Adaptive Randomization Ratios in Multi-Arm Clinical Trials

Michael P Garcia

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2015

Committee:

Scott S. Emerson, Chair

Noah Simon

Program Authorized to Offer Degree:

School of Public Health, Department of Biostatistics

University of Washington

Abstract

Adaptive Randomization Ratios in Multi-Arm Clinical Trials

Michael P Garcia

Chair of the Supervisory Committee:

Professor Scott S. Emerson

Department of Biostatistics

Ethics and economics are two of many motivations for streamlining the process of discovering new therapies for the treatment of human disease. Population ethics and economics dictate that any true benefit be demonstrated as quickly as possible to allow effective therapies to be marketed and distributed to patients in need, while individual ethics favor the protection of the human subjects actually enrolled in a specific research study. Adaptive designs address both of these issues by using accumulating results to modify the trial design during the trial. One statistical aspect of multi-arm clinical trial design that can be adaptively modified is the randomization ratio, such as has been proposed in the I-SPY 2 [2] and BATTLE [16] trials. This thesis explores the statistical properties of adaptively modifying randomization ratios to favor accrual to experimental arms whose interim results appear most promising. Biases and inflation of type I error that can arise from modifying randomization ratios, in particular in the presence of secular time trends in

patient characteristics, are described and demonstrated with simulations. In addition, we show that these issues can largely be mitigated by properly accounting for randomization ratios through stratified analyses.

Acknowledgements

I would like to extend my thanks to my advisor, Scott Emerson, for his insight, guidance, and enthusiasm while conducting this research; to Noah Simon for his insight and interest in joining this research; and to the staff of the UW Biostatistics Department, who were a tremendous help in many ways during my Master's studies at UW. And, of course, to my family for their unending patience and support while completing this thesis.

To my parents, Phil and Carol.

Contents

List of Figures

List of Tables

Chapter 1

Background

1.1 Classical clinical trial design

Research directed toward the identification of safe and effective interventions to improve human health and well-being requires a balance of ethical and scientific concerns. Concern for the protection of human subjects participating in research and for human patients receiving treatments identified by research dictate that research should minimize risk to participants, while identifying effective therapies as quickly as possible and efficiently utilizing available resources (e.g. patients, money, time). Scientific concerns include ensuring that data obtained from such research is accurate, generalizable, and informative for guiding decisions in future research, clinical practice, and public health.

The gold standard to determine the effect of a treatment is the randomized clinical trial (RCT), in which an experimental therapy is compared to a standard of care or placebo by randomly assigning experimental units (individuals, clinics, healthcare organizations, etc.) to receive either an experimental therapy or control therapy. In the most basic form of an RCT, all available experimental units are randomized and studied, and data are analyzed after observations have been collected on all experimental units. The randomization process is crucial for obtaining estimates of treatment effects that are both unbiased and consistent. Randomizing subjects to study arms removes systematic bias in the assignment, and ensures that on average, measured and unmeasured prognostic factors will be equally distributed across groups. Although imbalances may occur due to chance alone, the probability distribution of such imbalances are well understood. Furthermore, the

magnitude of any differences in prognostic factors between treatment groups decreases with larger sample sizes.

Ethical issues in the conduct of RCTs relate both to subjects on the trial as well as future patients who may receive therapies that are demonstrated to be safe and effective. In particular, randomizing subjects to study arms introduces many *individual ethical* concerns concerning humans participating in research. When there is true clinical equipoise, i.e. true uncertainty in the scientific and medical community regarding the benefit of a treatment, the ethical concerns over randomizing are significantly diminished. However, if during research, accumulating preliminary evidence indicates that one intervention is superior to another, either overall or in some subpopulation, the ethics of randomizing subjects to a intervention which data suggests may be inferior becomes more controversial. Uncertainties inherent in the *a priori* assessments of adequate treatment effect sizes and variability may lead to RCT designs that over- or underestimate the number of subjects necessary to investigate any true clinical benefit of the treatment. This creates ethical issues especially when true effects could be detected with fewer resources and exposing a lower number of RCT subjects to an inferior treatment.

In addition to individual ethics, the design and conduct of RCTs also impact the *group ethics* that relate to the larger population of patients that an adopted effective treatment might benefit. From this viewpoint, ethics demand that safe and effective interventions be identified and made available as quickly as possible to those who will benefit, and that the study of ineffective therapies be quickly discarded in favor of studying those more promising. This often aligns with the economic interest of cost-efficient trials by efficient use of resources (subjects and time) to bring effective interventions to the market quickly, thereby minimizing research expenditure and maximizing profit.

1.2 Group sequential designs

Methods addressing the problem of making decisions regarding efficacy using interim data have been well-studied [1, 19]. In these methods, RCT designs pre-specify the rules for determining the maximal statistical information, as well as for the conduct of interim analyses performed on the accruing data. These *group sequential designs* protect the experimentwise false positive rate and, depending on the exact monitoring rules, statistical power to detect a hypothesized treatment effect. An RCT will terminate at an interim analysis if the observed treatment effects are so extreme that the benefit or harm of the treatment has been

established with scientific and statistical rigor. Alternatively, an RCT might also terminate at an interim analysis if the observed results establish with statistical rigor that the experimental treatment does not provide a benefit that is clinically important. The only adaptive change to these designs is to potentially alter the sample size observed at the end of the study. Modifications of the group sequential nature change the sampling density of the test statistic, which for a variety of stopping rules can be computed directly [1]. Importantly, however, a classical group sequential design modifies only the statistical operating characteristics related to the sample size distribution and (perhaps) the statistical power, without changing other statistical aspects (e.g., the randomization ratio) or the scientific goals of the study.

1.3 Adaptive designs

More recently described adaptive designs [3, 10, 20] allow for more varied use of interim trial results to modify an RCT design. These modifications may be scientific in nature (e.g. dose level, experimental therapy, endpoint, eligibility criteria), or may be statistical in nature (e.g. randomization procedure, adjustment of the maximal sample size) [23]. In draft guidance released in 2010, the United States Food and Drug Administration (FDA) defines an *adaptive design clinical study* as 'a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study' [23]. To date, the FDA has not offered guidance on non-prospectively modified designs.

Group sequential designs are well understood in part because the sampling distribution is defined *a priori* through a full description of stopping rules. Designs which make further adaptive modifications to study aspects, also must have *a priori* specification of modifications in order to be effectively studied. An *a priori* specification must be in full detail for an accurate assessment of operating characteristics.

Adaptive modifications of many aspects of clinical trials have been discussed and advocated for a variety of ethical, scientific, and economic reasons. Proponents argue that more flexible designs could allow for more efficient use of resources (addressing group ethical concerns), fewer research subjects assigned to inferior therapies (addressing individual ethical concerns), and efficiently answering treatment response for a larger number of indications (such as identification of subgroups having the greatest benefit). The group sequential methods described in the previous section fall within the FDA's broad description of adaptive designs, but do not encompass all aspects of RCTs that can be adaptively modified.

Discussion arising from the release of the FDA draft guidance has identified some concerns with adaptive designs and suggestions for improvement of adaptive designs that have been implemented [4, 6]. Our present concerns with adaptively modified trials are three-fold: First, control of type I error must be maintained to avoid the further study or approval of truly ineffective therapies. Second, appropriate measures must be taken to avoid bias in estimation of treatment effects. Third, the data resulting from adaptively modified trials must be interpretable and relevant to answering a scientific question of interest [23].

1.3.1 Outcome dependent randomization

As previously mentioned, there are many aspects of trial design that can be modified based on accruing data. The focus of this thesis is specifically on adaptive modifications to the procedure in which subjects are randomized to study arms. Randomization is the key to maintaining balance between treatment groups and reducing bias in assessment of treatment effect. Although customary, equal randomization to experimental and control arms (1:1 randomization ratio) is not necessary to guarantee unbiased estimates. However, as we will show, changing randomization ratios on the basis of accruing trial data can create bias by disrupting covariate balance and introducing confounding. Relationships between covariates and clinical outcome can result in imbalance between treatment groups when this ratio is modified on the basis of observed outcomes during the course of the study, because the treatment groups may not be balanced with respect to important prognostic variables having trends over the course of patient accrual [6, 23].

1.4 Examples: past, present, and future

Several trials incorporating adaptive design have been conducted to date, with more planned. The following examples demonstrate a sustained interest in using various adaptive approaches to improve trial design with regard to various ethical and economic concerns. The past, present, and future use of adaptive designs illustrate a need for more thorough investigation of less well-understood designs.

1.4.1 ECMO Trial

An early example of adaptive design is the ECMO trial [9]. This trial, conducted in 1985, arose from case-reports and phase I trials that provided early evidence that in neonates with persistent pulmonary hy-

pertension (PPH), extracorporeal membrane oxygenation (ECMO), a mechanical heart-lung bypass, resulted in a large benefit in survival rates over standard of care (SOC) [24]. These preliminary data showing as much as a 60% absolute improvement in survival over SOC were strong enough to convince clinicians of a strong benefit (as high as 80% vs. 20% mortality rate in SOC vs. ECMO [24]), but not conclusive enough for clinicians to adopt ECMO as a new SOC. A sufficiently powered RCT with fixed randomization ratios would provide an adequate scientific solution, but due to the strong belief of the efficacy of ECMO, randomizing patients to ECMO or SOC throughout an entire trial was deemed unethical. This lack of clinical equipoise led to a 'play the winner' trial design [25, 27], in which a total of 12 neonates with persistent pulmonary hypertension were randomized to either ECMO or SOC. The study design called for randomization of the first subject to either ECMO or SOC with equal probability and subsequent subjects were randomized on the basis of accruing results, with randomization favoring arms displaying more promising results. The trial ended with 1 subject accrued to the SOC arm (this subject was the first to be randomized to SOC, and died), and 11 accrued to ECMO (all of whom survived).

The ethical concerns that drove the ECMO trial design were protection of the individual subjects involved in research. Preliminary evidence of benefit of ECMO and the severity of the disease were sufficient to discourage investigators from using 1:1 randomization. The immediate nature of the outcome of each subject (which was generally known before the following subject was randomized), reduced logistical difficulties. However, the resulting data was not sufficient for the broader clinical research community to definitively determine whether ECMO was superior to SOC [24]. Ultimately a follow-up trial, which also employed an adaptive randomization design to address ethical issues, was conducted in 1989 to more adequately assess the putative benefit of ECMO vs SOC [18, 24]. This study was conducted in two stages, with 1:1 randomization during the first stage, which was planned to end after observing 4 deaths in a single arm. During the second stage, new subjects were only accrued to the arm with fewer than 4 deaths until either 4 deaths were also observed in that arm, or 'the number of survivors was significantly larger than the number of survivors in the arm that had been discontinued first'. [18]. The SOC arm was discontinued first, and a total of 28 individuals were assigned to ECMO with 1 death, and a total of 10 assigned to SOC, with 4 deaths. The lack of concurrent controls raises the possibility that the comparison groups were not entirely equivalent, potentially due to background trends in disease severity or other covariates related to outcome. Furthermore, this study has been criticized for inference resulting from this design being highly sensitive to small differences in observed interim results, with p-values ranging from < 0.001 to 0.62 being assigned to

the data [6].

Another RCT was conducted in Great Britain in 1996, randomizing 185 infants with severe respiratory failure to either ECMO at a specialized hospital or to conventional therapy at the hospital where the infant initially presented [12]. Randomization was equal across arms and non-adaptive, and the primary endpoint was a composite of death or severe disability at one year of age. This study reported a relative risk of 0.55 (95% CI: [0.39, 0.77], p=0.0005), in favor of ECMO. In contrast to the mortality rate of 80% reported in the 1980s, the observed rate of the combined endpoint of mortality and severe disability was 59% on the SOC of the 1996 trial. This could indicate that estimates of mortality rate in the 1980s were incorrect, or that a trend in mortality rate occurred over the course of a decade.

In retrospect, with knowledge of the effectiveness of ECMO, the original ECMO trial was successful in protecting subjects on the study, as only one subject was randomized was assigned to the inferior treatment (unknown at the time the trial was conducted). However, the inadequacy of the scientific results resulted in a larger follow-up trials being conducted anyway. It is worth noting that there was a delay in obtaining scientifically sound results that could have led to an earlier wide-spread adoption of a highly effective, life-saving therapy, and that research subjects were again randomized to SOC after the initial trial. While there has been much speculation on what would have occurred had the original trial used equal randomization, it is not possible to determine with certainty the effect such a trial would have had on clinical care involving ECMO.

1.4.2 BATTLE

The Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE) trial in non-small cell lung cancer [16, 22] is a completed phase II randomized trial that used relationships between interim outcomes and biomarker data to adaptively modify the randomization procedure. The study outcome was 8-week disease control rate (DC), defined as the rate of non-progression at 8 weeks [15]. After pre-specified baseline biomarker data (tumor gene expression) was collected, 97 participants were initially assigned to one of four study arms with equal probability. An analysis of these 97 subjects examined associations between biomarker subgroups and DC, and this information was used to update the probability of a new subject being randomized to each arm, conditional on observed biomarkers. The data were modeled with a two parameter hierarchical Bayesian probit model, with vague prior distributions for the parameters. The model was designed with the intention for the the two parameters to reflect the amount of information

sharing occurring across biomarker groups within the same treatment arm, and across all treatment arms. Zhou et al. provide a full description of the model [28]. Subsequently, the randomization ratios were updated continuously, based on the updated posterior probability of mean DC in each treatment-biomarker group. Every 4 weeks, subjects had the opportunity to be re-randomized according to updated randomization ratios. Combinations of treatments and biomarker groups with low posterior probability of efficacy were allowed to be dropped from the study. Note that this study design allowed non-promising studying arms to be stopped, but did not allow for early conclusions to be made about arms showing promise. These arms were accelerated, but no conclusion was made. A treatment was considered efficacious if the posterior probability of the DC in a biomarker defined group being greater than the DC of 0.30 observed in historical controls was greater than 80% [16]. Comparisons among treatment arms were not considered.

While the primary goal was to identify efficacious treatment-biomarker combinations, this adaptive design favors the ethical treatment of individual subjects enrolled in the BATTLE trial by selectively randomizing subjects to treatment arms in which their biomarker subtype had shown superior results. A key feature of BATTLE was the use of a rapidly observable endpoint so that randomization could be continuously updated. An endpoint such as overall survival would limit such an adaptive design as all subjects may enroll before any events are observed and adaptation can occur.

1.4.3 I-SPY 2

The Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And moLecular Analysis (I-SPY) study was a non-randomized clinical study in the setting of breast cancer. Following tumor genotyping and chemotherapy, subjects were evaluated for pathologic complete response (pCR), defined as disappearance of invasive cancer at time of surgery. The goal of the study was to evaluate the predictive ability of pCR on recurrence free survival (RFS) and overall survival (OS), both overall and within four biomarker defined subgroups, and a benefit in survival for those with pCR was reported [8].

I-SPY set the stage for I-SPY 2, a currently ongoing multi-arm, multi-site phase II randomized clinical trial in the setting of breast cancer. I-SPY 2 has a goal of assessing the efficacy of several chemotherapeutic agents on the primary outcome of pCR, defined as disappearance of invasive cancer at time of surgery, and secondary endpoints of disease-free and overall survival at 10 years [2]. I-SPY 2 also aims to validate biomarkers and identify effective treatments within subgroups defined by 14 biomarker profiles [2]. Similarly to BATTLE, the randomization ratio is adaptively modified with the goal of identifying the optimal arm for

each subject on the basis of biomarkers. If a particular treatment is performing well in a biomarker defined subgroup, then that subgroup will be enriched by accelerating accrual of patients with that biomarker profile to that particular treatment arm. If statistical significance is achieved, the treatment may graduate (along with the corresponding biomarker signatures) to phase III studies. Similarly, if a treatment is doing poorly for a certain biomarker signature, the accrual rate of patients with that signature will decrease. If it performs poorly for all signatures, the treatment may be dropped. If a treatment either graduates or is dropped, a new experimental agent will be selected to enter the trial. Accrual to the control arm is kept constant, however.

This last feature allows for a constant pipeline of investigational therapies to be studied, and aims to reduce the time between completion of investigation of one therapy and the start of investigation of the next experimental therapy. Phase II studies are designed with the goal of establishing safety and gathering preliminary efficacy data, before advancing to Phase III studies. With estimates of failure rates for demonstrating efficacy in Phase III studies as high as 60% [13], continuing Phase II studies to maintain a constant supply of potential therapeutic agents for Phase III trials is prudent. In contrast to the ECMO trial, which focused on ethical issues surrounding the treatment of individual subjects on the trial, the adaptive nature of the I-SPY 2 trial also attempts to address on group ethics by aiming to demonstrate efficacy of the most promising experimental therapies as quickly as possible and reduce the time taken to bring effective therapies to market.

A non-adaptive feature that I-SPY 2 also employs is the use of a single control arm as a comparator for multiple experimental arms. This aims to reduce the total number of research subjects, which would be significantly higher if each arm had its own control, and reduces the logistical burden of running multiple clinical trials. There are unique considerations that must be taken with multi-arm, single control trials. For example, the effects of any spuriously unrepresentative control arms will be propagated and affect the inference for all experimental arms. Furthermore, the advancement of specific biomarker/treatment defined groups to Phase III trials introduces a differential time trend in the accrual of patients to both the control and experimental treatment arms. To the extent that the biomarkers are prognostic of clinical outcome, this may introduce a confounding imbalance in the treatment arms.

1.5 Specific aims

This thesis investigates adaptively modifying randomization ratios based on accumulating analysis of treatment effect in the setting of multi-arm Phase II clinical trials. This type of adaptive modification is also referred to as *outcome dependent randomization* and bears some similarity to traditional *play the winner* approaches, although the focus of accelerating effective arms through the discovery process rather than optimizing research subject assignment could perhaps be more appropriately described as a *grouped play the winner design*. Following the general design of the I-SPY 2 trial, a multi-arm trial with 4-experimental arms and a single control arm will be considered. The only adaptive aspect of the design considered here is the outcome dependent adaptive randomization; subgroup enrichment and incorporation of biomarker data will not be considered, although these are aspects of I-SPY 2 which are intertwined with adaptive randomization. As discussed above, a full specification of the prospective modifications is necessary for an accurate assessment of the operating characteristics of an adaptive trial. A feature of the data necessary to adaptively modify randomization ratios is the availability of the primary outcome while subject accrual is ongoing. If all subjects are accrued to a trial before the first outcome is observed, the randomization ratio cannot be modified.

As discussed above, one or more experimental arms may be observed to be performing well while monitoring the conduct of a multi-arm clinical trial, but it may be that not enough evidence has accumulated in the interim data to establish an effect of interest. If the trial is still accruing new subjects, both ethics and economics suggest preferentially accruing subjects to the arms which are performing well, to identify effective treatments quickly. To reduce the number of research subjects at risk, it is natural to consider maintaining a constant accrual rate to the control arm, regardless of the number of ongoing arms. However, not maintaining a constant ratio of each experimental arm to control leads to further imbalance between groups by accruing to various arms preferentially at different times.

Two statistical models are considered. In the first, a continuous outcome is assumed to be immediately known. The summary statistic of interest is the difference in mean outcome between experimental and control. For computational simplicity, the data generated in simulations is normal, although the model does not require this. The second model considered is a survival model, in which the hazard ratio comparing experimental to control is used to summarize the treatment effect. As some of the properties of the adaptive randomization strategy derive solely from the adaptive change in randomization ratio, both models are

considered in a 2-arm model. The continuous model is also studied in the general 5-arm design described above. In the 5-arm design, after a pre-specified interim analysis, arms with an interim test statistic exceeding a certain threshold are deemed promising and will complete Phase II first, and then be advanced to Phase III. Any arms with an interim test statistic not meeting the threshold are postponed, with stage 2 of Phase II completed after any promising arms complete Phase II.

The focus of these discussions will be on the effects of adaptively modifying randomization ratios on various statistical operating characteristics, including type I error, power, and bias. Central to these issues is the imbalance of subject assignment to trial arms with respect to time. When there are associations between time and outcome, there is confounding in the estimation of treatment effect. This will be shown directly to be an issue when there is a secular trend in the outcome measure. However, the Phase III graduation scheme can induce a trend in estimation of treatment effect in conjunction with biomarker-treatment specific analysis by altering the control group biomarker composition.

After demonstrating the existence of these issues, we will show that conducting analyses stratified on periods of constant accrual rates can ameliorate any confounding that might occur. However, there are some additional issues related to 'random high bias' that will have to be considered in these stratified analyses.

Chapter 2

Two-arm study

2.1 Design and notation

We begin by considering a two-arm clinical trial with one control arm, one experimental arm, and one planned interim analysis. Because response adaptive randomization in a multi-arm trial with several experimental arms and a single control arm will be discussed in Chapter 3, a general notation to accommodate both designs will be developed here. Suppose that J analyses of the data are planned ($J-1$ interim analyses, and 1 final analysis), and that K experimental arms are each tested against a single control arm. Let Y_{ijk} (independent) be the i^{th} observation during the j^{th} stage of arm k, and assume that $Y_{ijk} \sim \left(\mu_{ijk}, \sigma^2_{ijk}\right)$ are the first two moments of its distribution. Let $(\widehat{\mu}_{jk}, \widehat{\sigma}^2_{jk})$ and $(\widehat{\mu}_{j0}, \widehat{\sigma}^2_{j0})$ be the first two sample moments of the experimental and control arm, respectively, computed from the data accumulated by the j^{th} interim analysis, and let $\delta_j = \widehat{\mu}_{jk} - \widehat{\mu}_{j0}$, be the interim estimate of the treatment effect, where the control arm is denoted by $k = 0$. Under the assumption of equal variances across arms, a t-statistic T_{jk} can also be computed using a pooled variance estimate

$$T_{jk} = \frac{\widehat{\mu}_{jk} - \widehat{\mu}_{j0}}{\sqrt{\left(\frac{1}{n_{jk}} + \frac{1}{n_{j0}}\right) \cdot \frac{(n_{jk}-1)\cdot\widehat{\sigma}^2_{jk} + (n_{j0}-1)\widehat{\sigma}^2_{j0}}{n_{jk}+n_{j0}-2}}} \tag{2.1}$$

In a classical fixed-sample design, T_{jk} could be used to test the null hypothesis $H_0 : \mu_j = \mu_0$. In a group sequential design, this interim statistic could be used to modify the trial by potentially concluding early

efficacy, futility, or harm depending on the strength and direction of the interim evidence. In the present design, this interim statistic will be used solely in a decision rule to decide whether or not to accelerate accrual to the experimental arm(s) in the subsequent stage, and will not be used for any efficacy analysis.

For the two-arm design discussed in this chapter, assume, without loss of generality, a planned total sample size of 200 to the experimental arm and that a single interim analysis is conducted after 100 have completed the study on the experimental arm. Let the initial randomization ratio of experimental to control be 1:1, with blocked randomization guaranteeing 100 subjects are accrued to experimental and control during stage 1. If the statistic T_{1k} comparing experimental arm k to control is larger than some pre-specified threshold a, the experimental arm will be considered promising, and randomization in the second stage will be modified to accrue to experimental and control in a $4:1$ ratio. In the multi-arm trial with $K = 4$ experimental arms discussed in Chapter 3, this will correspond to a single arm being accelerated while maintaining accrual of 20% of the overall stage 2 sample to the control arm. In the present 2-arm design, this adaptation can be viewed as protecting individual ethics by reducing the number of research subjects who receive a therapy which is possibly inferior, as indicated by the interim analysis. Group ethics are also addressed since the total number of research subjects is reduced, resulting in studies of truly effective therapies completing with fewer subjects, while still having adequate safety data on the experimental arm. This is intended to allow effective treatments to be brought to market more quickly. If $T_{jk} < a$, then no adaptation will occur. After the planned 200 have completed the experimental arm, the difference of means $\delta = \mu_1 - \mu_0$ will be estimated from the cumulative sample means in each arm and the null hypothesis $H_0 : \delta = 0$ will be tested against the one sided alternative $H_A : \delta > 0$ at a significance level of 0.025, using a two-sample t-test assuming equal variances (substituting $j = 2$ in Equation 2.1). Later, in §2.8, we will also consider a t-test allowing for unequal variances. Data collected over all stages will be pooled for this final analysis.

While we choose to summarize the data on the normalized Z-statistic scale in making decisions, a variety of equivalent scales may be chosen, including maximum likelihood estimate, fixed-sample p-value, partial sum, conditional power, and Bayesian posterior probability [5].

2.2 Simple example

Consider a simple example in which the true expected outcome is zero in the control arm, and that the experimental arm has an average treatment effect of 2: $\mu_1 = 2$, $\mu_0 = 0$. With fixed total accrual of 200 to

the experimental arm and a planned maximum sample size of 200 to the control arm, an interim analysis is conducted after 100 subjects are accrued to each arm. For the first stage, $1:1$ randomization is used, but during the second stage, $4:1$ randomization of experimental to control is used. This simple example always forces the experimental arm to accelerate, regardless of the interim results, unlike the general design considered in §2.1 which only accelerates arms deemed promising. Table 2.2.1 displays the sample size and expected response in each arm at each stage. In this simple example with no trend, there is no bias in estimating the treatment effect, though the imbalance in accrual over time can be clearly seen in the stage-wise and overall ratios of experimental to control.

Table 2.2.1
Expected stage-wise sample sizes and means
Large effect in experimental and acceleration, but no trend

Stage	N_{con}	μ_0	σ_0^2	N_1	μ_1	σ_1^2	δ	Bias	N_1:N_{ctl}	N_j
1	100	0	1	100	2	1	2	0	1:1	200
2	25	0	1	100	2	1	2	0	4:1	125
Total	125	0	1	200	2	1	2	**0**	8:5	325

2.3 Secular trend in outcome

Now consider the same example, but assume a linear time trend in the mean, where the average response increases by 1 every 125 subjects. Note this increase occurs regardless of treatment group assignment. The mock data in Figure 2.3.1 illustrates this sampling scheme and extreme time trend. The individual outcomes are color coded by treatment stage. Stage 1 data is composed equally of experimental arm and control arm data, while 80% of Stage 2 data is accrued to the experimental arm, and only 20% to control arm. In other words, controls are more likely to be accrued during the first stage, while experimental arm subjects are balanced between stage 1 and stage 2. In the presence of a positive trend, this biases the average Arm 1 outcome $\overline{Y}_{\cdot 1}$ upwards relative to the control arm outcome $\overline{Y}_{\cdot 0}$.

Table 2.3.2 shows the expected stage-wise and cumulative means in the same scenario shown in Table 2.2.1, but with a linear trend in the mean response with slope of $\beta = \mu_{(i+1)jk} - \mu_{ijk} = \frac{1}{125}$. This increase in mean of 1 per 125 subjects accrued to the trial is assumed to begin immediately upon the start of accrual.

These results demonstrate that bias in estimation of the treatment effect can arise in the presence of a secular time trend in the outcome. In this situation, accelerating accrual to the experimental arm results in bias through an imbalance in accrual rates over time. Although the stage-wise means and variation

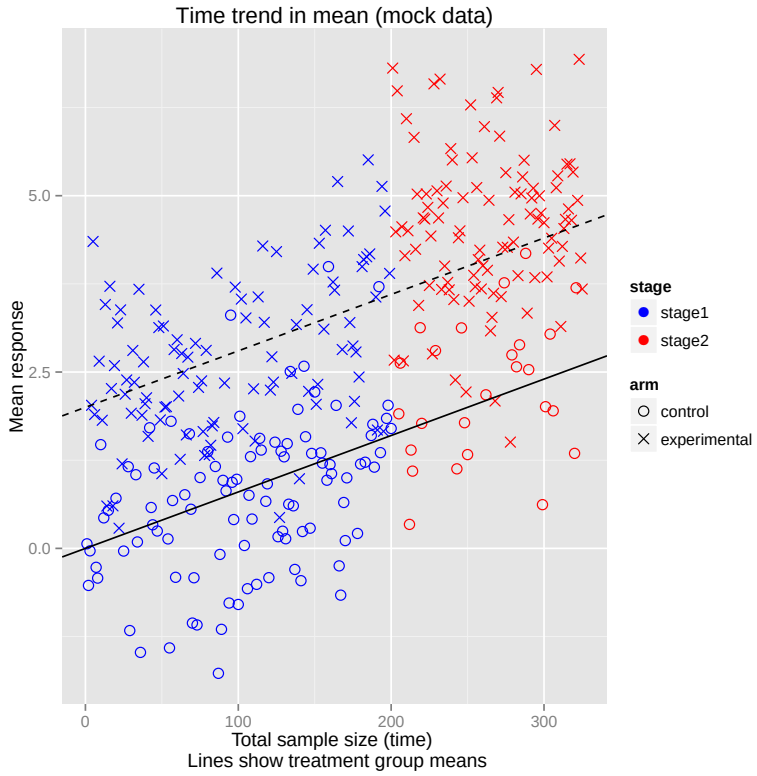

Figure 2.3.1
Response vs. sample size, colored by stage
Mock data

are balanced ($\mu_{11} = 2.8$, $\mu_{10} = 0.8$, $\mu_{21} = 2.3$, $\mu_{20} = 1.3$, and $\sigma^2_{11} = \sigma^2_{10} = 1.213$, $\sigma^2_{21} = \sigma^2_{20} = 1.083$), the corresponding pooled mean and variance are not balanced due to the imbalance in sample sizes ($\mu_1 = 3.05, \mu_0 = 0.9$). Note that the time trend considered refers to a background trend over time in the outcome measure Y in the population, unrelated to treatment effect. This is distinct from a time-by-treatment interaction, which will not be considered.

2.4 Random high bias

In the previous section, we discussed differences in the true mean when sampling is performed with a changing randomization ratio. However, when making decisions to accelerate accrual to one arm, those decisions will be based on estimates that might be spuriously higher or lower than the true value. Conducting analyses of accruing data and modifying the randomization ratio on these interim results bears some similarity to

Table 2.3.2
Expected stage-wise sample sizes, means, and bias
Large true effect in experimental and both trend and acceleration

Stage	N_{con}	μ_0	σ_0^2	N_1	μ_1	σ_1^2	δ	Bias	N_1:N_{con}	N_j
1	100	0.8	1.213	100	2.8	1.213	2	0	1:1	200
2	25	1.3	1.083	100	3.3	1.083	2	0	4:1	125
Total	125	0.9	1.219	200	3.05	1.205	2.15	**0.15**	8:5	325

classical group sequential designs, in which a decision is made to either stop or continue accrual, using interim data. In group sequential designs, bias arising from stopping when random highs are observed at the interim result has been well studied [11, 26]. Analogous to classical group sequential designs, we are concerned with bias arising from changing randomization ratios on the basis of random highs at an interim analysis. Both random high bias and regression to the mean are observed when the same data that is used to determine whether or not to modify the randomization ratio is subsequently included in a pooled efficacy analysis. Even in the absence of a true effect or trend, a spuriously significant, but truly ineffective, experimental arm could be inappropriately accelerated. Similarly, a spuriously unrepresentative control arm may result in all experimental arms appearing more or less favorable than they truly are. If the accrual rates are modified to favor accrual to these spuriously favorable experimental arms, there is less control arm data available to balance the aberrantly low interim estimates of the control outcome. This may lead to bias when the first stage results are over- or under-emphasized when included in the final pooled analysis of the data.

Consider the example presented in §2.2: if the mean of the first 100 control subjects accrued during stage one is spuriously low and a positive treatment effect is considered beneficial, then there is less opportunity for this random high bias to be balanced by additional data when only 25 subjects are accrued to control during stage 2 rather than the 100 subjects which would be accrued to the control arm in the absence of acceleration.

Table 2.4.3 presents simulation results of 1 million replicates of a trial with no true treatment effect ($\mu_0 = \mu_1 = 0$), and an O'Brien-Fleming level 0.20-level boundary used for acceleration. Conditional on acceleration, there is a marked bias of 0.256 in the stage 1 estimates for the 8.9 % of trials which are accelerated with this particular treatment effect and acceleration rule. In the 91.1% of trials not accelerated, there is a slight bias of -0.013 in the opposite direction. Weighting the stage 1 conditional bias by the probability of accelerating or not, we get an overall unbiased estimate of treatment effect from stage 1 ($.089 \cdot 0.256 - 0.911 \cdot 0.025 = 0$).

Of course, when conducting a trial, the probability of acceleration and the bias induced by the acceleration are unknown. Our primary concern with the random high bias focuses on how the stage 1 data will be pooled with unbiased stage 2 data. This will be further addressed when discussing stratified analyses in §2.7.

Table 2.4.3
Expected stage-wise sample sizes, means, and bias
Null of no treatment effect or trend, OBF20 acceleration
1,000,000 simulations

	Stage	N_{con}	μ_0	σ_0^2	N_1	μ_1	σ_1^2	Bias	$N_1 : N_{con}$	N_j
Accelerated	1	100	-0.128	1.000	100	0.128	1.000	0.256	1	200
(8.9% of simulations)	2	25	-0.000	1.000	100	-0.000	1.000	0.000	4	125
Non-accelerated	1	100	0.012	1.000	100	-0.013	1.000	-0.025	1	200
(91.1% of simulations)	2	100	0.000	1.000	100	0.000	1.000	0.000	1	200

2.5 Time as precision variable

As described in §2.2, if not accounted for, secular trends in the mean can increase the type I error rate, and result in biased coefficient estimates and anti-conservative inference. A second aspect of such a scenario is the effect on estimating study precision. In the presence of a secular time trend, we would need to adjust for calendar time to account for the prognostic value of that secular trend in predicting outcome. If randomization is blocked over time, then failure to adjust for that precision variable will lead to overestimation of standard errors. The loss of power due to the trend in the absence of any acceleration can easily be computed analytically (Appendix C). Figure 2.5.2 below shows the probability of rejecting the null hypothesis of no treatment effect in the absence of acceleration, as a function of the slope of the linear trend, for a variety of treatment effect sizes. Note that in the absence of a trend, type I error and power are achieved, and both of these operating characteristics decrease as the trend is increased.

2.6 Unequal Variances

In the model discussed so far, equal variances have been presumed for inference. However, as illustrated in 2.5, the presence of a secular trend will increase the observed variance of subjects collected over a period of time. When accrual is imbalanced between arms, this can lead to unequal variances between experimental and control arms (Table 2.3.2).

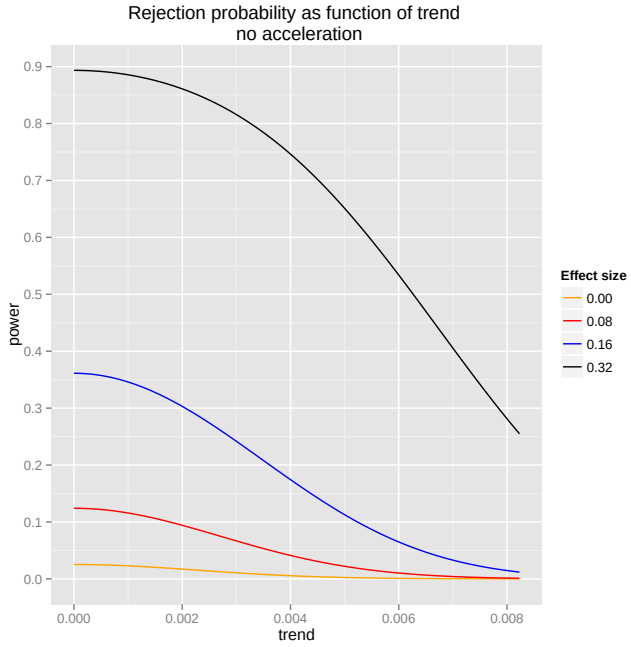

Figure 2.5.2
Rejection probability as a function of secular trend (analytic)

In the absence of acceleration, the secular trend affects both experimental and control arms equally, and the variance inflation is equal across arms. If the variances are equal, i.e. $\sigma^2_{ijk} = \sigma^2$ for all i, j, k, then the variances will be equal for the pooled analysis as well. However, when accrual is imbalanced across time, particularly when absolute sample sizes are modified to be unequal across arms, the observed variances between arms will also be unequal in the pooled analysis. For example, again consider the example discussed in §2.2. Consider the case with no trend and the experimental arm always accelerated. Table 2.2.1 displays both the stage-wise and cumulative variances. In the absence of a trend, there is no variance inflation in either arm, and the equal variances assumption holds.

However when a linear trend is present and the experimental arm is always accelerated, the observed variances of the two study arms are not equal. As discussed in §2.5, the increase in trend will inflate the observed variance. Stage-wise, the equal variances assumption holds. However, the variance during the first stage (in which 200 total are accrued) is larger than that of the second stage (in which only 125 are accrued). Similarly to how a trend induces bias in the estimation of the treatment effect by weighting the two stages differently for each of the two arms (§2.3), the pooled variance of the control arm will be weighted more heavily to the first stage than the experimental arm. Since the first stage has larger variance, the cumulative

control arm variance is thus larger than the cumulative experimental arm variance. This is reflected in Table 2.3.2.

The cumulative variance calculation is described in detail in Appendix A. The cumulative variance is not simply a weighted average of the two stage-wise estimates, but stage-wise sample sizes are directly proportional to the influence the stage-wise variances estimates have on the calculation of the cumulative variance. Appendix B details the calculation of the stage-wise variance associated with a linear trend in the mean of a sequence of identically distributed random variables.

A key feature of these data is the presence of staged sampling. An analysis that pools the data from both stages presumes that the variance to be used in computing the standard error is the variance of available data across the entire study. However, individuals were sampled within stages, and the true sampling variance is that of the within-stage data. As a subset of the larger set of study data, this variance will be less than or equal to the pooled variance. This overestimation of the sampling variance leads to decreased power if not properly accounted for.

2.7 Stratified analysis

After illustrating the existence of these biases with simulations, we show that a stratified analysis will correct for it. This does not alter the trial design, but stratifies the analysis on the stage of accrual. The bias resulting from the presence of a secular trend is primarily due to the imbalance in accrual rates over time across arms, and thus stratifying the analysis on treatment stage is expected to remove this source of bias by only estimating the treatment effect with concurrent experimental and control data. However, calendar time will still play the role of a precision variable, resulting in inflation of the observed variance within each stratum.

Stratifying the analysis by stage of accrual only addresses the bias related to imbalanced accrual rates, and will not correct for the random high bias discussed in §2.4. If the stage 1 estimate $\widehat{\theta}_1$ is spuriously high, then the pooled estimate will be biased as well. The extent of this bias depends on the weights used for the stratified analysis. Let $\epsilon_1 = \widehat{\theta}_1 - \theta$ be the residual of the estimate at the stage 1 analysis. Then given weights w_1, w_2, the bias b in the stratified estimate will be:

$$\begin{aligned} b &= E\left[\widehat{\theta}|\widehat{\theta}_1\right] - \theta \\ &= E\left[w_1\widehat{\theta}_1 + w_2\widehat{\theta}_2\right] - \theta \\ &= w_1\widehat{\theta}_1 - \theta + w_2 E\left[\widehat{\theta}_2\right] \\ &= w_1\epsilon_1 + 0 = w_1\epsilon_1 \end{aligned}$$

Conditional on the stage 1 residual being large enough to inappropriately accelerate an ineffective arm, $w_1 > w_2$ for both efficient and sample size weights. In contrast, conditional on no such bias being observed and no resulting acceleration, $w_1 = w_2$. Thus, the stage 1 data will always be weighted at least as much as the stage 2 data, and precisely when the observed estimate is spuriously high, the stage 1 data will be weighted more heavily than the stage 2 data. Thus the presence of random high bias leading to inappropriate acceleration can induce unequal weighting that favors the biased stage 1 estimate.

2.7.1 Weights

For the stratified analysis, two different sets of weights are considered. First, weighting by the ratio of stage-wise sample size to overall trial sample size (both experimental and control), and second using the most efficient weights, which can be derived as follows:

Let $j = 1, \ldots, J$ index the stage of the trial, let $k = 0, 1$ index control and experimental treatment arms respectively, and let $i = 1, \ldots, n_{jk}$ index the individuals accrued during stage j to arm k. Finally, let $Y_{ijk} \sim (\mu_{ijk}, \sigma^2_{jk})$ denote the observed treatment response for individual i, accrued to arm k during stage j, and let μ_{jk} be the mean outcome for all subjects accrued to treatment k during stage j. The stage-wise treatment effect parameter is then $\theta_j = \mu_{j1} - \mu_{j0}$, and can be estimated as $\widehat{\theta}_j = \overline{Y}_{j1\cdot} - \overline{Y}_{j0\cdot}$. By the CLT,

$$\widehat{\theta}_j \dot{\sim} N\left(\theta_j = \mu_{jk} - \mu_{j0}, V_j = \frac{\sigma^2_{jk}}{n_{jk}} + \frac{\sigma^2_{j0}}{n_{j0}}\right).$$

As discussed in §2.3, the overall treatment effect θ will be estimated with a stratified analysis, weighting the stages with weights w_j:

$$\widetilde{\theta} = \frac{\sum_{j=1}^{j} w_j \widehat{\theta}_j}{\sum_{j=1}^{j} w_j} \sim N\left(\theta, \frac{\sum w_j^2 V_j}{\left(\sum w_j\right)^2}\right)$$

We make a key assumption that $\theta_j = \theta$ for all j. This means that the treatment effect does not change over the course of the study, but does allow for the possibility of a secular trend that occurs independent of treatment (i.e. $\mu_{ijk} \neq \mu_{ij'k}$, for $j' \neq j$).

The optimal weights will be chosen as those which minimize the variance of the final treatment effect estimate $\widehat{\theta}$. Without loss of generality, impose the constraint that the weights sum to one: $\sum w_j = 1$. Then Lagrange multipliers can be used to minimize $\text{Var}_{\widehat{\theta}}$ with respect to $\mathbf{w} = (w_1, \ldots, w_J)$, subject to the following constraint:

$$Var_{\widehat{\theta}}(\mathbf{w}) = \sum_{j=1}^{J} \left[w_j^2 \left(\frac{\sigma_{jk}^2}{n_{jk}} + \frac{\sigma_{j0}^2}{n_{j0}} \right) \right] + \lambda \left[\left(\sum_{j=1}^{J} w_j \right) - 1 \right].$$

For the present purpose, it is sufficient to consider a single interim analysis ($J = 2$), and assume homoskedasticity ($\sigma_{jk}^2 = \sigma^2$). Then $w_2 = 1 - w_1$ and

$$V(\mathbf{w}) = w_1^2 \sigma^2 \left(\frac{1}{n_{1k}} + \frac{1}{n_{10}} \right) + (1 - w_1^2) \sigma^2 \left(\frac{1}{n_{2k}} + \frac{1}{n_{20}} \right)$$

The value of w_1 which maximizes this expression can be found by differentiating with respect to w_1, setting equal to zero, and solving for w_1. This solution can be expressed in terms of the harmonic means of the stage-wise sample sizes across arms, reflecting that the weights which minimize the variance of the estimate are proportional to the inverse variance of the stage-wise treatment effect estimates:

$$w_1 = \frac{\left(\frac{1}{n_{1k}} + \frac{1}{n_{10}} \right)^{-1}}{\left(\frac{1}{n_{1k}} + \frac{1}{n_{10}} \right)^{-1} + \left(\frac{1}{n_{2k}} + \frac{1}{n_{20}} \right)^{-1}}$$

The stratified treatment estimate $\widetilde{\theta}$ and test statistic are

$$\widetilde{\theta} = w_1\widehat{\theta}_1 + w_2\widehat{\theta}_2,$$
$$T = \frac{\widetilde{\theta}}{SE(\widetilde{\theta})}$$

where $SE(\widetilde{\theta}) = \sqrt{Var(\widetilde{\theta})} = \sqrt{\frac{\sum w_j^2 V_j}{(\sum w_j)^2}}$.

Substituting the appropriate samples sizes corresponding to the 2-arm design described in 2.1, give the following weights conditional on the experimental arm being accelerated. For trials not accelerated, two stages are weighted equally.

As there are only two acceleration paths considered (accelerated or not accelerated) and two types of weights, there are a total of four ways the trial results are weighted. For trials not accelerated, there is balance with respect to sample size, both between arms and between trial stages, and the two sets of weights are equivalent, placing equal emphasis on each stage. When the experimental arm is accelerated, more emphasis will be placed on stage 1 than on stage 2 for both sets of weights. Table 2.7.4 contains both sets of weights conditional on the trial being accelerated. Relative to the sample size weights, the efficient weights place more emphasis on the stage 1 results.

Table 2.7.4 again displays the results of the simulations shown in Table 2.4.3, and also shows the results of stratifying the analysis on accrual stage. Although, across acceleration patterns, the stage-wise estimates are unbiased, the stratified estimates are slightly biased. This bias arises from the heavier weights placed on the accelerated trials, which have more biased estimates relative to non-accelerated trials.

2.8 Two arm simulations

The two arm trial with the acceleration rule described in §2.1 was simulated $100,000$ times for each of 8 different acceleration thresholds. Outcomes were generated as $Y_{ijk} \sim N(\mu_{ijk}, 1)$. At the interim analysis, a t-statistic using a pooled variance estimate was computed, and if greater than a pre-specified threshold, the randomization ratio was changed to $4:1$ and an additional 100 were assigned to experimental and 25 to control for stage 2. Otherwise 100 were assigned to each arm for stage 2. The final analysis was conducted using pooled data and a two-sample t-test. Both Student's t-test (assuming equal variances) and Welch's

Table 2.7.4
Random High Bias conditional on acceleration (under H_0)
Acceleration probability of 0.152 under $H_1 : \mu_1 = 0.32$
(OBF 20 threshold)

		n_1	n_0	Sample size weights	Efficient weights	Bias
Accelerated	Stage 1	100	100	0.615	0.714	0.256
	Stage 2	100	25	0.385	0.286	0.000
	Strat (ss)					0.158
	Strat (eff)					0.183
Non-accelerated	Stage 1	100	100	0.500	0.500	-0.025
	Stage 2	100	100	0.500	0.500	0.000
	Strat (ss)					-0.0125
	Strat (eff)					-0.0125
Overall (weighted by acceleration probability)	Stage 1					0.000
	Stage 2					0.000
	Strat (ss)					0.0026
	Strat (eff)					0.0048

t-test (allowing for unequal variances) were used in final analysis for each simulation.

A variety of group sequential test sizes α were used to generate the acceleration threshold a. These thresholds are shown in Table 2.8.5 on both the sample mean and normalized Z-statistic scale. These α's do not have the usual false positive rate interpretation since a fixed sample test is used for inference at the final analysis, and the threshold α is only used to potentially modify the randomization procedure. However, the α's can be interpreted in the context of the probability of inappropriately accelerating an ineffective arm, with increasing α corresponding to lower acceleration thresholds and more aggressive (anti-conservative) acceleration. Note that all that truly matters is the value of the threshold. In the present 2-arm setting, the threshold is over-parameterized when looking at both the boundary shape and size. However, we include both parameters to appeal to those with intuition about such boundaries. Figure 2.8.3 shows the probability of accelerating an ineffective arm using both OBF and Pocock boundaries, for the group sequential test sizes simulated.

Table 2.8.5
Acceleration boundaries used for continuous model
(both sample mean and normalized Z scales)

Sample mean	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$
O'Brien-Fleming	0.47	0.38	0.27	0.19
Pocock	0.38	0.31	0.22	0.16
Normalized Z-statistic	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$
O'Brien-Fleming	2.37	1.90	1.35	0.97
Pocock	1.88	1.53	1.11	0.80

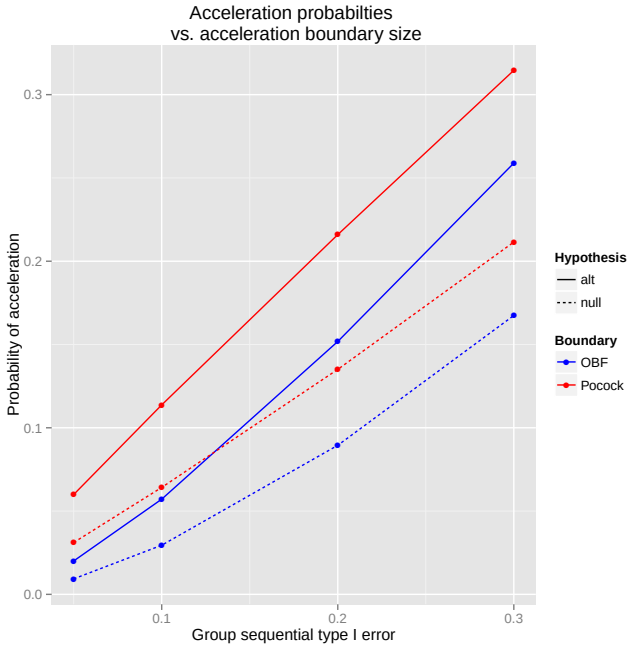

Figure 2.8.3
Probability of accelerating under both null and alternative
vs. Group sequential boundary size

Four efficacy scenarios are considered: all arms under the null hypothesis of no effect, and an effect in arm 1 only with effect sizes $\delta_1 \in \{0.32, 0.16, 0.08\}$. These effect sizes correspond to fixed sample t-test power of 90%, 36.7%, and 12.8% respectively. Table 2.8.6 displays these efficacy parameters.

Table 2.8.6
Efficacy parameters for two-arm normal model
corresponding power in absence of time trend

Efficacy	μ_0	μ_1	Fixed-sample power
Null	0	0	-
Strong effect	0	0.32	0.90
Medium effect	0	0.16	0.367
Weak effect	0	0.08	0.128

A variety of linear time trends in the mean outcome were considered, with $Y_{ijk} \sim \left(\mu_k + \beta i, \sigma^2\right)$, for $\beta \in \{0, 1/20000, 1/10000, 1/5000, 1/2000, 1/1000, 1/750, 1/500, 1/250, 1/125\}$. Note that a trend of $1/k$ corresponds to an increase in mean of 1 per k subjects accrued. This parameterization of the trend can be used to judge the magnitude of the trend with respect to the various alternatives with this design since $N \cdot \beta$ equals the observed increase in outcome over the course of the study, where N is the total study sample size. Thus, depending on acceleration, the trends considered range from an increase of $325 \cdot \frac{1}{20000} = 0.0163$

over the entire trial (5.08% of the effect detectable with 90% power) to $400 \cdot \frac{1}{125} = 3.2$ over the entire trial (1,000% of the effect detectable with 90% power). The more extreme trends are not intended to model any real-world scenario, but are included to more clearly illustrate issues that still arise at more reasonable trends, albeit at a more modest scale. The change in mean and proportion to effect size detectable with 90% power are shown for all trends in Table 2.8.7.

Table 2.8.7
Change in mean over study for various trends and proportion of effect size detectable with 90% power ($\Delta_{90} = 0.32$)

Trend	1/20000	1/10000	1/5000	1/2000	1/1000	1/750	1/500	1/250	1/125
	No Acceleration								
$\Delta\mu$	0.0200	0.0400	0.0800	0.2000	0.4000	0.5333	0.8000	1.6000	3.2000
$\Delta\mu/\Delta_{90}$	0.0625	0.1250	0.2500	0.6250	1.2500	1.6667	2.5000	5.0000	10.0000
	Acceleration								
$\Delta\mu$	0.0163	0.0325	0.0650	0.1625	0.3250	0.4333	0.6500	1.3000	2.6000
$\Delta\mu/\Delta_{90}$	0.0508	0.1016	0.2031	0.5078	1.0156	1.3542	2.0312	4.0625	8.1250

In total, all possible combinations of 4 unique efficacy scenarios, 10 linear trends, and 9 acceleration boundaries were simulated, for a total of 360 unique sets of simulation parameters. For each of these simulations, an unstratified analysis and two different stratified analyses were conducted, and for each of these analyses, two t-tests were used in the final analysis. Thus a total of 2,160 trials were simulated.

2.9 Discussion of simulations

For interpretability, only a portion of the simulation parameters are discussed in detail here. Observations and patterns discissed in this section are similar for the remaining simulations. Full results are shown in Appendix D (Figures D.0.1 and D.0.2).

Three acceleration boundaries are discussed in detail: no acceleration, OBF 20 which corresponds to moderate likelihood of acceleration, and P30 which corresponds to aggressive acceleration. Each of these trials are discussed for the extremes of the range of efficacy scenarios considered: under the null of $\mu_1 = 0$ and under the alternative of $\mu_1 = 0.32$. All results discussed in this section use a t-test assuming equal variance, except for the section on unequal variance.

2.9.1 No acceleration

For the simulation under the null of no treatment effect, without a trend or acceleration, results are as expected for a fixed sample test (Table 2.9.8). Type I error was maintained at 0.025, and no bias was observed. Similarly, under the alternative, 90% power was achieved and no bias was observed. Under both the null and alternative, when a trend was simulated without acceleration, loss of precision was observed to lower the rejection probability with increasing trend (Figure 2.9.4, Tables 2.9.9, 2.9.10). The two stratified analyses performed similarly, and corrected for the loss of precision at extreme trends observed in the unstratified analyses (Figure 2.9.4).

Table 2.9.8
Simulations under null ($\mu_1 = 0$) and alternative ($\mu_1 = 0.32$)
No trend or acceleration

Null	reject.prob	bias	accel.prob
unstratified	0.0243	0.0002	0
efficient	0.0244	0.0002	0
sample size	0.0244	0.0002	0
Alternative			
unstratified	0.8911	-0.0004	0
efficient	0.8930	-0.0004	0
sample size	0.8930	-0.0004	0

2.9.2 Acceleration

Figures 2.9.4 and 2.9.5 display the bias plotted against both trend and effect size, under the null hypothesis of $\mu_1 = 0$ and the alternative hypothesis of $\mu_1 = 0.32$. This same information is shown for select trends in Tables 2.9.9 and 2.9.10 When the trials were allowed to accelerate, both bias and inflation of rejection probability were observed for all efficacy scenarios when the analysis was not stratified. The observed bias increased with increasing trend, with a stronger relationship between trend and bias observed under the alternative of $\mu_1 = 0.32$ than the null. As the effect size increased, the observed bias increased for the trend of 1/1000, but when no trend was present, the bias was less pronounced, increasing initially and then declining at the larger effect sizes considered.

The rejection probability increased with trends for the lower trends considered, and then decreased as loss of precision from the magnitude of the trend overwhelmed the bias induced from accelerating in the presence of a trend. That is, the loss of precision due to not adjusting for an extremely important prognostic

variable of calendar time in a block randomized study proved to be more of a factor than the confounding bias introduced by the changing randomization ratio. As the acceleration became more aggressive, the magnitude of the bias and type I error increased. These results, along with the results of the stratified analyses, are summarized for select trends and acceleration boundaries in Table 2.9.9.

2.9.3 Stratified

In Figure 2.9.4, the bias of each of the three analyses are plotted against trend, under the null hypothesis of $\mu_1 = 0$ and the alternative hypothesis of $\mu_1 = 0.32$. When an analysis stratified on accrual stage was conducted, both bias and and type I error are significantly improved. Comparing the two stratified analyses, the sample size weights tend to perform slightly better than the efficient weights with respect to bias and type I error. Under the alternative, loss of power is observed when the analysis is stratified, but this loss appears steady over various trends. With more aggressive acceleration, the loss of power becomes more pronounced.

In Figure 2.9.5, bias is plotted against effect size for trends of 0, 1/10000, 1/1000, and 1/125, and for unstratified and stratified analyses. The bias from accelerating at the smaller of the two trends is small, and increases significantly with the the extreme trends.

2.9.4 Random high bias

In §2.4, we discussed the bias that can arise from spurious interim estimates being preferentially weighted in a stratified analysis Table 2.7.4. This random high bias can be estimated using the simulation results with no treatment effect or time trend. With no true effect or trend, any acceleration is due to spuriously favorable interim results, and there is no association between time and outcome to confound results. Table 2.9.11 shows the estimated bias and probability of accelerating under the null of no treatment effect, for the 8 acceleration boundaries simulated. The bias is shown within accelerated and non-accelerated arms, as well as overall. This random high bias, although present, appears to be relatively small compared to the alternative for which the study is powered. Among the scenarios considered, the highest bias under the null was 0.006 (simulation standard error 0.0003) for an acceleration threshold corresponding to a Pocock boundary with $\alpha = 0.2$. Such a boundary had an acceleration probabilty of 21% under the null. Similarly, the greatest bias of 0.08 under the alternative occurred when the acceleration probability was closest to 0.5

(thus having the greatest variability whether the treatment arm was accelerated). But, again, that bias due to using a stratified analysis in the absence of a trend over time is negligible compared to the bias seen in an unstratified analysis when a large secular trend in outcome exists.

2.9.5 Unequal variances

Figure 2.9.6 shows results from simulations using a t-test allowing for unequal variances. While unequal variances between experimental and control arms was shown to exist in §2.6, our simulations did not show a noticeable difference between when analyzing the data with a t-test assuming equal variances vs. a t-test allowing for unequal variances.

Table 2.9.9
Rejection probability, bias, and acceleration probability under null
Two-sample level 0.025 t-test assuming equal variances

		No acceleration		OBF20			P30		
Trend	Stratification	reject.prob	bias	reject.prob	bias	accel.prob	reject.prob	bias	accel.prob
No trend	unstratified	0.0243	0.0002	0.0372	0.0038	0.0893	0.0389	0.0062	0.2123
	efficient weights	0.0244	0.0002	0.0303	0.0053	0.0893	0.0278	0.0089	0.2123
	sample size weights	0.0244	0.0002	0.0252	0.0031	0.0893	0.0256	0.0048	0.2123
1/10000	unstratified	0.0246	-0.0002	0.0386	0.0038	0.0885	0.0433	0.0073	0.2128
	efficient weights	0.0246	-0.0002	0.0300	0.0048	0.0885	0.0288	0.0089	0.2128
	sample size weights	0.0246	-0.0002	0.0251	0.0026	0.0885	0.0259	0.0048	0.2128
1/1000	unstratified	0.0241	0.0000	0.0587	0.0080	0.0901	0.0786	0.0169	0.2120
	efficient weights	0.0250	0.0000	0.0299	0.0051	0.0901	0.0276	0.0091	0.2120
	sample size weights	0.0250	0.0000	0.0254	0.0028	0.0901	0.0251	0.0050	0.2120
1/500	unstratified	0.0228	0.0003	0.0766	0.0123	0.0879	0.1257	0.0272	0.2121
	efficient weights	0.0257	0.0003	0.0300	0.0051	0.0879	0.0277	0.0091	0.2121
	sample size weights	0.0257	0.0003	0.0251	0.0029	0.0879	0.0256	0.0050	0.2121
1/125	unstratified	0.0039	-0.0000	0.0698	0.0292	0.0682	0.1887	0.0794	0.1882
	efficient weights	0.0254	-0.0000	0.0303	0.0036	0.0682	0.0274	0.0081	0.1882
	sample size weights	0.0254	-0.0000	0.0252	0.0017	0.0682	0.0244	0.0043	0.1882

Table 2.9.10
Rejection probability, bias, and acceleration probability under alternative ($\mu_1 = 0.32$)
Two-sample level 0.025 t-test assuming equal variances

Trend	stratification	reject.prob	bias	reject.prob	bias	accel.prob	reject.prob	bias	accel.prob
No trend	unstratified	0.8911	-0.0004	0.8823	0.0056	0.8205	0.8661	0.0034	0.9263
	efficient weights	0.8930	-0.0004	0.8353	0.0081	0.8205	0.7941	0.0047	0.9263
	sample size weights	0.8930	-0.0004	0.7988	0.0044	0.8205	0.7672	0.0030	0.9263
1/10000	unstratified	0.8916	-0.0003	0.8867	0.0094	0.8182	0.8731	0.0075	0.9269
	efficient weights	0.8926	-0.0003	0.8349	0.0078	0.8182	0.7899	0.0041	0.9269
	sample size weights	0.8926	-0.0003	0.7986	0.0041	0.8182	0.7631	0.0022	0.9269
1/1000	unstratified	0.8885	0.0003	0.9172	0.0463	0.8208	0.9262	0.0485	0.9271
	efficient weights	0.8924	0.0003	0.8356	0.0084	0.8208	0.7898	0.0044	0.9271
	sample size weights	0.8924	0.0003	0.7994	0.0047	0.8208	0.7634	0.0025	0.9271
1/500	unstratified	0.8806	-0.0003	0.9237	0.0865	0.8183	0.9498	0.0936	0.9274
	efficient weights	0.8920	-0.0003	0.8368	0.0085	0.8183	0.7905	0.0039	0.9274
	sample size weights	0.8920	-0.0003	0.7994	0.0048	0.8183	0.7630	0.0020	0.9274
1/125	unstratified	0.6940	-0.0004	0.8683	0.3133	0.7834	0.9352	0.3621	0.9156
	efficient weights	0.8912	-0.0004	0.8454	0.0089	0.7834	0.7948	0.0042	0.9156
	sample size weights	0.8912	-0.0004	0.8079	0.0047	0.7834	0.7661	0.0020	0.9156

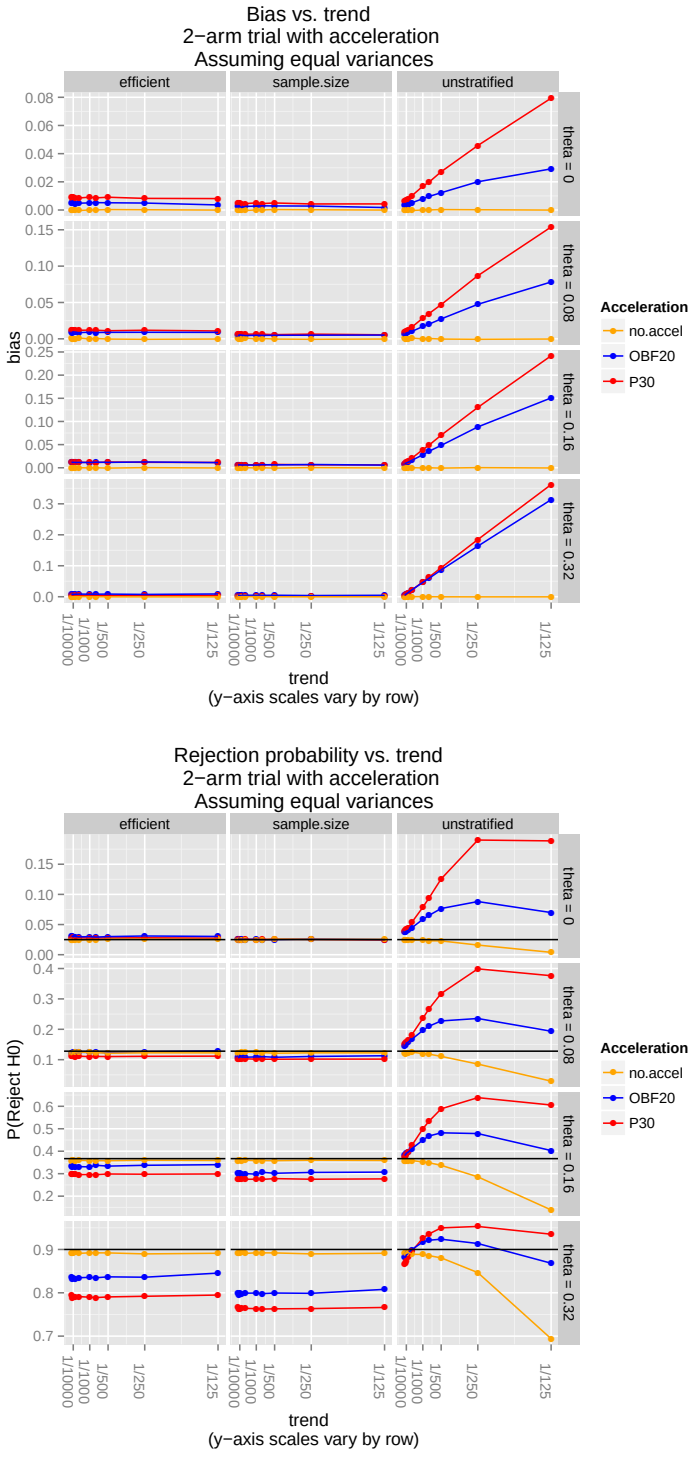

Figure 2.9.4
Bias and rejection probabilities as function of trend
Both unstratified and stratified analyses

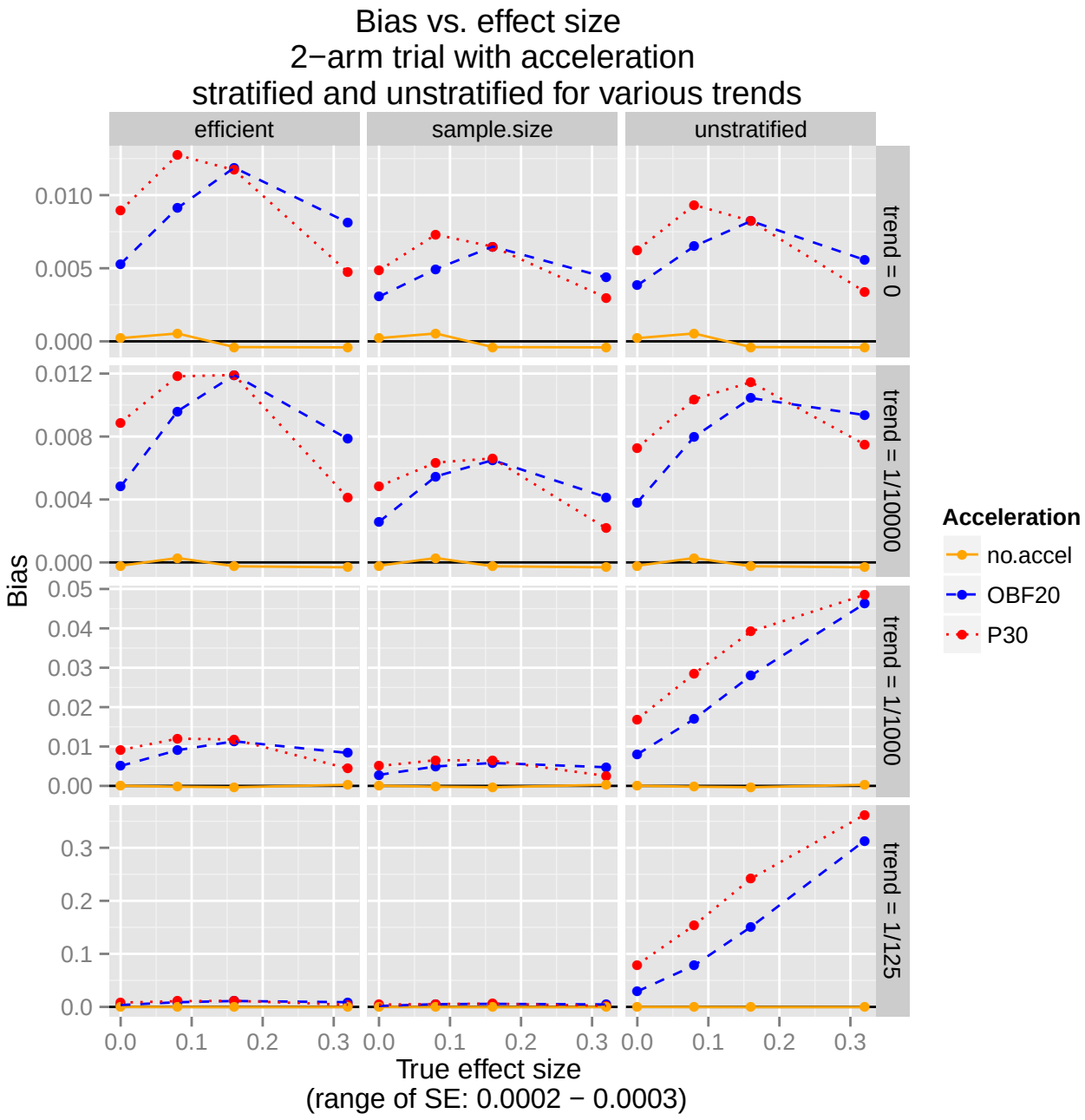

Figure 2.9.5
Bias versus effect size
Both unstratified and stratified analyses for four different trends

Table 2.9.11
Random high bias and acceleration probabilities with no trend
(100,000 simulations under each of null and alternative hypotheses)

Under Null ($\mu_1 = 0$)

Random High Bias	OBF05	OBF10	OBF20	OBF30	P05	P10	P20	P30	No Accel
Total	0.001	0.001	0.004	0.005	0.001	0.003	0.004	0.006	0.000
Accelerated	0.242	0.206	0.166	0.139	0.207	0.179	0.147	0.126	-
Non Accelerated	-0.002	-0.005	-0.012	-0.022	-0.006	-0.009	-0.018	-0.026	0.000
Acceleration probability									
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Accelerated	0.009	0.030	0.089	0.166	0.031	0.064	0.134	0.212	0.000
Non Accelerated	0.991	0.970	0.911	0.834	0.969	0.936	0.866	0.788	1.000
Standard Error									
Total	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
Accelerated	0.0002	0.0002	0.0002	0.0003	0.0002	0.0002	0.0002	0.0003	
Non Accelerated	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003

Under Alternative ($\mu_1 = 0.32$)

Random High Bias	OBF05	OBF10	OBF20	OBF30	P05	P10	P20	P30	No Accel
Total	0.008	0.008	0.006	0.004	0.008	0.006	0.004	0.003	-0.000
Accelerated	0.079	0.053	0.029	0.018	0.052	0.036	0.021	0.014	-
Non Accelerated	-0.052	-0.073	-0.102	-0.124	-0.074	-0.093	-0.117	-0.133	-0.000
Acceleration probability									
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Accelerated	0.454	0.644	0.820	0.903	0.653	0.766	0.878	0.926	0.000
Non Accelerated	0.546	0.356	0.180	0.097	0.347	0.234	0.122	0.074	1.000
Standard error									
Total	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
Accelerated	0.0002	0.0002	0.0002	0.0003	0.0002	0.0002	0.0002	0.0003	
Non Accelerated	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003

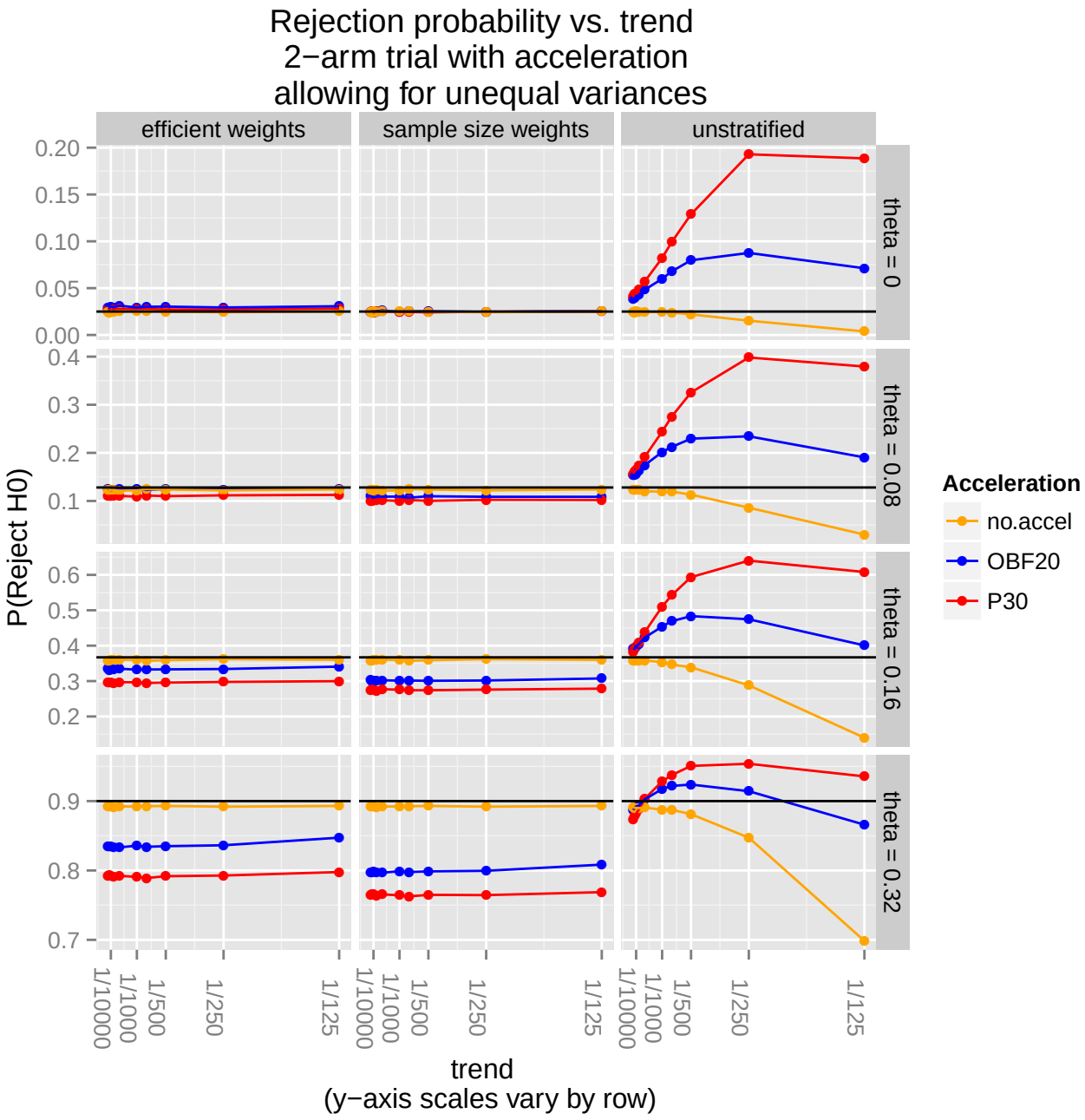

Figure 2.9.6
Rejection probabilities as function of trend
t-test allowing for unequal variances

Chapter 3

Multi-arm studies

3.1 Design and notation

Now we generalize the two-arm design discussed in the preceding section to a five-arm clinical trial with one control arm, four experimental arms, and one planned interim analysis. Again let Y_{ijk} (independent) be the i^{th} observation during the j^{th} stage of arm k ($k = 0, 1, 2, 3, 4$), and assume that $Y_{ijk} \sim \left(\mu_k, \sigma_k^2\right)$ are the first two moments of its distribution. Without loss of generality, assume a planned sample size of 200 per arm (to each experimental and control), and that a single interim analysis is conducted after 100 subjects are accrued to each experimental arm. Let the initial randomization ratio be 1:1:1:1:1, with blocked randomization guaranteeing 100 subjects are accrued to each experimental and control during stage 1.

As before, t-statistics T_{1k} using a pooled variance estimate will be computed at the interim analysis, comparing each experimental treatment to control. If any of the interim test statistics T_{1k} are larger than some threshold a, those experimental arms will be considered promising, and randomization in the second stage will favor their accrual. An additional 100 will be accrued to each promising arm to a total of 200, and 20% will be maintained on the control arm, regardless of the number of experimental arms continuing. Any arms not crossing the acceleration threshold will be postponed until the promising arms are completed, at which point accrual to 200 total per experimental arm will resume (again, 20% accrual to control will be maintained). If no arms cross the acceleration threshold, then no adaptation will occur and a total of 200 will be accrued to each arm.

In contrast to the two-arm design previously considered, this multi-arm design focuses more on group ethics than individual ethics. A total of 200 subjects are accrued to each arm, regardless of interim results. The modification alters the time at which each arm is ongoing, but each arm (including control) accruing 200 subjects. However the acceleration of promising arms to finish more quickly favors more rapid progression of truly effective therapies through the drug discovery process.

3.2 Simple example (multi-arm)

Now expand the illustration given in Section 2.2 to the five-arm trial. Assume the true expected outcome is zero in all arms except for arm 1, which is assumed to have a average treatment effect of 2: $\mu_1 = 2$, $\mu_k = 0$ $(k = 2, 3, 4)$. With a planned maximum sample size of 200 per arm, and interim analysis conducted after 100 subjects are accrued to each arm, initially randomize in a 1:1:1:1:1 ratio, but accrue 20% of subjects to control regardless of overall stage-wise sample size. Assume that arm 1 alone is accelerated (regardless of observed outcome) and that the rest are postponed. Table 3.2.1 displays the sample size and expected response in each arm at each stage.

Table 3.2.1
Expected stage-wise sample sizes and means with large effect in arm 1 and no trend

Stage	N_{con} (μ_0)	N_1 (μ_1)	N_2 (μ_2)	N_3 (μ_3)	N_4 (μ_4)	$N_{tot}(\mu)$	ratio (tx:ctl)
1	100 (0)	100 (2)	100(0)	100(0)	100(0)	500	1:1
2	25 (0)	100 (2)	0	0	0	125	4:1
3	75 (0)	0	100(0)	100(0)	100(0)	375	4:3
Total N (μ)	200 (0)	200 (2)	200 (0)	200 (0)	200 (0)	1000	-

Now consider a time trend in the mean, where the average response increases by 1 every 125 subjects. Note this increase occurs regardless of treatment group assignment. The mock data in Figure 3.2.1 illustrates this sampling scheme with an extreme time trend. The individual outcomes are color coded by treatment stage, and stage wise means are marked on the y-axis. Arm 1 data consists of 20% of the stage 1 data and 80% of the stage 2 data (100 from each stage). Control data consists of 20% of the first stage data and 20% of the second stage data. Controls are more likely to be accrued during the first stage, while arm 1 subjects are balanced between stage 1 and stage 2. In the presence of a trend, this biases the average Arm 1 outcome $Y_{\cdot\cdot 1}$ upwards relative to the control arm outcome $Y_{\cdot\cdot 0}$ accrued during the same time period.

Table 3.2.2 shows the expected stage-wise and cumulative means in a possible trial pathway, where Arm

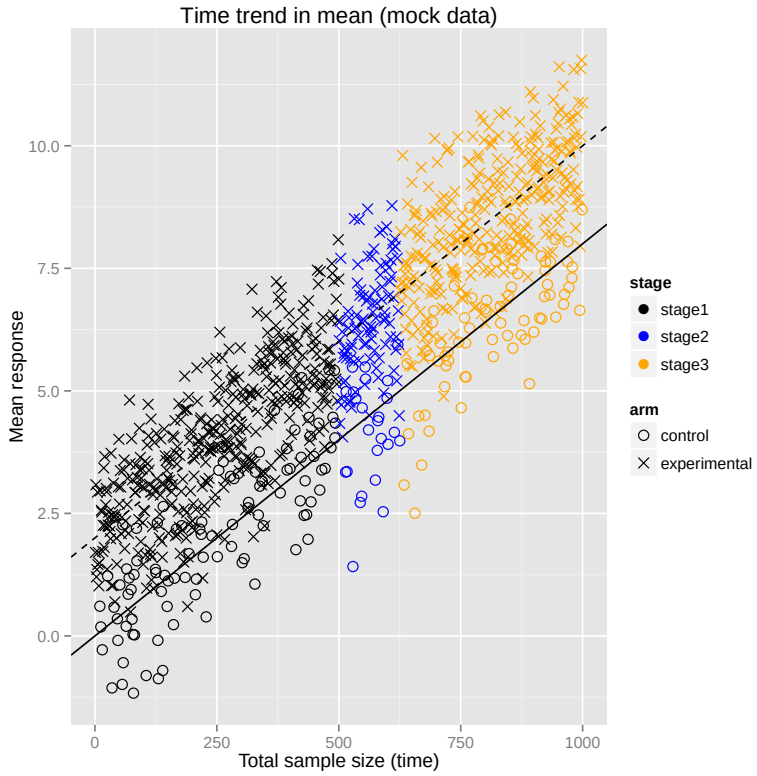

Figure 3.2.1
Outcome of interest vs. subject number, colored by stage
Mock data

1 advances to stage 2 while the remaining are postponed and continue in stage 3. A linear trend with a slope in the mean of $\beta = \mu_{(i+1)jk} - \mu_{ijk} = \frac{1}{125}$ (increase of 1 per 125 subjects accrued to the trial) is assumed to begin immediately upon the start of accrual, and experimental arms are only compared to control arm data accrued during the same stage.

As with the two-arm design, the presence of a trend exacerbates the issue of random high bias in the control arm. In the scenario above, if the mean of the first 100 control subjects accrued during stage one is spuriously high or low, there is less opportunity for this effect to be balanced out with only 25 subjects accrued during stage 2 than if 100 were accrued during stage 2.

Table 3.2.2
Expected stage-wise sample sizes and means with large effect in arm 1, with secular trend

Stage	N_{con} (μ_0)	N_1 (μ_1)	N_2 (μ_2)	N_3 (μ_3)	N_4 (μ_4)	$N_{tot}(\mu)$	ratio (tx:ctl)
1	100 (2)	100 (4)	100(2)	100(2)	100(2)	500	1:1
2	25 (4.5)	100 (6.5)	0	0	0	125	4:1
3	75 (6.5)	0	100(6.5)	100(6.5)	100(6.5)	375	4:3
Total N (μ)	-	200 (5.25)	200 (4.25)	200 (4.25)	200 (4.25)	1000	-
Control N	-	125 (2.5)	175 (3.93)	175 (3.93)	175 (3.93)		
True effect	-	2	0	0	0	-	-
Pooled effect	-	2.75	0.32	0.32	0.32	-	-
Bias	-	**0.75**	**0.32**	**0.32**	**0.32**	-	-

3.3 Simulations

To investigate the biases described above, 100,000 trials were simulated for various trends, treatment effects, and acceleration rules, using the design described in §3.1. The trends and treatment effects used are summarized in Table 3.3.3. Homoskedasticity is assumed, with $\sigma_k^2 = 1$ for all k. At the single interim analysis, arms were accelerated if the interim test statistic crossed a group sequential efficacy boundary. Both O'Brien-Fleming and Pocock acceleration boundaries were considered, each with sizes 0.05, 0.10, 0.20, and 0.10 (Table 2.8.5). Final decisions were based on a two-sample t-test assuming equal variance. Any control arm data accrued during a stage in which an experimental arm was not active was not used in the final analysis of that experimental arm.

The same linear time trends in the mean outcome as in the two-arm simulations were considered. Table 3.3.3 shows four efficacy scenarios used for simulations with normal data. As before, the effect sizes 0.32, 0.16, and 0.08 correspond to fixed sample t-test power of 90%, 36.7%, and 12.8%.

Table 3.3.3
Efficacy parameters for multi-arm normal model

Efficacy	μ_0	μ_1	μ_2	μ_3, μ_4
All arms under null	0	0	0	0
Arm 1 efficacious Arms 2,3,4 under null	0	0.32	0	0
Arms 1 and 2 efficacious Arms 3 and 4 under null	0	0.32	0.08	0
Arms 1 and 2 efficacious Arms 3 and 4 under null	0	0.32	0.16	0

3.3.1 Code

Simulations were performed using `R` code modified from the `RCTdesign` package [7, 21]. The function `rSeqMean2` was modified in two key ways. First,`rSeqMean5` simulates 5-arm clinical trials (4 experimental arms and 1 control), rather than 2-arm trials. Second, `rSeqMean5` only allows a single interim analysis, at which experimental arms may be postponed or accelerated depending on their performance in the first arm (`rSeqMean2` does not allow for any adaptive modification, other than group sequential stopping rules).

To improve efficiency of the code, rather than generating individual subject-level data, stage-wise sufficient statistics $(\widehat{\mu}_{jk}, \widehat{\sigma}_{jk})$, were generated for each simulated trial. Applying the Central Limit Theorem, the stage-wise sample averages were simulated from a $N(0,1)$ distribution using the `rnorm` function in `R`, and scaled to the appropriate means and variances: $\overline{Y}_{jk} \sim \frac{\sigma_{jk}}{\sqrt{n_{jk}}} N(0,1) + \mu_{jk}$. While the assumption of a normally distributed outcome is not needed to simulate sample means, it is necessary in order to rely on the asymptotic distribution of the sampling variance. The code is partly vectorized, and generates the data for each stage using matrix operations and no loops. However, since the sampling distribution of all but the first stage depends on previous stages, looping over the stages was unavoidable. There is also a loop over the simulations to determine which arms are accelerated and which are postponed.

At the interim analysis, the statistic T_{1k} is compared to the efficacy boundary of a group sequential design. As described in Section 2.8, a variety of designs were used to generate the acceleration boundaries including both O'Brien-Fleming and Pocock at several sizes. The `seqDesign` function from the `RCTdesign` package was used to generate these acceleration boundaries.

At the final analysis, the cumulative estimates of the means and variances were calculated from the stage-wise estimates and sample sizes [Appendix B]. For each experimental arm, only control data that was accrued at the same time was used. For example, control data accrued during a stage in which an arm is accelerated is not used as a comparison for the arms which were postponed.

The `R` code `rSeqMean5` was validated by comparison to analytic results, using an acceleration option built into `rSeqMean5` that forces arm 1 to continue to stage 2 while postponing the rest, regardless of observed outcomes. At the first stage, 100 are accrued to each arm. During the second stage, 100 are accrued to arm 1 and 25 to control to maintain a 20% accrual rate to control. These simulations match the analytic results over the range of trends considered (increase in mean per subject of 1/20000 to 1/125) [Appendix C].

3.4 Results

In the simulations, we see the patterns of bias described in Chapter 2: Unstratified analyses produce bias in the presence of a secular trend, but that bias is overwhelmed by the lack of precision when we fail to adjust for an extremely prognostic variable of calendar time that is used in blocked randomization. When two treatment arms each have some treatment effect, these patterns are observed in each of those arms. Plots of type I error and bias versus trend are shown in Figure 3.4 for the Pocock30 acceleration boundary and three sets of efficacy parameters (other boundaries displayed similar patterns). The efficacy parameters chosen correspond to the first, second, and fourth rows in Table 3.3.3. The two top panels show the observed rejection probability versus simulated trend, with an unadjusted analysis on the left and an adjusted analysis on the right. The bottom two panels show the observed bias versus simulated trend, again with an unadjusted analysis on the left and an adjusted analysis on the right.

All of the arms simulated under the null hypothesis of no effect display a similar trend in type I error as a function of trend. In the upper left panel of Figure 3.4 (unstratified analyses), the planned type I error is achieved ($\alpha = 0.025$) in the absence of a trend, is inflated at the weaker trends of 1/2000 and 1/1000, and drops to nearly 0 at the stronger trends of 1/250 and 1/125. The type I error curves for arms 3 and 4 (always simulated under null) show similar patterns but vary in magnitude with acceleration boundary type and size (data not shown), as well as in the number of effective arms and the magnitude of the effect in arm 2 (data not shown).

In the simulations with an non-zero treatment effect in arm 1 (eff1), power of nearly 0.90 was achieved in the absence of a trend. As the strength of the linear trend increased to 1/250, the power to detect a difference fell to 0.60. Treatment arm 1 behaved similarly in each of the two scenarios with an effect in arm 2, with power achieved in absence of trend, and declining to 0.84 as the trend increased to 1/250 (Figure 3.4 shows the results when the effect in treatment arm 2 is half that of treatment arm 1).

In the absence of a trend, arm 2 achieves the expected power of 0.367 and 0.128 for effect sizes of $\delta_2 = 0.16$ (Figure 3.4) and $\delta_2 = 0.08$ (data not shown) respectively. As trend increases to 1/250, arm 2 is rejected less often: $26\% - 28\%$ of the time for effect size of $\delta_2 = 0.16$ and $11.5\% - 16.0\%$ of the time for effect size of $\delta_2 = 0.08$ (increasing with size of acceleration boundary used).

Controlling for all other simulation parameters, the O'Brien-Fleming acceleration boundaries resulted in a lower inflation of type I error than the Pocock boundaries of the same size, reflecting the early conservatism

of the O'Brien-Fleming boundary. For both acceleration boundaries, increasing the size of the acceleration boundary resulted in increased type I error as the threshold for acceleration is lowered with increasing size. The increased type I error observed in the ineffective arms 3 and 4, as the effect size of arm 2 is increased also illustrates the effect of the imbalanced randomization over time in the presence of a trend on the type I error. When only one arm is accelerated, there is a $200 : 175$ ratio (1:1 and 8:7 per stage) for each postponed experimental arm to the control. However, when two arms are accelerated, there is a $200 : 150$ (1:1 and 4:3).

The righthand column of Figure 3.4 shows the results of stratified analyses. A plot of the stratified rejection rates and bias versus trend for the most anti-conservative acceleration boundary used is shown in the right column of Figure 3.4. It should be noted that the y axis scale for the bias is an order of magnitude lower for the stratified analyses (lower right panel) compared to that for the unstratified analyses (lower left panel). The type I error rate of 0.025 is attained for arms under the null hypothesis, and appropriate power is achieved for the arms with non-zero effects. As expected, the effects of random high bias from a stratified analysis in the absence of a trend (see §2.7.1) is not discernible on these graphs.

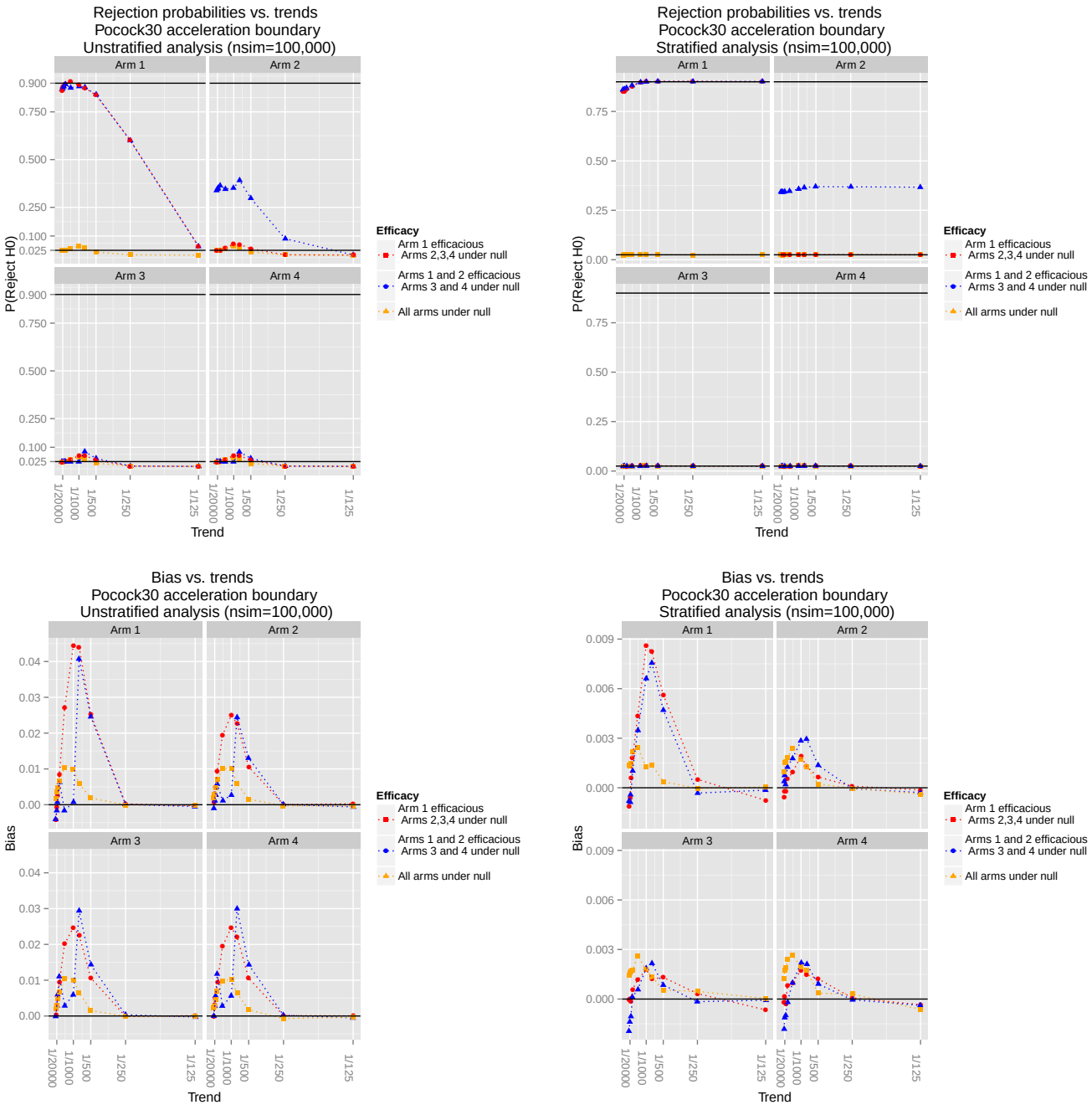

Figure 3.4.2
Normal model: Pocock 30 acceleration boundary for all panels
Top left: rejection probability v. trend, unstratified analysis
Top right: rejection probability v. trend, stratified analysis
Bottom left: Bias v. trend, unstratified analysis
Bottom right: Bias v. trend, stratified analysis

Chapter 4

Survival data

In the previous chapters, we considered the effect on RCT operating characteristics of adaptive changes to randomization ratios in the setting of immediately observed outcomes. With time to event data, however, there will be some differences due to the delayed ascertainment of outcomes: some subjects accrued prior to the adaptive interim analysis will have censored observations when making the decision to change the randomization ratio. Later in the study, those subjects may have an observed event. Thus even when there is a secular trend in the types of patients accrued over time, the statistical information added after the adaptive analysis will come in part from subjects accrued prior to the change in randomization ratio. We would thus expect that any biasing effect of adaptively changing randomization ratios will be less with a time to event endpoint. In this chapter, we briefly illustrate that behavior in unadjusted analyses.

4.1 Data and design

Similar to the continuous outcome case discussed in Chapter 2, consider a 2-arm clinical trial with one experimental arm, one control arm, and 1 planned interim analysis. Let Y_{ijk} be a binary event indicator for the i^{th} observation during the j^{th} stage of treatment arm k. Assume that a total of n_E events are to be observed, and that the interim analysis is conducted when one-third of the planned number of events are observed. Prior to the interim analysis, blocked randomization in a 1:1 ratio will be used.

At the interim analysis, a Cox proportional hazard model will be be used to compare the experimental

arm to the control. As with the continuous outcome model, a group sequential efficacy boundary will be used for accelerating arms: promising arms will be accelerated by modifying the randomization ratio in favor of accrual to promising experimental arms. Subjects will continue to be followed for events until a total of n_E events are observed (across both arms).

Assume uniform accrual, so that the entry time on arm k is uniformly distributed $T_{Ek} \sim U(0,1)$. Assuming survival times on arm k of $T_{Sk} \sim \exp(\lambda_k)$, the calendar time from start of study (not entry on study) to event is then $T_{Dk} \sim T_{Ek} + T_{Sk}$. Furthermore, assume independent entry and survival times. Note that the assumption of uniform entry times only holds in the absence of acceleration. When accelerating, the the distribution of entry times overall is still uniform, but conditional on the treatment arm, the entry times may be skewed.

Let p_k be the probability that a subject on arm k has an event during the first half of the study, i.e. $p_k = P\left(T_{Dk} < \frac{1}{2}\right)$. Let $V_i \sim \text{Bernoulli}(p_1)$ and $W_i \sim \text{Bernoulli}(p_0)$ so that $V = \sum_{i=1}^{n_1} V_i \sim \text{Bin}(n_1, p_1)$ is the number of experimental subjects experiencing an event during the first half of the study. Similarly define $W \sim \text{Bin}(n_0, p_0)$. For simplicity, assume that $n_0 = n_1$. Then the desired number of subjects to accrue so that n_E events are expected to be observed during the study period is

$$E[V+W] = \frac{n}{2}(p_1+p_0) = \frac{2}{3}n_E \Rightarrow n = \frac{2 \cdot n_E}{3(p_1+p_0)}.$$

The probabilities p_k can be derived by integrating the joint density of T_{Ek} and T_{Sk}, which is simply the product of the individual densities (by independence). For notational ease, let $E = T_{Ek}$, $S = T_{Sk}$, $D = T_{Dk}$, and $\lambda_k = \lambda$. Then $D < \frac{1}{2} \Leftrightarrow E + S < \frac{1}{2}$.

$$\begin{aligned} f_E(t) &= 1, t \in (0,1) \\ f_S(s) &= \lambda e^{-\lambda s}, s \in (0,\infty) \\ f_{E,S}(t,s) &= f_E(t) \cdot f_S(s) = \lambda e^{-\lambda s}, t \in (0,1), s \in (0,\infty) \end{aligned}$$

The desired probability is

$$p = \int_0^{\frac{1}{2}} f_D(t)dt = \int_R f_{E,S}(t,s)\text{dt ds} = \frac{e^{-\frac{\lambda}{2}} - 1}{\lambda} + \frac{1}{2},$$

where the latter integration is over the region $R = \left\{(t,s)|t \in (0,\frac{1}{2}), s \in (0,\frac{1}{2}-t)\right\}$. The full details of this calculation can be found in Appendix E.

4.2 Sources of bias

4.2.1 Random high bias

The sources of bias for the time-to-event model are similar to those in the continuous outcome model. For example, random high bias may cause truly ineffective arms to be accelerated. Also, if the randomization ratio is dependent on calendar time, the a secular trend may confound an analysis of pooled data at the end of the study and lead to bias analogous to that observed in the continuous data model (§2.3). However, unlike the immediate outcome observed with the continuous data model, the accumulating events observed during stage 2 from individuals censored at the interim analysis may ameliorate spurious interim results, attenuating this bias.

4.2.2 Accumulating events

In contrast to the continuous outcome model, events from individuals censored at the interim analysis continue to accumulate over stage 2 and should be properly accounted for in the final analysis. In the design we are considering, the decision to adaptively modify the study is based on the primary endpoint used in the efficacy analysis, and this correlation is not a concern. Concerns do arise, however, in the presence of a secular trend. The analysis of data from individuals with a low hazard rate in stage 1 (relative to individuals accrued during stage 2) is used to decide whether or not to modify the randomization procedure. During stage 2, events from individuals with low baseline hazard who are censored at stage 1 will continue to accumulate. Under the null, equal number of accumulating events expected from both treatment and control. But if the randomization ratio is modified to favor accrual to the experimental arm, then an equal number of accumulated events from individuals with lower baseline hazard are expected to be pooled with stage 2 data, which contains more experimental arm data than control arm data. Thus when accelerating, the final

control arm data will contain a disproportionately higher number of individuals with events accumulating from the stage 1 accrual, relative to the experimental arm.

We note that there are additional issues that would have to be considered in an adaptive design using time to event data. Bauer and Posch pointed out that information available at the adaptive analysis about secondary endpoints for censored observations may lead to correlations between the first stage data and the second stage data. Jenkins, et al. and Irle and Schafer have proposed solutions to this issue [14, 17]. In our investigations we pretend there are no such concerns, although in practice this would not be true.

4.3 Code

The function `rSeqHaz` was written to simulate an arbitrary number of trials `nsim` with Weibull data, and analyzes the data with either exponential regression, weibull regression, or Cox proportional hazards regression (using either `survreg` or `coxph` from the `survival R` package). The following must be specified in the function call: baseline hazard λ_0, Weibull shape parameter $k = 1$, true log-hazard ratio β, desired `power`, and the type of regression to use in the analysis. Optional arguments are the slope for a linear trend in the baseline hazard and an acceleration threshold (on the normalized z-statistic scale). If $\beta = 0$, then the `hypothesis` to be used for sample size calculations must also be specified.

4.3.1 Sample size and data generation

A complete derivation can be found Appendix E. Briefly, the number of events need to detect a log-hazard ratio of β with desired power is computed as follows:

$$n_E = \frac{4\left(Z_{1-0.025} + Z_{\text{power}}\right)^2}{\beta^2} \tag{4.1}$$

This is then used to compute the number of subjects which should be accrued to observe an expected $n_E/3$ events halfway through the accrual period, given exponential survival times with mean survival times λ_1 and λ_0 for the experimental and control arms, respectively.

$$n = \frac{2n_E}{3}\left[\lambda_1 e^{-\frac{\lambda_1}{2}-1} + \lambda_0 e^{-\frac{\lambda_0}{2}-1} + 1\right]^{-1} \tag{4.2}$$

4.3.2 Notes on parameterization

Accrual times are generated as Unif$(0, 1)$ for all subjects and simulations using `runif()`. The survival times for all subjects and simulations are initially generated as Exp(1) using `rexp()` and scaled appropriately throughout the simulation. We choose to parameterize the exponential distribution by the mean λ. This is done to match the parameterization of the Weibull distribution implemented in the `R` function `rweibull()` when $k = 1$, but does not match the `R` function `rexp()`, which parameterizes by the rate. The two probability density functions used are shown below:

$$f_{exp}(x) = \frac{1}{\lambda} \exp\{-x/\lambda\} \tag{4.3}$$

$$f_{weib}(x) = \frac{k}{\lambda^k} x^{k-1} \exp\left\{-(x/\lambda)^k\right\} \tag{4.4}$$

Let $Y \sim Exp(1)$. Then a random variable $X \sim Weib(\lambda, k)$ can be generated using the following transformation:

$$X = \lambda \left(Y^{1/k}\right) \tag{4.5}$$

Note that the median ratio (MR) and the hazard ratio (HR) are the same for Weibull data when the shape parameter $k = 1$. However, when $k \neq 1$, these two summary measures differ, but are perfectly correlated:

$$median(X) = \lambda \log(2)^{1/k} \Rightarrow MR = \frac{\lambda_1}{\lambda_0}$$

$$\begin{aligned} \text{hazard}(X) &= -\frac{f_X(x)}{S_X(x)} = -\frac{\frac{1}{\lambda^k} k x^{k-1} \exp\{-(x/\lambda)\}^k}{\exp\{-(x/\lambda)\}^k} = -\frac{1}{\lambda^k} k x^{k-1} \\ \Rightarrow HR &= \frac{-\frac{1}{\lambda_1^k} k x^{k-1}}{-\frac{1}{\lambda_0^k} k x^{k-1}} = \frac{\lambda_0^k}{\lambda_1^k} = \left(\frac{\lambda_0}{\lambda_1}\right)^k \end{aligned}$$

Thus, $\log(HR) = k \cdot \log(\frac{\lambda_0}{\lambda_1}) = -k \cdot \log(\frac{\lambda_1}{\lambda_0}) = -k \log(MR)$. The function `survreg()`, gives estimates for the MR (for both exponential and Weibull distributions), while the function `coxph()` gives estimates for

the HR. When $k = 1$, the data is exponential and the MR and HR coincide for the parametric models. But when $k \neq 1$, there is a non-unity scaling factor between the two summary measures (on the log-scale), which is equal to the negative of the shape parameter. The function `rSeqHaz` uses the hazard ratio to summarize treatment effect both in the argument (β) and the estimates of the hazard ratio are returned. If $k \neq 1$, then the specified coefficient β is transformed as above and data is generated to yield the median ratio corresponding to the specified hazard ratio.

When the Weibull shape parameter $k = 1$, the simulated data is exponential with the specified rate parameter λ. However, if $k \neq 1$, the Weibull rate parameter is re-computed so that the median of the resulting Weibull data is the same as that of the planned exponential data. This is to demonstrate that issues can arise with non-constant hazard, even when the median of the two comparison groups are equal.

4.3.3 Coding details

First, the total maximum sample size n is computed using equation 4.2. The stage 1 sample size is set to $n/2$, rounding down if n is odd. Blocked randomization with a 1:1 ratio is used to assign subjects to treatment arm for stage 1, with the extra randomly assigned if the stage 1 sample size is odd.

The code then loops over simulations, and conducts each simulation completely before moving to the next. The trend is added to the baseline hazard λ_0 for the stage 1 data. Then, the Exp(1) data is transformed using equation 4.5 to generate the appropriate Weibull data for stage 1, and the interim analysis is conducted. The interim analysis occurs after $n_E/3$ events have been observed, rounding up if $n_E/3$ is odd. The time t_1 at which this occurs is noted, and only individuals with entry time less than t_1 are included in the interim analysis. The rest are included in stage 2 (re-scaling survival times appropriately). The acceleration decision is then made based on the interim analysis z-statistic.

The remaining subjects are randomized in stage 2. If an acceleration threshold is specified, these subjects are assigned to treatment according to a 4:1 ratio if the interim z-statistic is above the threshold, and in a 1:1 ratio otherwise. If an acceleration threshold is not specified, these subjects are randomized in a 1:1 ratio. The data is scaled appropriately according to the same formula used in stage 1. The time of the n_E^{th} event is noted, and only subjects accrued before this point are considered for the final analysis. The only censoring considered is administrative, at the the time of the n_E^{th} analysis.

Analyses are performed on stage 2 data, the stage 1 data with accumulated events, and on pooled data

including all subjects accrued before the n_E^{th} analysis is observed. Summaries of each of these analyses are then generated and returned.

4.4 Simulations

Simulations were performed to mimic the scenarios discussed and simulated in Chapter 2 to maintain as much comparability as possible. Events are assumed to be harmful (e.g. death, disease progression) and treatment is assumed to be beneficial (i.e. HR < 1 is evidence in favor of treatment). An increasing trend in the hazard is considered. In contrast to the continuous data model, subjects accrued towards later periods of the survival model have less favorable outcomes. Thus, true beneficial treatment effects seen early in the study are expected to be attenuated towards the null of no effect when accelerated in the presence of a trend.

Three effect sizes were considered: no effect ($\beta = 0$, HR=1), a small effect ($\beta = -0.2$, HR $= 0.82$), and large effect ($\beta = -0.4$, HR $= 0.67$). Both no trend and an increase in hazard of $1/250$ per subject accrued were simulated. Survival times were generated with three distributions: exponential, Weibull with increasing hazard (k=1.2), and Weibull with decreasing hazard (k=0.8). Both no acceleration and an O'Brien-Fleming level 0.20 boundary (Table 2.8.5) were considered. All simulations were conducted with $50,000$ replications and data were analyzed with a Cox proportional hazards model. All combinations of these parameters were simulated, for a total of 36 simulations.

4.5 Results

The results are described below for each combination of acceleration and trend, with all results aggregated in a single table afterwards (Table 4.5.1).

4.5.1 No acceleration, no trend

With no acceleration or trend, coefficient estimates are unbiased, and both type I error and power are attained (Table 4.5.1, part I). Note that both non-zero effect sizes reach 90% power since simulations were coded to accrue until the number of events yielding 90% power was reached.

4.5.2 No acceleration, with trend

From Table 4.5.1 part II, when a trend is added but no acceleration takes place, there is a slight attenuation in the coefficient estimates toward the null, as expected from an unadjusted proportional hazard regression model that does not adjust for an important prognostic variable. This attenuation is more pronounced for exponential data than for the two Weibull options with non-constant hazard. A slight decrease in power is also observed, but type I error is attained. This result is in contrast to the results observed with the continuous data model when a trend is present, but no adaptive modification. This can be explained by differences between summarizing a continuous outcome with the mean and summarizing survival distributions with the hazard ratio.

When a secular trend is present in either case, calendar time is a precision variable. Failure to account for it in the continuous data model inflates the estimate of the standard error, resulting in a decreased type I error rate, but coefficient estimates remain unbiased. However, failure to account for a precision variable with Cox proportional hazards will yield coefficient estimates attenuated toward the null, but preserve the type I error. In the presence of a true treatment effect, this attenuation of the hazard ratio will lead to decreased power.

4.5.3 Acceleration, no trend

When accelerating in the absence of a trend (Table 4.5.1, part III), there is essentially no bias, loss of power, or inflation of type I error. The only change being made is to the randomization ratios, and there is no relationship between adapting the trial and outcome.

4.5.4 Acceleration, with trend

When accelerating in the presence of a trend, there is a small amount of bias in coefficient estimates under the null hypothesis, and a significant amount of bias under both of the alternatives (Table 4.5.1, part IV). This bias is again more extreme for exponential data than either Weibull distribution simulated. For both alternatives, the bias is towards the null (as expected, as described in §4.4). The bias observed is such that a more deleterious effect of treatment is estimated too often, thus leading to a lower probability of rejecting the null hypothesis under both the null hypothesis and the alternative hypothesis. However, the relative bias is not as great as had been observed with the immediate outcomes in Chapter 2.

Table 4.5.1
Summary of survival simulations

I. No acceleration, without trend									
β	Trend	Acceleration	Data	$\widehat{\beta}$	$Var(\widehat{\beta})$	P(Rejection)	Bias	Relative bias	P(accelerate)
0	0	None	Exponential	-0.000	0.015	0.025	-0.000	-	0.000
0	0	None	Weibull increasing	-0.000	0.015	0.025	-0.000	-	0.000
0	0	None	Weibull decreasing	0.000	0.015	0.025	0.000	-	0.000
-0.2	0	None	Exponential	-0.200	0.004	0.899	-0.000	-0.001	0.000
-0.2	0	None	Weibull increasing	-0.200	0.004	0.898	0.000	0.001	0.000
-0.2	0	None	Weibull decreasing	-0.200	0.004	0.899	0.000	0.001	0.000
-0.4	0	None	Exponential	-0.401	0.016	0.895	-0.001	-0.002	0.000
-0.4	0	None	Weibull increasing	-0.400	0.016	0.896	-0.000	-0.001	0.000
-0.4	0	None	Weibull decreasing	-0.401	0.016	0.897	-0.001	-0.003	0.000
II. No acceleration, with trend									
β	Trend	Acceleration	Data	$\widehat{\beta}$	$Var(\widehat{\beta})$	P(Rejection)	Bias	Relative bias	P(accelerate)
0	1/250	None	Exponential	-0.000	0.015	0.025	-0.000	-	0.000
0	1/250	None	Weibull increasing	0.001	0.015	0.025	0.001	-	0.000
0	1/250	None	Weibull decreasing	0.000	0.015	0.025	0.000	-	0.000
-0.2	1/250	None	Exponential	-0.187	0.004	0.858	0.013	0.064	0.000
-0.2	1/250	None	Weibull increasing	-0.193	0.004	0.876	0.007	0.036	0.000
-0.2	1/250	None	Weibull decreasing	-0.197	0.004	0.888	0.003	0.015	0.000
-0.4	1/250	None	Exponential	-0.389	0.016	0.879	0.011	0.027	0.000
-0.4	1/250	None	Weibull increasing	-0.398	0.016	0.892	0.002	0.006	0.000
-0.4	1/250	None	Weibull decreasing	-0.399	0.016	0.894	0.001	0.003	0.000
III. Acceleration, without trend									
β	Trend	Acceleration	Data	$\widehat{\beta}$	$Var(\widehat{\beta})$	P(Rejection)	Bias	Relative bias	P(accelerate)
0	0	OBF20	Exponential	0.000	0.015	0.026	0.000	-	0.089
0	0	OBF20	Weibull increasing	0.000	0.015	0.025	0.000	-	0.088
0	0	OBF20	Weibull decreasing	-0.000	0.015	0.026	-0.000	-	0.088
-0.2	0	OBF20	Exponential	-0.200	0.004	0.896	0.000	0.002	0.697
-0.2	0	OBF20	Weibull increasing	-0.200	0.004	0.898	0.000	0.001	0.695
-0.2	0	OBF20	Weibull decreasing	-0.200	0.004	0.895	0.000	0.000	0.698
-0.4	0	OBF20	Exponential	-0.400	0.015	0.897	-0.000	-0.001	0.695
-0.4	0	OBF20	Weibull increasing	-0.400	0.015	0.895	0.000	0.001	0.696
0	0	OBF20	Weibull decreasing	-0.400	0.016	0.897	0.000	0.001	0.697
IV. Acceleration, with trend									
β	Trend	Acceleration	Data	$\widehat{\beta}$	$Var(\widehat{\beta})$	P(Rejection)	Bias	Relative bias	P(accelerate)
0	1/250	OBF20	Exponential	0.011	0.013	0.012	0.011	-	0.088
0	1/250	OBF20	Weibull increasing	0.005	0.014	0.016	0.005	-	0.091
0	1/250	OBF20	Weibull decreasing	0.004	0.014	0.018	0.004	-	0.086
-0.2	1/250	OBF20	Exponential	-0.083	0.005	0.288	0.117	0.583	0.683
-0.2	1/250	OBF20	Weibull increasing	-0.116	0.003	0.458	0.084	0.419	0.691
-0.2	1/250	OBF20	Weibull decreasing	-0.122	0.003	0.501	0.078	0.388	0.699
-0.4	1/250	OBF20	Exponential	-0.301	0.012	0.703	0.099	0.247	0.685
-0.4	1/250	OBF20	Weibull increasing	-0.353	0.013	0.837	0.047	0.117	0.695
0	1/250	OBF20	Weibull decreasing	-0.358	0.013	0.843	0.042	0.105	0.692

Chapter 5

Discussion

5.1 Two-arm continuous data

The simulations we conducted with the two arm continuous model demonstrate the existence of inflation of type I error in assessing efficacy and bias in estimating treatment effect when the randomization procedure is adaptively modified in the presence of a secular trend. The underlying issue is unequal randomization of subjects to experimental versus control, with more subjects assigned to the experimental arm at later times in the study. This becomes problematic when the subjects accrued towards the beginning of the study are different that those accrued near the beginning. The bias resulting from accelerating in the presence of a trend was shown to persist over a range of acceleration rules, and increased with more aggressive acceleration rules, tending to be worse as the acceleration probability was closer to 50%

Stratifying on analysis stage removes this bias and inflation of type I error when a secular trend does exists. However, if no trend exists, then a stratified analysis will actually introduce a small amount of bias relative to a fixed sample test and reduced precision. This results from weighting the first stage favorably when accelerating (and spuriously high estimates are observed) compared with non-accelerated trials, in which the estimates are biased slightly downward, but both stages are weighted equally. This held true in our simulations when we weighted the stages both efficiently and by sample size. Of course, when designing a study, it is unknown whether or not a trend in outcome will be present or not. However, the bias and loss of precision resulting from unnecessarily stratifying the analysis is minute relative to the bias that can be

induced if a trend is present.

We also demonstrated the existence of random high bias when the randomization ratio is adaptively modified, even in the absence of a trend or treatment effect. When no trend is present, the adaptive design we present introduces less bias than classical group sequential designs, because additional data is accrued which regresses the spurious estimates towards the mean. In contrast, group sequential designs would allow for stopping with spuriously extreme estimates.

By accruing unevenly between arms over time when a secular trend present, the variance of the outcome is unequal. Our simulations did not show that this had a significant impact on the operating characteristics of the trial, even though a test presuming equal variances might be used.

5.2 Multi-arm continuous data

The performance of the multi-arm design is complicated by resuming treatment arms which are postponed when other arm(s) show promise at the interim analysis. In addition, the effects of random bias in the control arm extend to comparisons involving all experimental arms.

When a linear trend is present, estimates of treatment effect in accelerated arms are biased. As with the two-arm design, this result from unbalanced accrual rates with respect to time, when a relationship between time and outcome exists. The experimental arms postponed are also biased for the same reason, but the magnitude of this bias is less because the randomization ratio is closer to 1:1 than for the accelerated arms.

As with the two-arm design, conducting a stratified analysis removed the bias induced by the linear trend by only comparing arms during periods of constant accrual between the two arms.

5.3 Two arm survival data

The same issues demonstrated in the continuous data models also arise with time-to-event data. Significant bias and loss of power were shown to exist when a secular trend in outcome is present over calendar time and randomization ratio is adaptively modified. However, these data have the additional feature of accumulating events from individuals censored at the interim analysis, which ameliorates many of the issues seen with continuous data. While in principle random high bias could be observed in the time-to-event designs

considered, the simulations we conducted did not demonstrate this, perhaps suggesting that any such bias is not detectable with the $50,000$ simulations we ran. This is likely due to the continued accrual of statistical information from the first stage subjects who were censored at the interim adaptive analysis.

Significant bias and loss of power were shown to exist when a trend is present and randomization is consequently associated with both outcome and time.

5.4 Implications for current practice

Trial designs such as BATTLE and I SPY 2 have proposed the implementation of adaptive modifications of randomization ratios as the accruing data suggest treatment benefit in some or part of the study population. Proponents of such designs have not yet addressed the potential confounding that can occur due to secular calendar time trends. Our research suggests more attention needs to be paid to these issues.

The time trends explored in this research were at times extreme, and perhaps the most extreme trends would not be likely in nonadaptive studies. However, the idea that adaptive enrichment of particular subgroups might be combined with adaptive modifications to the randomization ratio may potentiate such trends: to the extent that changes to eligibility criteria are based on variables that are prognostic of outcome, whether or not they are truly predictive of better treatment effect.

Table 5.4.1
Scenarios and issues discussed
- no issues; + present but not major concern; +++ major concern

Δ in ratios	Trend	Analysis	Random high bias	Confounding	Unequal variances	Loss of precision
No	No	Equal variance	-	-	-	-
		Unequal variance	-	-	-	-
		Stratified (ss)	-	-	-	-
		Stratified (efficient)	-	-	-	-
Always	No	Equal variance	-	-	-	-
		Unequal variance	-	-	-	-
		Stratified (ss)	-	-	-	-
		Stratified (efficient)	-	-	-	-
Adaptive	No	Equal variance	+	-	-	-
		Unequal variance	+	-	-	-
		Stratified (ss)	+	-	-	-
		Stratified (efficient)	+	-	-	-
No	Yes	Equal variance	-	-	-	+++
		Unequal variance	-	-	-	+++
		Stratified (ss)	-	-	-	+
		Stratified (efficient)	-	-	-	+
Always	Yes	Equal variance	-	+++	+++	+++
		Unequal variance	-	+++	-	+++
		Stratified (ss)	-	-	-	+
		Stratified (efficient)	-	–	-	+
Adaptive	Yes	Equal variance	+	+++	+++	+++
		Unequal variance	+	+++	-	+++
		Stratified (ss)	+	-	-	+
		Stratified (efficient)	+	-	-	+

Appendix A

Calculating sufficient statistics

Notation

- $\widetilde{\widehat{\sigma^2}}_{jk}$: cumulative variance of treatment arm k at end of stage j.
- $\widetilde{\widehat{\mu}}_{jk}$: cumulative mean of treatment arm k at end of stage j.
- N_{jk}: cumulative sample size accrued to treatment arm k at end of stage j.
- $\widehat{\sigma^2}_{jk}$: stage-wise variance of observations accrued to treatment arm k during stage j.
- $\widehat{\mu}_{jk}$: stage-wise mean of observations accrued to treatment arm k during stage j.
- n_{jk}: stage-wise sample size accrued to treatment arm k during stage j.
- Y_{ijk}: i^{th} observation accrued to treatment arm k during stage j (i resets at each stage).

Calculation

The quantities of interest are the standard unbiased estimator of the cumulative variance at stage j in treatment arm k. and the cumulative sample mean at the same time. Note that with a single interim analysis, $j = 2$ is primary of interest (the cumulative estimates at $j = 1$ are simply the stage-wise estimates). The estimates of the cumulative mean and the cumulative variance can be written as follows:

$$\widetilde{\widehat{\sigma^2}}_{jk} = \frac{1}{N_{jk} - 1} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} \left(Y_{ilk} - \widetilde{\widehat{\mu}}_{jk} \right)^2 .$$

$$\widetilde{\widehat{\mu}}_{jk} = \frac{1}{N_{jk}} \sum_{l=1}^{j} n_{lk} \widehat{\mu}_{lk}$$

It is more favorable to work with the standard biased estimator of the cumulative variance and and then correct for the bias. Denote the biased estimator by

$$\widetilde{\widehat{\sigma^2}}^{*}_{jk} = \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} \left(Y_{ilk} - \widetilde{\widehat{\mu}}_{jk} \right)^2 .$$

Let $\beta = \frac{N_{jk}-1}{N_{jk}}$ be reciprocal of the bias correction factor, so that $\beta \cdot \widetilde{\widehat{\sigma^2}}_{jk} = \widetilde{\widehat{\sigma^2}}^{*}_{jk}$. Then

First, rewrite the biased estimator of the cumulative variance in terms of the first two sample moments:

$$\begin{aligned}
\beta \cdot \widetilde{\widehat{\sigma^2}}_{jk} &= \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} \left(Y_{ilk} - \widetilde{\widehat{\mu}}_{jk} \right)^2 \\
&= \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} \left(Y_{ilk}^2 - 2\widetilde{\widehat{\mu}}_{jk} \cdot Y_{ijk} + \widetilde{\widehat{\mu}}^2_{jk} \right) \\
&= \left[\frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} Y_{ilk}^2 \right] - \left[\frac{2\widetilde{\widehat{\mu}}_{jk}}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} Y_{ilk} \right] + \left[\widetilde{\widehat{\mu}}^2_{jk} \right] \\
&= \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} Y_{ilk}^2 - 2\widetilde{\widehat{\mu}}^2_{jk} + \widetilde{\widehat{\mu}}^2_{jk} \\
&= \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} Y_{ilk}^2 - \widetilde{\widehat{\mu}}^2_{jk}.
\end{aligned}$$

Now, write the cumulative mean in terms of the stage-wise means and sample sizes, substituted, and expand the resulting quadratic (making use of the fact that $j = 2$).

$$\begin{aligned}
\beta \cdot \widetilde{\widehat{\sigma^2}}_{jk} &= \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} Y_{ilk}^2 - \left(\frac{1}{N_{jk}} \sum_{l=1}^{j} n_{lk} \widehat{\mu}_{lk} \right)^2 \\
&\underbrace{=}_{j=2} \frac{1}{N_{jk}} \sum_{l=1}^{j} \sum_{i=1}^{n_{lk}} Y_{ilk}^2 - \frac{1}{N_{jk}^2} \left(n_{1k} \widehat{\mu}_{1k} + n_{2k} \widehat{\mu}_{2k} \right)^2 \\
&= \frac{1}{N_{jk}} \left(\sum_{i=1}^{n_{1k}} Y_{i1k}^2 + \sum_{i=1}^{n_{2k}} Y_{i2k}^2 \right) - \frac{1}{N_{jk}^2} \left(n_{1k}^2 \widehat{\mu}_{1k}^2 + n_{2k}^2 \widehat{\mu}_{2k}^2 + 2 n_{1k} n_{2k} \widehat{\mu}_{1k} \widehat{\mu}_{2k} \right).
\end{aligned}$$

This expression contains the sum of squared observations Y_{ilk}^2 and the estimates of the stage-wise means $\widehat{\mu}_{lk}$, for $l = 1, 2$. Next, with a bit of straightforward algebraic gymnastics (not shown), rearrange and factor this expression so it is written in terms of the biased stage-wise variance estimates.

$$\begin{aligned}
\beta \cdot \widetilde{\widehat{\sigma^2}}_{jk} &= \left(\frac{1}{N_{jk}} \sum_{i=1}^{n_{1k}} Y_{i1k}^2 - \frac{n_{1k}^2 \widehat{\mu}_{1k}^2}{N_{jk}^2} \right) + \left(\frac{1}{N_{jk}} \sum_{i=1}^{n_{2k}} Y_{i2k}^2 - \frac{n_{2k}^2 \widehat{\mu}_{2k}^2}{N_{jk}^2} \right) - \left(\frac{2 n_{1k} n_{2k} \widehat{\mu}_{1k} \widehat{\mu}_{2k}}{N_{jk}^2} \right) \\
&= \sum_{l=1}^{2} \left[\frac{n_{lk}}{N_{jk}} \cdot \left(\frac{1}{n_{lk}} \sum_{i=1}^{n_{lk}} Y_{ilk}^2 - \hat{\mu}_{lk}^2 \right) + \frac{n_{lk} \widehat{\mu}_{lk}^2}{N_{jk}} \left(1 - \frac{n_{lk}^2}{N_{jk} n_{lk}} \right) \right] - \frac{2 n_{1k} n_{2k} \widehat{\mu}_{1k} \widehat{\mu}_{2k}}{N_{jk}^2} \\
&= \left[\sum_{l=1}^{2} \frac{n_{lk}}{N_{jk}} \widehat{\sigma^2}^*_{lk} + \frac{n_{lk} \widehat{\mu}_{lk}^2}{N_{jk}} \left(1 - \frac{n_{lk}^2}{N_{jk} n_{lk}} \right) \right] - \frac{2 n_{1k} n_{2k} \widehat{\mu}_{1k} \widehat{\mu}_{2k}}{N_{jk}^2}
\end{aligned}$$

This is easily expressed in terms of the unbiased estimate for the stage-wise variance, using the relationship

$$\widehat{\sigma^2}^*_{lk} = \beta \widehat{\sigma^2}_{lk} = \frac{n_{jk} - 1}{n_{jk}} \widehat{\sigma^2}_{lk}$$

Hence,

$$\beta \cdot \widetilde{\widehat{\sigma^2}}_{jk} = \left[\sum_{l=1}^{2} \frac{n_{lk}(n_{jk} - 1)}{n_{jk} N_{jk}} \widehat{\sigma^2}_{lk} + \frac{n_{lk} \widehat{\mu}_{lk}^2}{N_{jk}} \left(1 - \frac{n_{lk}^2}{N_{jk} n_{lk}} \right) \right] - \frac{2 n_{1k} n_{2k} \widehat{\mu}_{1k} \widehat{\mu}_{2k}}{N_{jk}^2}$$

This last expression is written entirely in terms of stage-wise sufficient statistics and stage-wise sample sizes. Multiplying the above quantity by $\frac{1}{\beta}$ yields the desired estimator of the cumulative variance:

$$\widetilde{\widehat{\sigma^2}}_{jk} = \frac{1}{\beta} \left[\sum_{l=1}^{2} \frac{n_{lk}(n_{jk} - 1)}{n_{jk} N_{jk}} \widehat{\sigma^2}_{lk} + \frac{n_{lk} \widehat{\mu}_{lk}^2}{N_{jk}} \left(1 - \frac{n_{lk}^2}{N_{jk} n_{lk}} \right) \right] - \frac{2 n_{1k} n_{2k} \widehat{\mu}_{1k} \widehat{\mu}_{2k}}{N_{jk}^2}$$

Appendix B

Variance adjustment

Let Y_{ijk} denote the i^{th} subject on treatment arm k, accrued during stage j. It is enough to consider a single stage and treatment arm, so the indices j and k are dropped for simplicity. The model for the data is

$$Y_i = \mu_0 + \beta i + \epsilon_i,$$

where μ_0 represents the mean observation at the beginning of the stage, β is the increase in observation per subject accrued (i.e. the mean time trend), and the $\epsilon_i \sim N(0, \sigma^2)$ are i.i.d. errors. When $\beta \neq 0$, there is an inflation in the variance of the observations. In other words, $Var(\mathbf{Y}) > \sigma^2$. This variance can be expressed in terms of σ^2, β, and n. Assume that the accrual times of the n subjects are uniformly distributed.

By the law of total variance,

$$\begin{aligned} Var\left(Y\right) &= E\left[Var\left(Y|i\right)\right] + Var\left(E\left[Y|i\right]\right) \\ &= \sigma^2 + Var\left(\mu_0 + \beta i\right) \\ &= \sigma^2 + \beta^2 Var(i) \\ &= \sigma^2 + \beta^2 \left(\frac{n^2 - 1}{12}\right) \end{aligned}$$

Where $Var(i)$ is the variance of the discrete uniform distribution on the set $\{1, 2, \ldots, n\}$.

Appendix C

Code

Below are descriptions and results from validation of code used to perform simulations which were presented in this thesis.

C.1 2-arm code

The `R` function `rSeqMean2arm` was written to simulate the 2-arm trial design described above. This code assumes an expected control response of $\mu_0 = 0$, and accepts arbitrary specification of baseline experimental mean μ_1, trend β, and acceleration threshold a. This code simulates outcomes for the first stage of each trial from a $N(0, 1)$ distribution using `rnorm`, which are then scaled to have the appropriate mean and trend. An interim test statistic comparing the single experimental arm to control is computed using Equation 2.1, and is compared to the provided acceleration boundary a. The sample size is hard-coded: 100 are accrued to both experimental and control during stage 1, 100 to experimental during stage 2, and either 25 or 100 to control, depending on whether or not the trial accelerates. Stage 2 data is generated separately for accelerated and non-accelerated trials, and pooled for inference at the final analysis. There is no option for conducting stratified analyses with `rSeqMean2arm`, but the necessary stage-wise and cumulative sufficient statistics and simulation parameters are returned in the function output and were saved and used to conduct stratified analyses.

C.2 Two-arm analytic calculation with forced acceleration

The same calculation shown in 2.2 can be done for a single experimental arm of a multi-arm trial, making a simplification that the experimental arm is always the only arm to accelerate. This proved useful for validation purposes, and an option in `rSeqMean5` was included to match this scenario, i.e. always force arm 1 to accelerate while postponing the rest. Again, the type I error for the normal model is derived as a function of the trend β, which is parameterized as the increase in mean response per additional subject accrued, and the number of subjects n accrued to the control group during stage 2. It is assumed that 100 are accrued to each of the 4 experimental arms and the single control during stage 1. The only acceleration rule used is that arm 1 is advanced to stage 2 regardless of outcome, while arms 2, 3, and 4 are always postponed (this is the same randomization scheme shown in Tables 2.2.1 and 2.3.2).

As in §2.1, Let Y_{ijk} indicate the i^{th} subject accrued to treatment arm k during stage j. The data-generating model for the experimental arm is

$$Y_{ij1} \sim \begin{cases} \left(\mu_1 + 5\beta i, \sigma^2\right) & \text{for } i = 1, \ldots, 100;\ j = 1 \\ \left(\mu_1 + 500\beta + \frac{\beta(100+n)}{100} i, \sigma^2\right) & \text{for } i = 1, \ldots, 100;\ j = 2 \end{cases}$$

The 5β in the stage 1 distribution reflects the interpretation of β as the increase in mean of per individual subject accrued. Since during the first stage, 500 are accrued, the average increase in mean per subject accrued to either experimental or treatment is 5β. Likewise, the $\frac{\beta(100+n)}{100}$ reflects the arm-specific trend: $\beta \cdot (100 + n)$ is the total increase in mean over the time period, and spread over the 100 subjects accrued to experimental during stage 2 gives a arm/stage specific trend of $\frac{\beta(100+n)}{100}$. Similarly, the data generating model for the control arm is

$$Y_{ij0} \sim \begin{cases} \left(\mu_0 + 5\beta i, \sigma^2\right) & \text{for } i = 1, \ldots, 100;\ j = 1 \\ \left(\mu_0 + 500\beta + \frac{\beta(100+n)}{n} i, \sigma^2\right) & \text{for } i = 1, \ldots, n;\ j = 2 \end{cases}$$

Thus, the cumulative means for the experimental and control arm are, respectively,

$$\begin{aligned}
\overline{Y}_1 &= \frac{1}{200}\left[\sum_{i=1}^{100} Y_{i11} + \sum_{i=1}^{100} Y_{i21}\right] \\
&\sim \frac{1}{200}\left(200\mu_1 + \frac{5\beta(100)(101)}{2} + 500(100)\beta + \frac{\beta(100+n)}{100}\frac{(100)(101)}{2}, 200\sigma^2\right) \\
&\sim \left(\mu_1 + \beta\left[\frac{5(100)(101)}{2(200)} + \frac{500(100)}{200} + \frac{(100+n)(101)}{400}\right], \frac{\sigma^2}{200}\right) \\
&\sim \left(\mu 1 + \beta\left[\frac{5(101)}{4} + 250 + \frac{(100+n)(101)}{400}\right], \frac{\sigma^2}{200}\right) \\
&\sim \left(\eta_1, \tau_1^2\right)
\end{aligned}$$

$$\begin{aligned}
\overline{Y}_0 &= \frac{1}{100+n}\left[\sum_{k=1}^{100} Y_{i10} + \sum_{l=1}^{n} Y_{i20}\right] \\
&\sim \frac{1}{100+n}\left((100+n)\mu_0 + \frac{5\beta(100)(101)}{2} + 500n\beta + \frac{\beta(100+n)}{n}\cdot\frac{n(n+1)}{2}, (100+n)\sigma^2\right) \\
&\sim \left(\mu_0 + \beta\left[\frac{5(100)(101)}{2(100+n)} + \frac{500n}{100+n} + \frac{n+1}{2}\right], \frac{\sigma^2}{100+n}\right) \\
&\sim \left(\eta_0, \tau_0^2\right)
\end{aligned}$$

A naïve approach that ignores the trend in the data would assume that the following test statistic Z has an asymptotic standard normal distribution under the null, when in fact it has the following normal distribution:

$$Z = Z(\beta, n) = \frac{\overline{Y}_1 - \overline{Y}_0}{\sqrt{\frac{\sigma_{1*}^2}{200} + \frac{\sigma_{0*}^2}{100+n}}} \rightarrow N\left(\eta_1 - \eta_0, \frac{\tau_1^2 + \tau_0^2}{\left(\frac{\sigma_{1*}^2}{200} + \frac{\sigma_{0*}^2}{100+n}\right)}\right),$$

where σ_{1*}^2 and σ_{0*}^2 are the cumulative variances in presence of trend for experimental and control, respectively. The cumulative variance for a given arm k and stage j is $\sigma_{*jk}^2 = \sigma^2 + \eta\left(\frac{n_{jk}-1}{12}\right)$ [Appendix B]. The cumulative variance for a single arm throughout both stages can then be computed from the stage-wise means and variances for that arm. Note that the parameters σ_{1*}^2 and σ_{0*}^2 depend on the trend β and the control stage-2 sample size n. The probability of obtaining a test statistic higher than the critical value $Z_{1-\alpha/2}$ can then be computed for various choices of β and n. Note that while β represents the trend present, n controls the degree of acceleration. When $n = 100$, no acceleration occurs (as in Section 2.5). Values of $n < 100$ correspond to accelerating the treatment arm, and values of $n > 100$ correspond to decelerating the

experimental arm (not considered in this thesis.

C.3 Validation of two-arm code with forced acceleration

The function `rSeqMean5` was validated by comparison to both analytic results and to more straightforward, but much slower, implementations of the function in `R`.

In addition to the implementation used for simulations, three functions were used for validation: one implements the analytic results derived in §2. The other two generate outcomes rather than relying on sufficient statistics. Of these, one generates data for all arms, and loops over stages and simulations. The other only generates data for a single arm that is accelerated and control, but computes the time trend to match that of a 5-arm trial in stage 1.

The basic `dsn` object used for the simulations was generated using `seqDesign`, testing a null hypothesis of equal means versus an greater alternative with two interim analyses. A standard deviation of 1 is used in all arms, 1:1 initial randomization ratio (treatment arm to control), and a planned sample size of 100 subjects per stage. The size of the tests was set at $\alpha = 0.025$.

A variety of linear time trends in the mean are considered, ranging from no trend to an increase in mean of 1 per 125 subjects accrued to the trial. This simulates a secular trend, rather than any time $\times$ treatment effect interaction. All simulations were conducted with all arms under the null and with a strong effect in arm 1. All simulations were repeated 20000 times (30000 and above resulted in memory issues with `rSM5v3`, which generates data for all arms and simulations).

In the arm which was accelerated, the simulated type I error matches the analytic simulations.

C.4 Multi-arm code

Simulations were performed using `R` code modified from the `RCTdesign` package [7, 21]. The function `rSeqMean2` was modified in two key ways. First,`rSeqMean5` simulates 5-arm clinical trials (4 experimental arms and 1 control), rather than 2-arm trials. Second, `rSeqMean5` only allows a single interim analysis, at which experimental arms may be postponed or accelerated depending on their performance in the first arm (`rSeqMean2` does not allow for any adaptive modification, other than group sequential stopping rules).

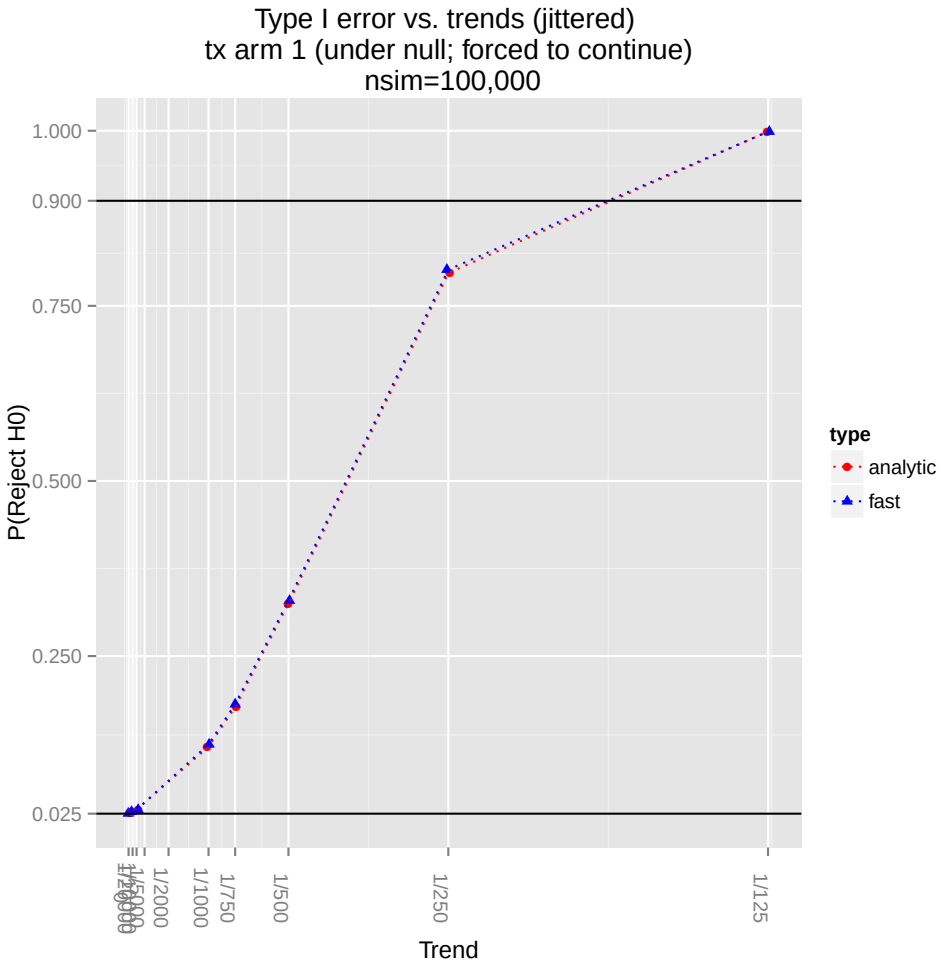

Figure C.3.1
Type 1 error vs. trend for both analytic and simulation

To improve efficiency of the code, rather than generating individual subject-level data, stage-wise sufficient statistics $(\widehat{\mu}_{jk}, \widehat{\sigma}_{jk})$, were generated for each simulated trial. Applying the Central Limit Theorem, the stage-wise sample averages were simulated from a $N(0,1)$ distribution using the `rnorm` function in `R`, and scaled to the appropriate means and variances: $\overline{Y}_{jk} \sim \frac{\sigma_{jk}}{\sqrt{n_{jk}}} N(0,1) + \mu_{jk}$. While the assumption of a normally distributed outcome is not needed to simulate sample means, it is necessary in order to rely on the asymptotic distribution of the sampling variance. The code is partly vectorized, and generates the data for each stage using matrix operations and no loops. However, since the sampling distribution of all but the first stage depends on previous stages, looping over the stages was unavoidable. There is also a loop over the simulations to determine which arms are accelerated and which are postponed.

At the interim analysis, the statistic T_{1k} is compared to the efficacy boundary of a group sequential design. As described in Section 2.8, a variety of designs were used to generate the acceleration boundaries including both O'Brien-Fleming and Pocock at several sizes. The `seqDesign` function from the `RCTdesign` package was used to generate these acceleration boundaries.

At the final analysis, the cumulative estimates of the means and variances were calculated from the stage-wise estimates and sample sizes [Appendix B]. For each experimental arm, only control data that was accrued at the same time was used. For example, control data accrued during a stage in which an arm is accelerated is not used as a comparison for the arms which were postponed.

The `R` code `rSeqMean5` was validated by comparison to analytic results, using an acceleration option built into `rSeqMean5` that forces arm 1 to continue to stage 2 while postponing the rest, regardless of observed outcomes. At the first stage, 100 are accrued to each arm. During the second stage, 100 are accrued to arm 1 and 25 to control to maintain a 20% accrual rate to control. These simulations match the analytic results over the range of trends considered (increase in mean per subject of 1/20000 to 1/125) [Appendix C].

Appendix D

2-arm simulations

The plots below show the full results of the 2-arm simulations. In Figure D.0.1, the horizontal line has intercept 0.032, which corresponds to bias of 10% of the effect size giving 90% power ($\mu_1 = 0.32$).

Figure D.0.2 shows the rejection probability as a function of trend, for each of the four efficacy scenarios simulated ($\mu_1 \in \{0, 0.08, 0.16, 0.32\}$). A subset of this plot is also shown in figure 2.9.5, for $\mu_1 = 0, 0.32$, with no acceleration, OBF 0.20, and Pocock level 0.30 acceleration boundaries.

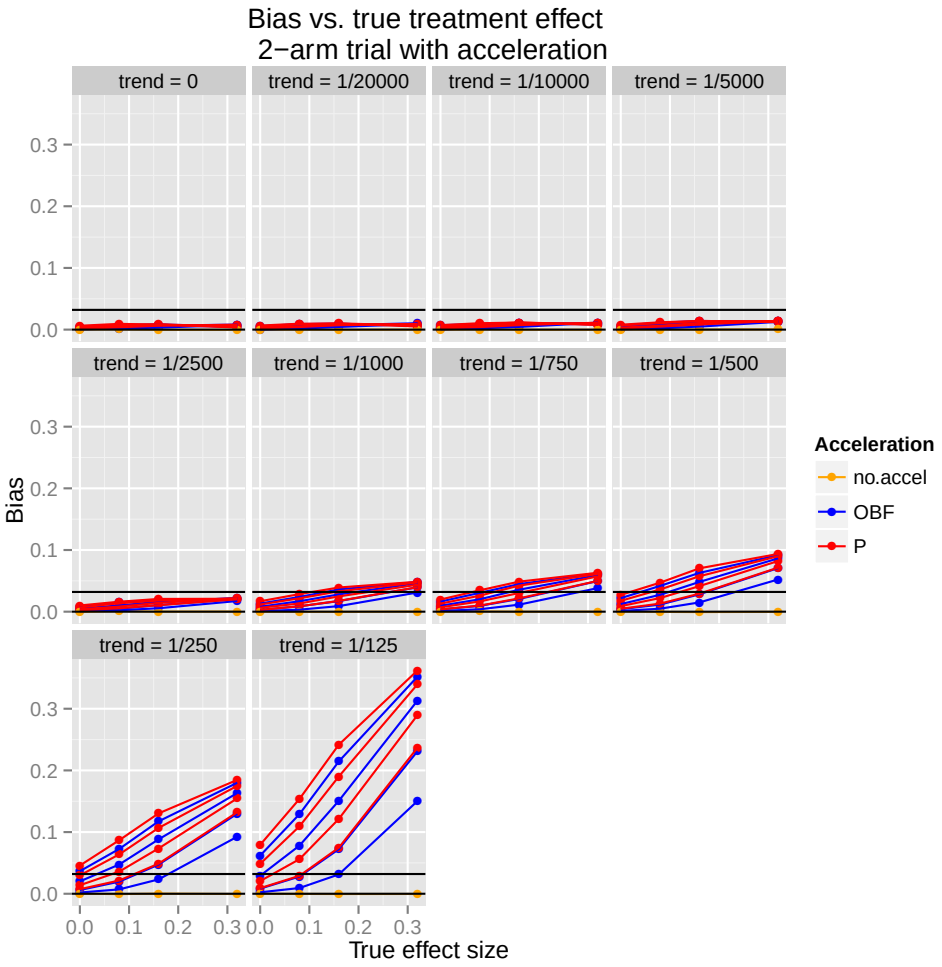

Figure D.0.1
Bias as function of effect size

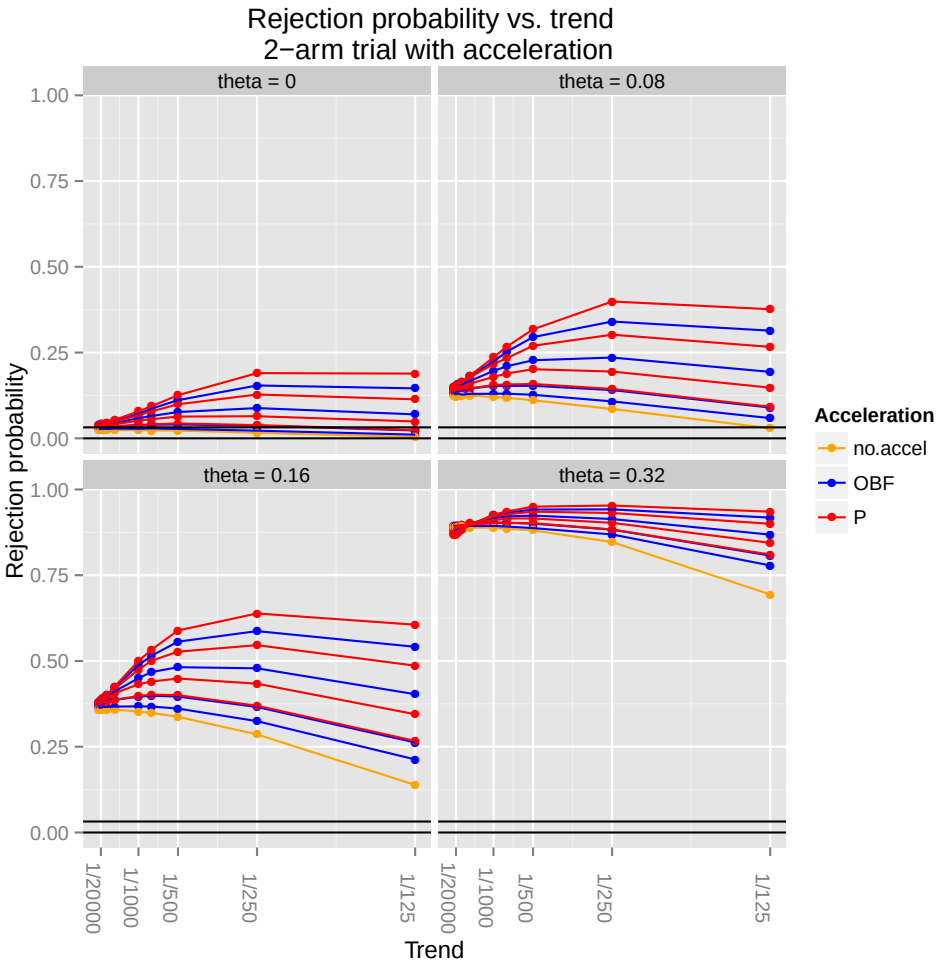

Figure D.0.2
Bias as function of trend

Appendix E

Survival calculations

E.1 Sample size

Below are the full details of the sample size calculation used for two-arm survival simulations. This calculation is specifically for the probability p than an individual experiences an event during the first half of the study when accrual is Unif(0,1) and survival times are Exp(λ). This probability can then be used to compute the number of individuals to accrue so that halfway through the study, the expected number of total events is equal to $\frac{1}{3}n_E$, where n_E is the planned number of events (across both arms). Note that this calculation parameterizes the exponential distribution by the rate, while the simulation code parameterizes by the mean. The resulting equation can be transformed to the mean parameterization by substituting λ by $1/\lambda$.

$$
\begin{aligned}
p &= \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-s} \lambda e^{-\lambda s} dt ds \\
&= \lambda \int_0^{\frac{1}{2}} e^{-\lambda s} \int_0^{\frac{1}{2}-s} 1 dt ds = \lambda \int_0^{\frac{1}{2}} e^{-\lambda s} \left[\frac{1}{2} - s\right] ds \\
&= \lambda \int_0^{\frac{1}{2}} \frac{1}{2} e^{-\lambda s} - \lambda s e^{-\lambda s} ds \\
&= \underbrace{\lambda \int_0^{\frac{1}{2}} \frac{1}{2} e^{-\lambda s} ds}_{A} \underbrace{- \lambda \int_0^{\frac{1}{2}} s e^{-\lambda s} ds}_{B}
\end{aligned}
$$

Then A and B can be computed as follows, using integration by parts for B with $u = s; du = ds$ and $v = -\frac{1}{\lambda}e^{-\lambda s}; dv = e^{-\lambda s}ds$.

$$\begin{aligned}
A &= \lambda \int_0^{\frac{1}{2}} \frac{1}{2} e{-}\lambda s ds = \frac{\lambda}{2}\left(-\frac{1}{\lambda}e^{-\lambda s}\Big|_0^{\frac{1}{2}}\right) \\
&= -\frac{1}{2}\left(e^{-\frac{\lambda}{2}} - 1\right) \\
&. \\
B &= uv - \int v du = -\frac{s}{\lambda}e^{-\lambda s}\Big|_0^{\frac{1}{2}} - \int_0^{\frac{1}{2}} -\frac{1}{\lambda}e^{-\lambda s}ds \\
&= \left[-\frac{1}{2\lambda}e^{-\frac{\lambda}{2}} + 0\right] + \frac{1}{\lambda}\int_0^{\frac{1}{2}} e^{-\lambda s}ds \\
&= \left[-\frac{1}{2\lambda}e^{-\frac{\lambda}{2}}\right] - \frac{1}{\lambda^2}\left[e^{-\lambda s}\Big|_0^{\frac{1}{2}}\right] \\
&= -\frac{1}{2\lambda}e^{-\frac{\lambda}{2}} - \frac{1}{\lambda^2}\left[e^{-\frac{\lambda}{2}} - 1\right] \\
&= e^{-\frac{\lambda}{2}}\left(-\frac{1}{2\lambda} - \frac{1}{\lambda^2}\right) + \frac{1}{\lambda^2}
\end{aligned}$$

Thus,

$$\begin{aligned}
p = A - \lambda B &= -\frac{1}{2}\left(e^{-\frac{\lambda}{2}} - 1\right) - \lambda\left[e^{-\frac{\lambda}{2}}\left(-\frac{1}{2\lambda} - \frac{1}{\lambda^2}\right) + \frac{1}{\lambda^2}\right] \\
&= -\frac{1}{2}\left(e^{-\frac{\lambda}{2}} - 1\right) + e^{-\frac{\lambda}{2}}\left(\frac{1}{2} + \frac{1}{\lambda}\right) - \frac{1}{\lambda} \\
&= e^{-\frac{\lambda}{2}}\left(-\frac{1}{2} + \frac{1}{2} + \frac{1}{\lambda}\right) + \frac{1}{2} - \frac{1}{\lambda} \\
&= \frac{e^{-\frac{\lambda}{2}} - 1}{\lambda} + \frac{1}{2} = \frac{2\left(e^{-\frac{\lambda}{2}-1}\right) + \lambda}{2\lambda}.
\end{aligned}$$

Hence recruiting

$$n = \frac{2n_E}{3}\left[\frac{2\left(e^{-\frac{\lambda_1}{2}-1}\right) + \lambda_1}{2\lambda_1} + \frac{2\left(e^{-\frac{\lambda_0}{2}-1}\right) + \lambda_0}{2\lambda_0}\right]^{-1} = \frac{2n_E}{3}\left[\frac{e^{-\frac{\lambda_1}{2}-1}}{\lambda_1} + \frac{e^{-\frac{\lambda_0}{2}-1}}{\lambda_0} + 1\right]^{-1}$$

individuals uniformly over the study period will result in an expected number of total events equal to $\frac{1}{3}n_E$ halfway through the study, where n_E is the planned number of events (across both arms). Again, note that the simulation code parameterizes the exponential distribution by the mean rather tha the rate, in which case the equation above is valid when λ_k is substituted by $1/\lambda_k$.

Bibliography

[1] P. Armitage, C.K. McPherson, and B.C. Rowe. Repeated signficance tests on accumulating data. *Joural of the Royal Statistical Society*, 1969.

[2] AD Barker. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Nature Clinical Pharmacology & Therapeutics*, 2009.

[3] P. Bauer and K. Kohne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, 1994.

[4] Thomas Cook and David L. DeMets. Review of draft FDA adaptive design guidance. *Journal of Biopharmaceutical Statistics*, 2010.

[5] Scott S. Emerson. *Boundary Scales in RCTdesign*, August 2012.

[6] Scott S. Emerson and Thomas R. Fleming. *Adaptive Methods: Telling 'The Rest of the Story'*. Department of Biostatistics, University of Washington, May 2010.

[7] Scott S. Emerson, Daniel L. Gillen, John K. Kittelson, Sarah C. Emerson, and Gregory P. Levin. *RCTdesign: Group Sequential Trial Design*, 2012. R package version 1.0.

[8] Laura J. Esserman and Donald A. Berry et al. Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: Results from the I-SPY 1 TRIAL. *Journal of Clinical Oncology*, 2012.

[9] Robert H. Bartlett et al. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics*, 1985.

[10] Lloyd D. Fisher. Self-designing clinical trials. *Statistics in Medicine*, 1998.

[11] Thomas R. Fleming. Clinical trials: Discerning hype from substance. *Annals of Internal Medicine*, 2010.

[12] UK Collaborative ECMO Trial Group. UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation. *The Lancet*, 1996.

[13] Michael Hay and David W Thomas et al. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 2014.

[14] Sebastian Irle and Helmut Schäfer. Interim design modifications in time-to-event studies. *Journal of the American Statistical Association*, 2012.

[15] Primo N. Lara Jr. and Mary W. Redman et al. Disease control rate at 8 weeks predicts clinical benefit in non-small cell lung cancer: Results from southwest oncology group randomized clinical trials. *Journal of Clinical Oncology*, 2008.

[16] Edward S. Kim, Roy S. Herbst, and Ignacio I. Wistuba et al. The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery*, 2011.

[17] Edward L. Korn and Boris Friedlin. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 2010.

[18] P. Pearl O'Rourke and Robert K. Crane et al. Extracorporeal membrane oxygenation and convential medical therapy in neonates with persistant pulmonary hypertension of the newborn: A prospective randomized study. *Pediatrics*, 1989.

[19] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 1977.

[20] Michael A. Proschan and Sally Hunsberger. Designed extension of studies based on conditional power. *Biometrics*, 1995.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[22] Lecia V. Sequist, Alona Muzikansky, and Jeffrey A. Engelman. A new BATTLE in the evolving war on cancer. *Cancer Discovery*, 2011.

[23] US Food and Drug Administration. *Guidance for Industry, Adaptive Design Clinical Trials for Drugs and Biologics*, February 2010.

[24] James H. Ware. Investigating therapies of potentially great benefit: ECMO. *Statistical Science*, 1989.

[25] L. J. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 1978.

[26] John Whitehead. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 1986.

[27] M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 1969.

[28] Xian Zhou, Suyu Liu, Edward S. Kim, Roy S. Herbst, and J Jack Lee. Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. *Clinical Trials*, 2008.