbol{\beta}) = 0.$$

Earlier in Appendix B.1, we showed that $S_C(t|Z=z) = \sum_{k=1}^{K} w_k^{\mathcal{AR}}(z) S_C(t|Z=z, A=k)$ and also stated that our proposed $W_i^{\mathcal{AR}}(t)$ is an $\mathcal{F}_t$-predictable process, as are the weights for Boyd et al. (2012). Because of this, asymptotic normality of the reweighted partial likelihood score then follows by arguments analogous to those by Boyd et al. (2012) where, here, our weights $W_i^{\mathcal{AR}}(t) = 1/S_C^{\mathcal{AR}}(t|Z_i)$ incorporate the adaptive randomization rule $\mathcal{AR}$. We also extended the Boyd et al. (2012) robust variance estimator by incorporating our weights $W_i^{\mathcal{AR}}(t)$. This requires, as they did, using an asymptotically equivalent form of the reweighted partial likelihood score process that has independent and identically distributed mean zero terms. In our notation, let $\overline{N}_W(t) = \sum_{i=1}^{n} W_i(t) N_i(t)$. Then, the asymptotically equivalent form of (4.6) is

$$\begin{aligned}
\mathcal{U}_W^*(\boldsymbol{\beta}, t) &= \sum_{i=1}^{n} \mathcal{U}_{W,i}^*(\boldsymbol{\beta}, t) \\
&= \sum_{i=1}^{n} W_i^{\mathcal{AR}}(t) \Delta_i \boldsymbol{Z}_i - \int_0^{\tau} \frac{\boldsymbol{s}_{\mathcal{AR}}^{(1)}(x;\boldsymbol{\beta})}{\boldsymbol{s}_{\mathcal{AR}}^{(0)}(x;\boldsymbol{\beta})} d\overline{N}_W(t) \\
&\quad - \int_0^{\tau} \frac{\boldsymbol{S}_{\mathcal{AR}}^{(1)}(x;\boldsymbol{\beta})}{\boldsymbol{s}_{\mathcal{AR}}^{(0)}(x;\boldsymbol{\beta})} d\tilde{N}_W(t) + \int_0^{\tau} \frac{\boldsymbol{S}_{\mathcal{AR}}^{(0)}(x;\boldsymbol{\beta}) \boldsymbol{s}_{\mathcal{AR}}^{(1)}(x;\boldsymbol{\beta})}{\{\boldsymbol{s}_{\mathcal{AR}}^{(0)}(x;\boldsymbol{\beta})\}^2} d\tilde{N}_W(t)
\end{aligned}$$

Following the arguments of Boyd et al. (2012) that show the robust variance estimator for the censoring-robust estimator, which we denoted by $\hat{\boldsymbol{V}}_W(\hat{\boldsymbol{\beta}}_{CRE}^{\mathcal{AR}})$, is consistent for

$$\boldsymbol{V}(\boldsymbol{\beta}^*) = \lim_{n\to\infty} n^{-1} \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1}$$

where

$$\begin{aligned}
\boldsymbol{A} &= -\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \partial \mathcal{U}_W(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} \,|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\
\boldsymbol{\Sigma} \equiv \boldsymbol{B} &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathcal{U}_W^*(\boldsymbol{\beta}^*) \mathcal{U}_W^*(\boldsymbol{\beta}^*)'.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE}$ solves the reweighted estimating equation and taking a first-order Taylor expansion around $\boldsymbol{\beta}^*$

$$\begin{aligned} 0 &= n^{-1/2}\mathcal{U}_W(\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE}) \\ &= n^{-1/2}\mathcal{U}_W(\boldsymbol{\beta}^*) - \{\hat{\boldsymbol{V}}_W(\tilde{\boldsymbol{\beta}})/n\}\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE} - \boldsymbol{\beta}^*) \end{aligned}$$

for $\tilde{\boldsymbol{\beta}}$ between $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE}$ and $\boldsymbol{\beta}^*$. Using earlier results of $\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE} \to_p \boldsymbol{\beta}^*$ (consistency of proposed estimator), $n^{-1/2}\mathcal{U}_W(\boldsymbol{\beta}^*) \to_d \mathcal{N}(0, \boldsymbol{B}(\boldsymbol{\beta}^*))$ (asymptotic normality of score process), and provided $\hat{\boldsymbol{V}}_W(\tilde{\boldsymbol{\beta}} \to_p \boldsymbol{V}(\boldsymbol{\beta}^*)$ (consistency of robust variance estimator), it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathcal{AR}}_{CRE} - \boldsymbol{\beta}^*) \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}^*))$ as $n \to \infty$.

Appendix C

For Chapter 5

C.1 Deriving the treatment-subperiod specific censoring distribution assuming no dropout

Let E denote a random variable for the entry time of a participant during an accrual period from trial start to τ_A (with respect to calendar time since trial start). Consider a randomization scheme in which the treatment arm allocation probability changes K during accrual. Then, define a K-partition of the accrual period $(0, \tau_A) = (e_0, e_1) \cup \{\cup_{k=2}^{K}[e_k, e_{k+1})\}$ where $e_1 = 0$ and $e_{K+1} = \tau_A$. Let τ_j represent the administrative censoring time (with respect to calendar time since the start of the trial) at analysis j. Assuming no dropout, censoring is captured fully by administrative censoring at τ_j such that $C = \tau_j - E$. Then, the treatment-subperiod-specific censoring distribution does not depend on the treatment arm and can be

expressed as

$$
\begin{aligned}
S_C(t;\tau_j|Z=z,A=k) &= S_C(t;\tau_j|A=k) \\
&= \Pr[C>t|A=k,E<\tau_j] \\
&= \Pr[C>t|E\in[e_k,e_{k+1}),E<\tau_j] \\
&= \Pr[\tau_j-E>t|E\in[e_k,e_{k+1}),E<\tau_j] \\
&= \Pr[E<\tau_j-t|E\in[e_k,e_{k+1}),E<\tau_j] \\
&= \frac{\Pr[E<\tau_j-t,\ E\in[e_k,e_{k+1}),\ E<\tau_j]}{\Pr[E\in[e_k,e_{k+1}),E<\tau_j]} \\
&= \frac{\Pr[e_k\le E<\{\tau_j-t\}\wedge e_{k+1}]}{\Pr[e_k\le E<e_{k+1}]} \\
&= \frac{F_E(\{\tau_j-t\}\wedge e_{k+1})-F_E(e_k)}{F_E(\tau_j\wedge e_{k+1})-F_E(e_k)}
\end{aligned}
$$

where accrual in subperiod k corresponds to entry times (with respect to calendar time) between e_k and e_{k+1}.

C.2 Proposed correction for timing of analyses and final bound to maintain overall type I error in a GSD with fixed randomization

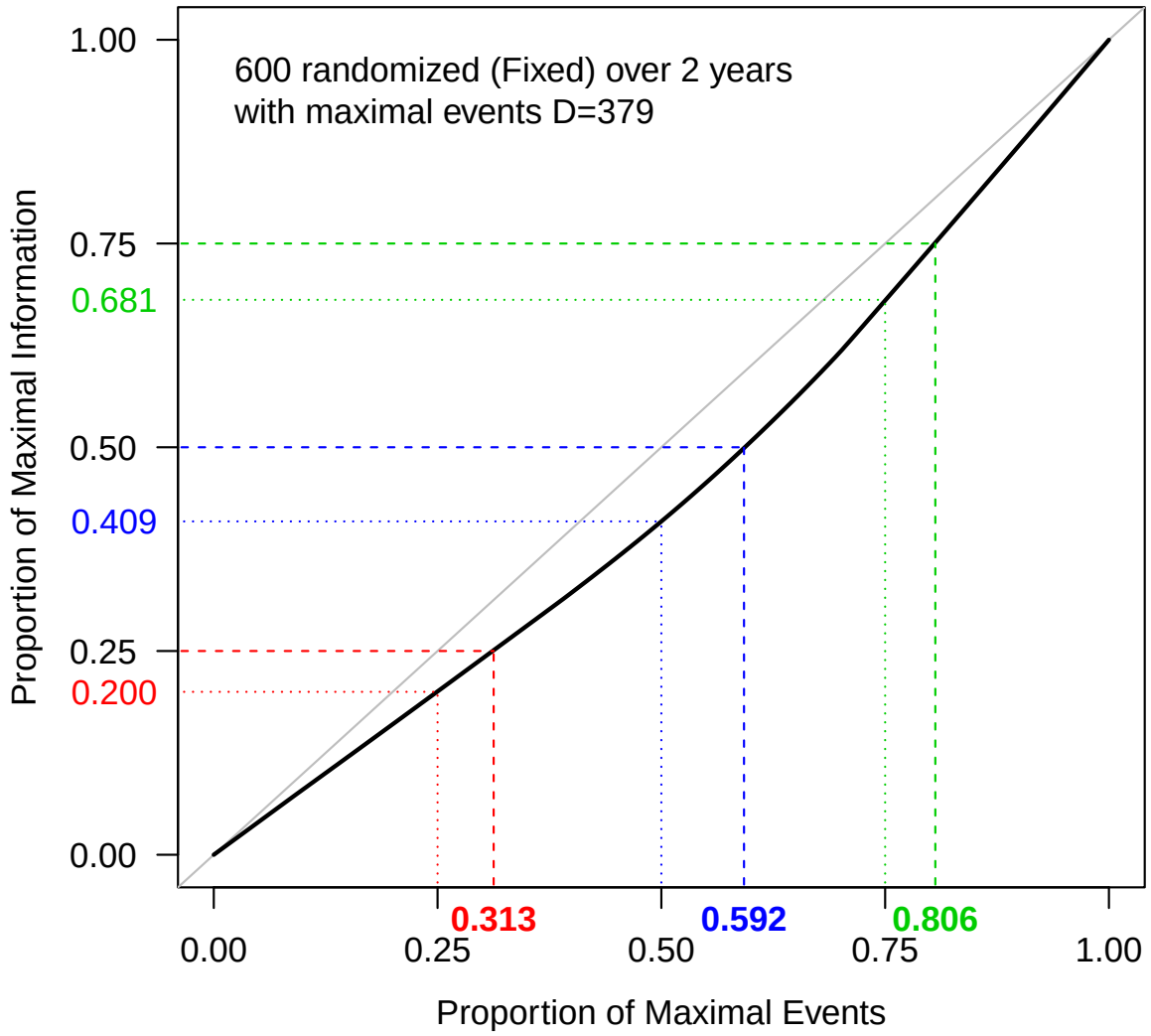


Figure C.1: Empirical information growth for our adaptive randomization censoring-robust estimator (black curve) under the strong null with Exponential(rate=0.30) true event times for uniform accrual of 600 participants over 2 years with a fixed (0.65:0.35) randomization scheme. Gray solid line indicates proportion of maximal information equals proportion of maximal events, D_j/D_J. Dotted lines map equally-spaced proportion of maximal events (0.25, 0.50, 0.75) to the respective proportion of maximal information, Π_j. Dashed lines map equally-spaced Π_j to the respective D_j/D_J.

Table C.1: Operating characteristics based on 2000 simulations for the group sequential design with symmetric O'Brien-Fleming boundaries with uniform accrual of 600 participants over 2 years and a fixed (0.65:0.35) randomization scheme. Boundaries are displayed on the hazard ratio and Z statistic scales, along with the corrected final boundary. Proportion of maximal events at each analysis is summarized according to naively assuming information growing proportional to events and the empirical AR-CRE information growth from Figure C.1. Operating characteristics include stopping probabilities at each analysis and power (cumulative probability of stopping) for futility and efficacy, along with the average and 75th percentile of events and maximum follow-up (years) at stopping. Power for efficacy in this setting is the overall type I error rate where nominal level is $\alpha = 0.025$, with Monte Carlo error (0.018, 0.032).

Analysis j	1	2	3	$4 = J$		
Boundaries at equally spaced Π_j on the hazard ratio scale at j						
					(J corrected)	
Futility	1.5398	1.0000	0.8660	0.8059	(0.8288)	
Efficacy	0.4218	0.6494	0.7499	0.8059	(0.8288)	
Boundaries at equally spaced Π_j on Z scale at j						
					(J corrected)	
Futility	-4.006	-2.833	-2.313	-2.003	(-1.744)	
Efficacy	2.003	0.000	-1.157	-2.003	(-1.744)	
Proportion of maximal events D_j/D_J at j with $D_J = 379$						
Naive: assume $\Pi_j \propto D_j$	0.250	0.500	0.750	1.000		
Map from AR-CRE Π_j	0.313	0.592	0.806	1.000		
	Stopping Probability at j				**Power**	**Events [Max F-U]**
Cox PH						
Futility	0.020	0.489	0.371	0.096	0.976	245 (284)
Efficacy	0.000	0.002	0.007	0.015	0.024	[]
AR-CRE (naive)						
Futility	0.034	0.477	0.395	0.080	0.986	241 (284)
Efficacy	0.000	0.002	0.008	0.004	0.014	[]
AR-CRE (only correct Π_j)						
Futility	0.032	0.490	0.384	0.079	0.985	266 (305)
Efficacy	0.000	0.002	0.006	0.007	0.015	[]
AR-CRE (correct Π_j + corrected final bound)						
Futility	0.032	0.490	0.384	0.074	0.980	266 (305)
Efficacy	0.000	0.002	0.006	0.013	0.020	[]

Events and maximum follow-up (F-U, in years) summarized with mean (75th percentile);
AR-CRE = adaptive randomization censoring-robust estimator