# Accounting for the Presence of Surrogate Data in Adaptive Clinical Trials

Cesar Torres

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

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Scott Emerson, Chair

Thomas Fleming

Lurdes Inoue

Program Authorized to Offer Degree:

Biostatistics

University of Washington

**Abstract**

Accounting for the Presence of Surrogate Data in Adaptive Clinical Trials

Cesar Torres

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

Some adaptive designs for randomized clinical trials (RCTs) allow for flexibility in modifying the sequential sampling plan using results from unblinded interim analyses. However, care must be taken to ensure that desired statistical operating characteristics are preserved when allowing for such adaptations. Approaches to analyzing results from RCTs with adaptations have been proposed in the literature, such as methods by Bauer & Köhne (1994), Proschan & Hunsberger (1995), Fisher (1998), and Cui, Hung, & Wang (1999), but these generally assume the use of adjustments that incorporate sufficient statistics derived from the data available at the time of adaptation. Bauer & Posch (2004) noted that this assumption may be violated in settings involving time-to-event data, where observed surrogate outcomes that are potentially informative about future events are not reflected in the adjustment for adaptation. For instance, in an analysis of overall survival, the adaptive analysis typically does not account for the way disease progression in censored subjects might be informative for the eventual death times of those subjects. The impact of falsely assuming unavailability of partial knowledge regarding future data needs to be explored.

Via simulation, we found that under an extreme scenario the type I error might be inflated from 0.025 to 0.205 when surrogate data can be used to predict future event times. Factors contributing to this inflation include minimal spending of error before adapting, early adaptations, minimal restrictions on sample size modifications, and stopping patient recruitment after the time of adaptation. Interestingly, in some situations where

surrogate outcomes are informative but not accounted for, use of the standard methods for adaptation (e.g., Cui, Hung, & Wang) can behave worse than ignoring adaptation completely. Generally, we find that adjusting for surrogate data in the analyses does not control type I error unless strong or conservative assumptions are made. It is not immediately clear that adaptive designs which correctly account for surrogate data hold any significant advantages to comparable group sequential designs, and thus we explore situations where the methods to adjust for this surrogacy might decrease the efficiency so much as to suggest that unblinded adaptations be avoided.

# Contents

# Acknowledgements

I would like to thank Professor Scott Emerson for his guidance and patience throughout the process of writing this dissertation. During my time as Scott's advisee, the feedback and thoughts he has shared with me have largely shaped the way I think of statistics today, and have been instrumental in ensuring that my research was relevant, thorough, and correct. I am also grateful to Professors Thomas Fleming, Lurdes Inoue, Susanne May, and Johanna Lampe for taking time out of their busy schedules to participate on my Supervisory Committee, and for providing valuable feedback regarding my research. I appreciate the opportunity and funding the University of Washington Department of Biostatistics provided for me to complete my graduate studies. I am grateful to my friends for their help during my journey of progressing through the different stages of the Ph.D. program. Finally, I thank my girlfriend Fan, my parents Ignacio and Rosa, and my brother Diego for supporting me in everything I do.

# Chapter 1

# Introduction

Randomized clinical trials (RCTs) are the gold standard for investigating the causal relationship between candidate treatments and clinically meaningful outcomes, because statistical results from observational studies generally cannot be used to make causal statements without exceptionally large treatment effects or strong assumptions about lack of (unmeasured) confounding. However, conducting a RCT is an expensive task, with one estimate putting the cost of an industry-sponsored RCT between $2,098 and $19,285 per enrolled patient, before accounting for overhead expenses[1]. With these financial costs along with ethical concerns of subjecting patients to potentially harmful treatments, it is important that RCTs are carried out correctly.

## 1.1 Fixed Sample Designs

The simplest type of RCT uses a Fixed Sample Design, where a sample of predetermined size $n$ is collected during the course of the RCT, and the primary analysis is carried out after all $n$ subjects have been collected.

For instance, consider the setting of immediately observed outcomes, where it is of interest to compare two treatments: the experimental treatment and the control treatment. Treatment indicator $X_i = 1$ ($X_i = 0$) means that the $i^{th}$ patient is assigned

to the experimental treatment (control treatment). We can define group sample sizes $n_1 \coloneqq \sum_{i=1}^{n} X_i$ and $n_0 \coloneqq n - n_1$, and denote independent and identically distributed outcomes as $Y_1, Y_2, Y_3, \ldots, Y_n$, with group means and variances denoted as $\mathrm{E}(Y_i|X_i = x) = \mu_x$ and $\mathrm{Var}(Y_i|X_i = x) = \sigma^2$, where for the purposes of testing, $\sigma^2$ is assumed or estimated using a consistent estimator. We presume the RCT aims to investigate the parameter $\theta \coloneqq \mu_1 - \mu_0$, a difference in means between the two treatment groups.

A distribution-free estimator of $\theta$ is the difference of the sample means of the treatment groups, $\hat{\theta} \coloneqq \hat{\mu}_1 - \hat{\mu}_0$, with $\hat{\mu}_1 \coloneqq \frac{1}{n_1} \sum_{i=1}^{n} Y_i X_i$ and $\hat{\mu}_0 \coloneqq \frac{1}{n_0} \sum_{i=1}^{n} Y_i (1 - X_i)$. If patients are randomized according to a 1:1 randomization scheme, the sample sizes for the experimental treatment and control treatment groups are $n_1 = n_0 = \frac{n}{2}$. We presume a model in which the Central Limit Theorem provides a good approximation such that $\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta, \frac{V}{n}\right)$, with $V = 4\sigma^2$. The approximate statistical information is $\mathcal{I} = \frac{n}{V}$. A test of the null hypothesis $H_0 : \theta = \theta_0$ can be performed using statistic $Z \coloneqq \frac{\sqrt{n}\left(\hat{\theta} - \theta\right)}{\sqrt{V}} \dot{\sim} \mathcal{N}(\delta, 1)$, with $\delta = \frac{\sqrt{n}(\theta - \theta_0)}{\sqrt{V}}$. We might design a RCT such that a one-sided level $\alpha$ test of $H_0$ would detect $H_a : \theta \geq \theta_a$ with power $\beta$. In that case we choose the sample size $n = \frac{\left(z_{1-\alpha} + z_\beta\right)^2 V}{\left(\theta_a - \theta_0\right)^2}$, where $z_\alpha = \Phi^{-1}(\alpha)$ is the $\alpha$-th quantile of the standard normal distribution.

## 1.2 Group Sequential Designs

As mentioned previously, RCTs have a very high per-patient financial cost. It is therefore of interest to investigators to design efficient RCTs, where efficiency might reflect the number of subjects accrued and the calendar time required. Investigators also prefer efficient RCT designs for ethical reasons: Minimizing the average number of patients in a RCT can reduce the number of patients exposed to potentially harmful experimental treatments. Group ethics are also a concern: Efficient designs maximize the number of patients who will benefit from adoption of beneficial treatments. While a fixed sample design is straightforward, there exist designs that are more efficient, on average.

Notationally, a Group Sequential Design (GSD) has $J$ planned analyses, where $J \geq 2$,

and the cumulative sample sizes for these analyses are $N_1, N_2, \ldots, N_J$. Such a RCT may stop early at one of the analyses according to a prespecified "stopping rule". The GSD can be implemented as a part of the safety monitoring of an RCT, wherein investigators or members of an independent Data and Safety Monitoring Board (DSMB) periodically examine accruing data for excess adverse events or unsuitable risk/benefit tradeoffs.

Analyzing data from a RCT utilizing a GSD is not as straightforward as when using a fixed sample design. Naively applying fixed sample design analysis methods to GSD data can lead to inflated type I error rates. For example, consider the scenario where there are three analyses that are equally spaced with respect to the number of patients accrued into the RCT. If a two-sided hypothesis test is carried out at each analysis by comparing the absolute value of the standardized cumulative test statistic to $z_{0.05} \approx 1.64$, the probability of rejecting the null hypothesis in this RCT when the null hypothesis is true is more than doubled, inflating from 0.05 to 0.10726[2].

### 1.2.1 Stopping Rules and Continuation Regions

Stopping rules effectively partition the sample space for some test statistic $T_j$ computed at the $j^{th}$ analysis. We define continuation sets $\mathscr{C}_j$ and stopping sets $\mathscr{S}_j$ at each of the $j = 1, 2, \ldots, J$ analyses. The stopping rule for the $j^{th}$ analysis would then be the following: If $T_j \notin \mathscr{C}_j$ stop the RCT, and continue on to the next planned analysis otherwise. To ensure that the RCT stops no later than the $J^{th}$ analysis, $\mathscr{C}_J$ is typically specified so that $P(T_J \in \mathscr{C}_J) = 0$. At each analysis $j$, $\mathscr{S}_j$ is defined to be the complement of $\mathscr{C}_j$ so that $\mathscr{C}_j \cap \mathscr{S}_j = \varnothing$ and $\mathscr{C}_j \cup \mathscr{S}_j = (-\infty, \infty)$.

Given that the RCT stops at analysis $m$, the sufficient statistic for the data collected is $(M = m, T_m = t_m)$. Different scales may be considered for the continuation sets, and these scales are discussed in more detail below. Since these scales are one-to-one transformations of each other in the presence of $\sigma^2$, it is sufficient to discuss continuation sets in the context of general scale $T$.

In the general one-sided testing case, continuation sets $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_J$ are defined so that the RCT is stopped upon having enough confidence to make a decision among the following hypotheses regarding the parameter of interest $\theta$:

- $H_0 : \theta \leq \theta_0$
- $H_+ : \theta > \theta_0$,

where $\theta_0$ is a known, fixed value. If larger values of $\theta$ lead to higher values of the outcome of interest, on average, then upon stopping the RCT at analysis $j$ $H_0$ would be rejected in favor of $H_+$ if $T_j$ was "sufficiently large". This framework can easily be extended to two-sided tests.

In the two-sided testing case, $T_j$ may be considered to be "sufficiently large", "sufficiently small", or "sufficiently average". This can be reflected in how the continuation regions are defined, in that if $T_j$ cannot be categorized into one of these three groups then more data must be collected. A reasonable form of these continuation regions is

$$\mathscr{C}_j \equiv \mathscr{C}_{T_j} \coloneqq \left(a_{T_j}, b_{T_j}\right] \cup \left[c_{T_j}, d_{T_j}\right),$$

with $a_{T_j} \leq b_{T_j} \leq c_{T_j} \leq d_{T_j}$. To ensure that $P\big(T_j \in \mathscr{C}_{T_j}\big) > 0$ for all analyses before analysis $J$, it must be the case that $\left|b_{T_j} - a_{T_j}\right| + \left|d_{T_j} - c_{T_j}\right| > 0$ for $j \in \{1, 2, 3, \ldots, J-1\}$. Additionally, so that the probability of continuing the RCT beyond analysis $J$ is zero, it must be that $|b_{T_J} - a_{T_J}| + |d_{T_J} - c_{T_J}| = 0$.

If it is of interest to only stop the RCT when the test statistic is "sufficiently large" or "sufficiently small", then $\mathcal{C}_{T_j}$ can be defined with $b_{T_j} = c_{T_j}$ for $j \in \{1, 2, 3, \ldots, J\}$, so that each continuation region consists of one continuous interval, not two disjoint intervals.

There are two common special cases of this scenario where all the continuation regions consist of single intervals. If it is determined that there will be no early stopping for futility, $a_{T_1} = a_{T_2} = \ldots = a_{T_{J-1}} = -\infty$. Here, it would only be possible to terminate the RCT early for large enough values of $T_j$, given that at least one of $d_{T_1}, d_{T_2}, \ldots, d_{T_{J-1}}$ was finite. Similarly, setting $d_{T_1} = d_{T_2} = \ldots = d_{T_{J-1}} = \infty$ would only allow for early stopping

if at least one of $a_{T_1}, a_{T_2}, \ldots, a_{T_{J-1}}$ was finite. The group sequential design framework also has the fixed sample design as a special case, where $a_{T_j} = -\infty$, $d_{T_j} = \infty$, and $b_{T_j} = c_{T_j}$, for $j \in \{1, 2, \ldots, J-1\}$.

### 1.2.2 Scales for Test Statistics and Boundaries

Emerson, Kittelson, & Gillen[3] describe different scales of the test statistic on which boundaries $\left\{a_{T_j}, b_{T_j}, c_{T_j}, d_{T_j}\right\}$ can be derived. The following are brief descriptions of these scales at analysis $j$, in the setting of immediately available outcomes with equal sample sizes between the treatment groups under $H_0 : \theta = 0$:

- Partial Sum Scale: $S_j \coloneqq \frac{N_j}{2}\hat{\theta}_j$, where $\hat{\theta}_j$ is the distribution-free estimator from Section 1.1 evaluated at the $j^{th}$ analysis.
- Sample Mean Scale: $\overline{X}_j \coloneqq \hat{\theta}_j$. This is a crude estimate of the treatment effect.
- Z Scale: $Z_j \coloneqq \sqrt{N_j}\frac{\overline{X}_j}{\sigma}$.
- Fixed Sample P Value Scale: $P_j \coloneqq 1 - \Phi(Z_j)$. This is a one-sided p-value if the observed data was the result of a fixed sample design where the sample size was $N_j$. This fixed sample p-value does not provide accurate inference in a GSD with early stopping, but can still be useful in a group sequential setting with respect to implementing stopping rules.
- Bayesian Posterior Probability Scale: Assuming a prior distribution for $\theta$ of $\mathcal{N}(\zeta, \tau^2)$, the posterior distribution of $\theta$ given $Y_1, Y_2, \ldots, Y_{N_j}$ is $\mathcal{N}\left(\frac{N_j\tau^2\overline{Y}_j + \sigma^2\zeta}{N_j\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{N_j\tau^2+\sigma^2}\right)$. The Bayesian posterior probability statistic is then $B_j(\zeta, \tau^2, \theta_*) \coloneqq$ $P\left(\theta > \theta_* \middle| Y_1, Y_2, \ldots, Y_{N_j}\right) = 1 - \Phi\left(\frac{\theta_*\left(N_j\tau^2+\sigma^2\right) - N_j\tau^2\overline{Y}_j - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2+\sigma^2}}\right) \longrightarrow_{\tau^2\to\infty}$ $1 - \Phi\left(\sqrt{N_j}\frac{\theta_* - \overline{Y}_j}{\sigma}\right)$. Though the interpretation is different, letting $\tau^2$ go to $\infty$ produces a form similar to to the fixed sample p-value.
- Conditional Power Scale: $C_j(d_J, \theta_*) \coloneqq P\left(\overline{Y}_J > d_J \middle| \overline{Y}_j, \theta = \theta_*\right) =$ $1 - \Phi\left(\frac{N_J(d_J - \theta_*) - N_j\left(\overline{Y}_j - \theta_*\right)}{\sigma\sqrt{N_J - N_j}}\right)$. The conditional power depends on the value of $\theta_*$.

- Predictive Probability Scale: Assuming a prior distribution for $\theta$ of $\mathcal{N}(\zeta, \tau^2)$, the predictive probability statistic is $H_j(d_J, \zeta, \tau^2) \coloneqq \int P\big(\overline{Y}_J > d_J \big| \overline{Y}_j, \theta\big) \lambda\big(\theta \big| \overline{Y}_j\big) d\theta = 1 - \Phi\left( \frac{N_J\big(N_j\tau^2+\sigma^2\big)\big(d_J - \overline{Y}_J\big) + \sigma^2(N_J - N_j)\big(\overline{Y}_j - \zeta\big)}{\sigma\sqrt{(N_J - N_j)(N_J\tau^2+\sigma^2)(N_j\tau^2+\sigma^2)}} \right) \longrightarrow_{\tau^2 \to \infty} 1 - \Phi\left( \frac{N_J\big(d_J - \overline{Y}_j\big)}{\sigma\sqrt{\frac{N_J}{N_j}(N_J - N_j)}} \right)$, where $\lambda\big(\theta \big| \overline{Y}_j\big)$ is the posterior distribution of $\theta$.

These scales, along with the Error Spending Scale (discussed in Section 1.2.4), are 1:1 transformations of each other when $\sigma^2$ is known. Therefore, with known $\sigma^2$, deriving boundaries on one scale yields corresponding boundaries on all of the other scales. However, for computational reasons it may be more convenient to work on certain scales.

The above definitions of the different scales reflect cumulative statistics. For some scales, it makes sense to also consider incremental test statistics calculated from data collected in between two adjacent analyses. Incremental sample sizes are defined as $\tilde{N}_1 \coloneqq N_1$, and $\tilde{N}_j \coloneqq N_j - N_{j-1}$ for $j = 2, 3, \ldots, J$. For a given analysis $j$, the incremental statistic would then be $\tilde{S}_j \coloneqq S_j - S_{j-1}$ on the Partial Sum Scale, $\tilde{\overline{X}}_j = \frac{\tilde{S}_j}{\tilde{N}_j}$ on the Sample Mean Scale, and $\sqrt{\tilde{N}_j}\frac{\tilde{\overline{X}}_j}{\sigma}$ on the Z Scale.

Though Bayesian procedures for GSDs exist, they are not considered in great detail in this dissertation. Published articles on this topic include Lewis & Berry[4] and Emerson et al.[5].

### 1.2.3 Sampling Density of GSD Test Statistic

Calculating appropriate boundaries satisfying operating characteristics requires knowledge of the joint distribution of the test statistics across interim analyses. This is made easier when the test statistics are known to have the "independent increment structure", a phrase which can be illustrated within the context of cumulative score statistics.

If $L(\theta|\boldsymbol{Y})$ is the likelihood (full or partial) of $\theta$ given observed data $\boldsymbol{Y}$, then the score is defined as $\mathcal{U}(\theta) \coloneqq \frac{\partial L(\theta|\boldsymbol{Y})}{\partial\theta}$, and the Fisher information is $I(\theta) \coloneqq -\mathrm{E}\Big(\frac{\partial \mathcal{U}(\theta)}{\partial\theta}\Big)$. For $H_0 : \theta = \theta_0$, the score statistic is $\mathcal{S}(\theta_0) \coloneqq \frac{\mathcal{U}(\theta_0)}{\sqrt{I(\theta_0)}}$.

For cumulative score statistics $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K$, the "independent increment structure" is said to be in place if

$$\begin{aligned} &\text{Cov}(\mathcal{S}_{k_1}, \mathcal{S}_{k_2}) = \text{Var}(\mathcal{S}_{k_1}) \\ \iff &\text{Cov}(\mathcal{S}_{k_1}, \mathcal{S}_{k_2} - \mathcal{S}_{k_1}) = 0, \end{aligned}$$

with $1 \leq k_1 \leq k_2 \leq K$. In the context of a RCT with sequential analyses, this means that the design has an "independent increment structure" if the statistical information obtained before any given analysis is independent of that collected after the analysis.

In the presence of the "independent increment structure", Armitage, McPherson, and Rowe demonstrated that on the Partial Sum Scale (as described in Section 1.2.2) given $\theta$, the Partial Sum test statistic $S_j$ at analysis $j$ has sampling density

$$p(j, s; \theta) = \begin{cases} f(j, s; \theta) & s \notin \mathcal{C}_{S_j} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{C}_{S_j}$ is the continuation region on the Partial Sum Scale, and $f(j, s; \theta)$ has the following recursive definition:

$$\begin{aligned} f(1, s; \theta) &= \frac{1}{\sqrt{n_1}\sigma}\phi\left(\frac{s - n_1\theta}{\sqrt{n_1}\sigma}\right) \\ f(j, s; \theta) &= \int_{\mathcal{C}_{j-1}} \frac{1}{\sqrt{\tilde{n}_j}\sigma}\phi\left(\frac{s - u - \tilde{n}_j\theta}{\sqrt{\tilde{n}_j}\sigma}\right) f(j-1, u; \theta) du, \end{aligned}$$

for $j = 2, 3, \ldots, J$, with $n_j$ being the cumulative sample size by analysis $j$, $\tilde{n}_j$ being the incremental sample size accrued between the $(j-1)^{st}$ and $j^{th}$ analyses, and $\phi(\cdot)$ is the density function for the standard normal distribution. We note that $f(j, s; \theta)$ generally does not have a closed form, and is therefore typically calculated numerically.

A number of approaches have been proposed in the literature that demonstrate how to use this sampling density to calculate critical values for test statistics in a group sequential design, so that type I error can be controlled. Among the most well known of these critical

values are Pocock boundaries[6] and O'Brien-Fleming boundaries[7]. Pocock boundaries and O'Brien-Fleming boundaries are constant on the Z Scale and the Partial Sum Scale, respectively. These boundaries and some other boundaries are described in a unified manner by Kittelson & Emerson[8].

### 1.2.4 Error Spending Functions

Error spending functions dictate how much cumulative type I error is spent according to the amount of statistical information gathered so far, relative to the maximal statistical information allowed at the last possible analysis. A given error spending function $\mathcal{E}(\cdot)$ is defined on $[0, 1]$ and typically continuous and non-decreasing, with $\mathcal{E}(0) = 0$ and $\mathcal{E}(1) = \alpha$, where $\alpha$ is the desired type I error rate. The following are some common error spending functions as described by Demets & Lan[9]:

- $\mathcal{E}_1(t) = 2 \times \left(1 - \Phi\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{t}}\right)\right)$
- $\mathcal{E}_2(t) = \alpha \times \log(1 + (e-1) \times t)$
- $\mathcal{E}_3(t) = \alpha \times t$,

where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the cumulative distribution function and quantile function of the standard normal distribution, respectively.

Stopping boundaries derived from $\mathcal{E}_1(\cdot)$ are sometimes described as "O'Brien-Fleming-like" because if analyses in a GSD are equally spaced with respect to statistical information, they may be similar to O'Brien-Fleming boundaries, which are conservative early. Similarly, stopping boundaries derived from $\mathcal{E}_2(\cdot)$ may be similar to Pocock boundaries when the analysis times are evenly spaced, being approximately efficient with respect to average sample sizes. However, O'Brien-Fleming boundaries and Pocock boundaries can look very different from boundaries derived from these error spending functions if the analysis times are not evenly spaced[3]. In addition to this, the way O'Brien-Fleming boundaries spend error during the RCT can vary depending on the overall type I error level $\alpha$. Because no one error spending function adequately approximates O'Brien-

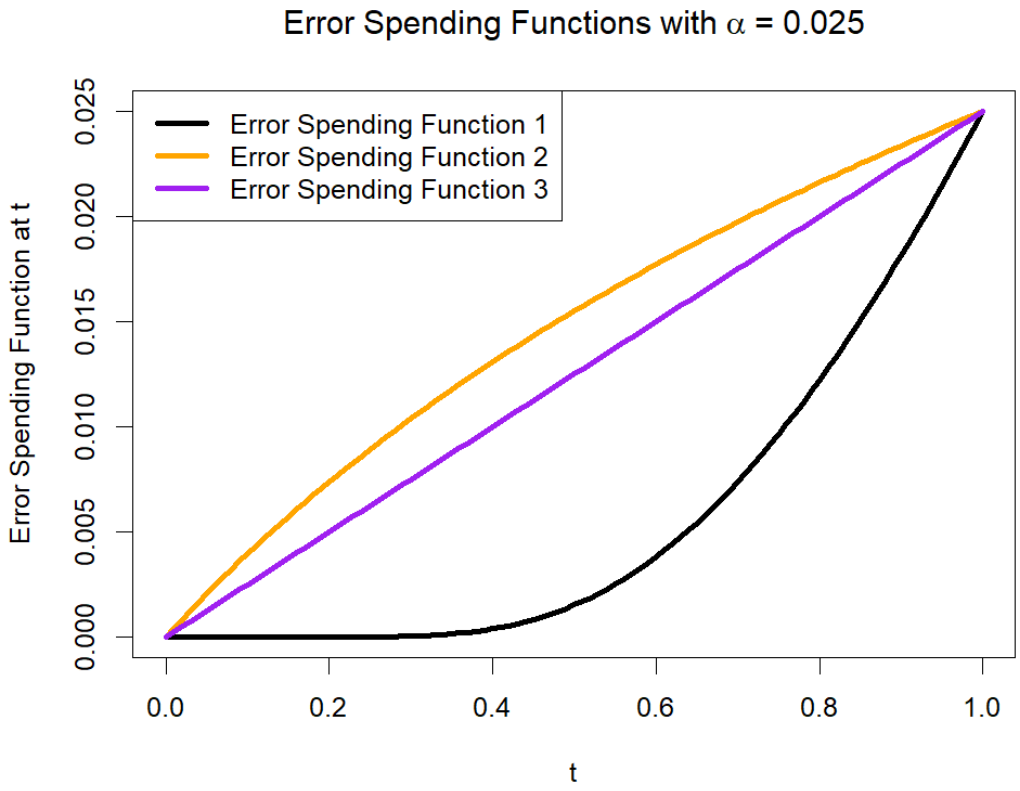



Figure 1.1: The black, orange, and purple lines correspond to $\mathcal{E}_1(\cdot)$, $\mathcal{E}_2(\cdot)$, and $\mathcal{E}_3(\cdot)$, respectively. $t = 0$ corresponds to the beginning of the RCT, where no data has been collected yet, while $t = 1$ corresponds to the end of the trial, where the maximal amount of statistical information has been collected. Early on, $\mathcal{E}_1(\cdot)$ spends very little error compared to the other two error spending functions.

Fleming or Pocock boundaries across all scenarios, it is incorrect to refer to any one error spending function as "the O'Brien-Fleming error spending function" or "the Pocock error spending function". Nevertheless, $\mathcal{E}_1(\cdot)$ is useful in that boundaries derived from it are conservative early like O'Brien-Fleming boundaries, and $\mathcal{E}_2(\cdot)$ is useful in that boundaries derived from it are not as conservative early, in a manner similar to Pocock boundaries. Pocock boundaries tend to be approximately efficient with respect to the average sample size.

The choice of error spending function reflects the kind of GSD desired for the RCT in question. For example, a DSMB may favor the use of an error spending function that is conservative early because the low probability of stopping early implies that the amount of safety data collected will likely be greater than if a less conservative error spending

function had been used.

## 1.3 Adaptive Designs

A key feature of the GSD is that the design (including rules determining $J$ and $N_1, N_2, \ldots, N_J$) is prespecified independent of any estimates of the treatment effect calculated using data from the RCT. Adaptive designs that go beyond the GSD allow for accruing data to affect arbitrary aspects of the RCT, and this is an area of ongoing research which is motivated by the desire for more flexible RCTs.

However, the statistical properties of adaptive designs have not yet been explored thoroughly. From a regulatory perspective, this is an issue because RCTs whose results are included in new drug applications must have strict type I error control. The FDA affirms this point in its draft guidance for industry in regard to adaptive designs for drugs and biologics when it states that ensuring control of the type I error rate remains critical[10]. In this draft guidance, the FDA considers fixed sample designs and GSDs to be "well-understood". The following is a list of designs considered to be "less well-understood":

- Adaptations for dose selection studies
- Adaptive randomization based on relative treatment group responses
- Adaptation of maximal sample size based on interim effect size estimates
- Adaptation of patient population based on treatment-effect estimates
- Adaptation for endpoint selection based on interim estimate of treatment effect
- Adaptation of multiple study design features in a single study
- Adaptations in non-inferiority studies

While a number of these "less well-understood" designs need further investigation into their control of operating characteristics, the focus of this dissertation is on adaptation of a GSD's schedule and timing of interim and final analyses based on interim data. The

choice to restrict the focus to this type of adaptation is justified in Section 1.3.1. From Section 1.3.2 on, the phrase "adaptive design" is reserved to refer to designs with this specific kind of adaptation that are not GSDs.

### 1.3.1 Parameter of Interest

It is of interest to develop notation to describe classification of adaptations. When summarizing the effect $\theta$ of the experimental treatment compared to the control treatment across the entire study population, a treatment effect is typically estimated for each patient participating in the RCT. For patient $i$, the treatment effect $\theta_i$ is a function of the following vectors of variables/summaries:

- $\boldsymbol{W}_i$: Characteristics defining disease and subject characteristics
- $\boldsymbol{X}_i$: Treatment characteristics
- $\boldsymbol{Y}_i$: Primary outcome variables
- $\boldsymbol{A}_i$: Auxiliary outcome variables

The population-level parameter of interest is $\theta = \theta(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{A})$. Often we define within-group summary measures $w(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{A})$ so that $\theta$ is a difference or ratio of $w(\cdot|X = 1, \cdot)$ and $w(\cdot|X = 0, \cdot)$.

A number of the different types of adaptations described in Section 1.3 correspond to modifying the distributions of some of these vectors for either patients accrued after the adaptation or all patients accrued into the RCT, and/or $w(\cdot)$. The following is a list of some examples:

- Adaptive sample size re-estimation: The sample size at the final analysis is adapted according to results at an interim analysis.
- Adaptive randomization ratios: At an interim analysis, the randomization ratios for the treatment arm assignments moving forward are adapted, according to results observed at that analysis.

- Adaptive enrichment: Eligibility criteria is adapted at an interim analysis, so that moving forward patients accrued into the clinical trial are more likely to respond positively to the experimental treatment.
- Adaptive selection of doses/treatment: At an interim analysis, one or more dose/treatment arms may be closed off to new patients, due to interim results suggesting harm or a lack of effect for those arms.

While it is certainly of interest to explore the operating characteristics of different types of adaptive designs, it is simply not feasible to complete an exploration of this size in a satisfactory manner for this dissertation. Our investigation is therefore restricted to adaptive sample size re-estimation. The results and discussion that follow will not pertain to the entire class of adaptive designs, but can provide both insight as to how other adaptive designs might behave and a good starting point for research into the operating characteristics of other types of adaptive designs.

### 1.3.2 Type I Error Control in Adaptive Designs

Analyzing data from a RCT with an adaptive design can lead to inflated type I error if the adaptations are not adjusted for correctly. Proschan & Hunsberger[11] demonstrated that with immediately observed outcomes, an adaptive design with just two analyses aiming for a level of one-sided type I error of $\alpha$ using critical value $z_\alpha$ can have the type I error inflated by as much as $\frac{1}{4}e^{-\frac{z_\alpha^2}{2}}$ when modifying the final sample size at the interim analysis. This is accomplished by choosing the (possibly infinite) amount of additional data to collect after the interim analysis such that the type I error conditional on the interim results is maximized. If $\alpha = 0.025$, then the true type I error rate can roughly be as high as 0.062, over twice the level that is desired. This amount of type I error inflation is nowhere near what is acceptable in the regulatory setting, so analyzing data from an adaptive design as if it came from a fixed sample design or a GSD is inadequate.

The following are a number of procedures proposed in the literature to control type I

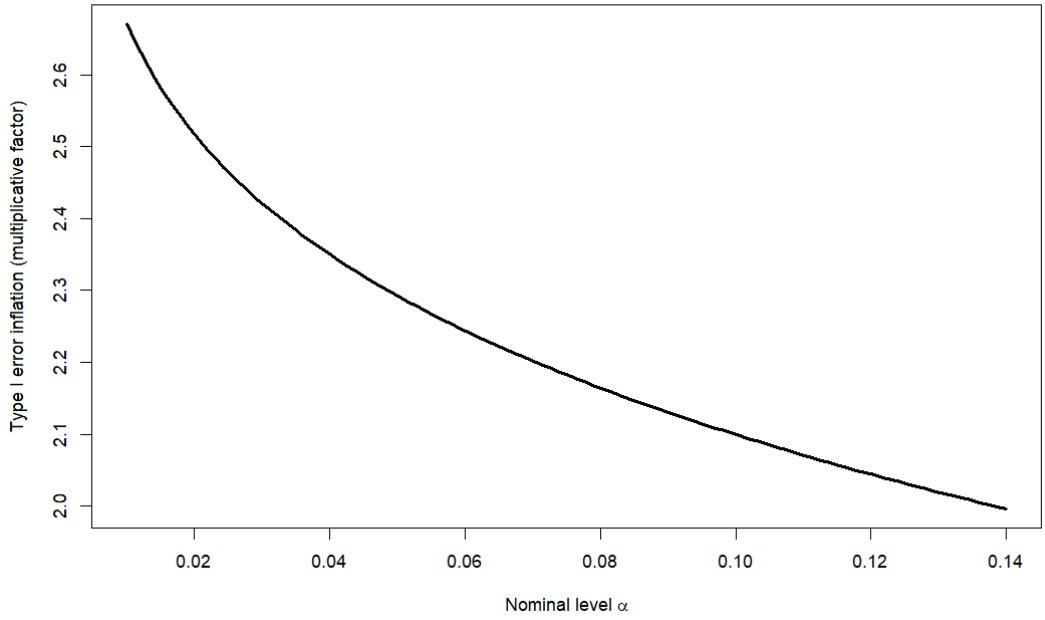



Figure 1.2: For $\alpha$ arbitrarily close to 0 up to about 0.138, the type I error is over double the nominal value.

error after the final sample size has been modified. Jennison & Turnbull[12] note that under certain settings, these methods are all equivalent.

#### 1.3.2.1 Combining P-Values

One approach to controlling the type I error rate is to combine independent p-values pertaining to patients enrolled into the RCT before and after the time of adaptation. If the distribution of this combination of p-values is known under the null hypothesis, then an appropriate rejection region can be defined, according to the desired level of the type I error rate.

Bauer & Köhne[13] present one way of doing this, based on R.A. Fisher's method[14] of combining independent p-values for meta-analyses. Suppose that a type I error level $\alpha$ is desired, and $c_\alpha$, $\alpha_0$, and $\alpha_1$ are chosen so that $c_\alpha \leq \alpha_1 \leq \alpha \leq \alpha_0 \leq 1$. $p_1$ is the p-value calculated at the interim analysis. If $p_1 \leq \alpha_1$, the interim analysis stops and the null hypothesis is rejected. If $p_1 \geq \alpha_0$, the null hypothesis is not rejected with early stopping. At the final analysis, the null hypothesis is rejected if $p_1 p_2 \leq c_\alpha$, where $p_2$ is the p-value

calculated from the data accrued after the interim analysis. Because the overall type I error of this procedure is $\alpha_1 + c_\alpha\,(\log\alpha_0 - \log\alpha_1)$, the three constants should be chosen so that this expression is equal to $\alpha$.

A different approach is to convert independent p-values to normal scores that are combined as a weighted average. If $p_1$ and $p_2$ are stagewise incremental p-values, $w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)$ is compared to $z_\alpha$, where $w_1$ and $w_2$ are prespecified such that $w_1^2 + w_2^2 = 1$.

#### 1.3.2.2 Conditional Error Preservation

Proschan & Hunsberger[11] noted that at an interim look of the data, modification of the amount of additional data to be collected can lead to the conditional probability of rejecting $H_0$ changing if the final analysis proceeds as if no adaptation had occurred. They proposed an approach to control the overall type I error by controlling the conditional type I error.

Suppose that under the original analysis plan, $H_0$ will be rejected at the final analysis if test statistic $T_{final}$ exceeds some fixed critical value $c_0$. If test statistic $T_{interim}$ is observed at an interim look of the data, the conditional type I error is $P_{H_0}(T_{final} > c_0 | T_{interim})$. Under the procedure proposed by Proschan & Hunsberger, the total sample size can be modified as long as the appropriate critical value $c^*$ (defined below using equation 1.1) is used at the final analysis. The procedure in its most general form only requires the definition of some prespecified increasing function $A(t_{interim})$ with range $[0,1]$, satisfying

$$\int_{-\infty}^{\infty} A(t_{interim}) f(t_{interim}) dt_{interim} = \alpha, \tag{1.1}$$

for type I error level $\alpha$ and density $f(\cdot)$. After observing test statistic $T^*_{final}$ with final sample size $n^*$, $c^*$ is chosen so that $P_{H_0}\left(T^*_{final} > c^* | T_{interim}\right) = A(T_{interim})$. While not required, $A(t_{interim})$ is often prespecified to be $P_{H_0}(T_{final} > c_0 | t_{interim})$.

#### 1.3.2.3 Variance Spending

In a setting with no early stopping and no adaptation, the distribution of the test statistic $T_{final}$ at the time of the final analysis often has a standard normal distribution under the null hypothesis. If at some time strictly between the beginning of data accrual and the time of the final analysis there is an interim look at the data, the data collected by the time of the final analysis can be partitioned into two groups pertaining to that collected before and after the time of the interim look. Test statistics calculated solely from these groups can be normalized according to their relative contributions to the final test statistic so that their sum $T_{before}+T_{after}$ is equal to $T_{final}$. Under $H_0$, $\text{Var}(T_{final}) = 1$ implies that $\text{Var}(T_{before}) < 1$ and $\text{Var}(T_{after}) < 1$.

Fisher[15] introduced the "variance spending" concept where investigators have spent $\text{Var}(T_{before})$ of the total variance by the time of the interim look. At the time of the interim look, investigators may desire to change the amount of additional data to be collected before the time of the final analysis. Fisher demonstrated that the type I error can be controlled as long as the incremental test statistic calculated from the data collected after the interim analysis is appropriately normalized. Under $H_0$, the appropriately normalized statistic $T^*_{after}$ has variance equal to $1 - \text{Var}(T_{before})$.

#### 1.3.2.4 Weighted Averages of Incremental Test Statistics

Recall the group sequential setting with $J$ total analyses. The test statistic at the $J^{th}$ analysis can be expressed as the following weighted average:

$$Z_J = Z_{J-1}\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J}} + \tilde{Z}_J\sqrt{\frac{\tilde{\mathcal{I}}_J}{\mathcal{I}_J}},$$

where $\mathcal{I}$ is typically proportional to the sample size, but may correspond to a different measure in some settings. Suppose that at the $(J-1)^{st}$ analysis, it is decided that $\mathcal{I}_J$ will be modified to $\mathcal{I}^*_J$. Without any adjustment, the test statistic at the $J^{th}$ analysis is

still a weighted average of two incremental statistics, now expressed in the following way:

$$Z_J^* = Z_{J-1}\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J^*}} + \tilde{Z}_J^*\sqrt{\frac{\tilde{\mathcal{I}}_J^*}{\mathcal{I}_J^*}},$$

where $\tilde{\mathcal{I}}_J^* := \mathcal{I}_J^* - \mathcal{I}_{J-1}$ and $\tilde{Z}_J^*$ is the incremental test statistic based on the last incremental sample pertaining to statistical information $\tilde{\mathcal{I}}_J^*$. As discussed previously, the use of an unadjusted test statistic such as this one in the immediately observed outcomes setting can result in an increase in the type I error by a factor greater than 2.

Cui, Hung, & Wang[16] demonstrated that by using a specific set of weights, type I error control can be achieved. Specifically, they suggest keeping the incremental test statistics corresponding to the unadjusted test statistic after modifying $\mathcal{I}_J$, but using the weights from before the modification, so that the test statistic using the CHW adjustment is the following:

$$Z_J^{CHW} = Z_{J-1}\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J}} + \tilde{Z}_J^*\sqrt{\frac{\tilde{\mathcal{I}}_J}{\mathcal{I}_J}}.$$

If $\mathcal{I}_J^* > \mathcal{I}_J$, then each of the patients enrolled into the RCT before adapting has more weight than each of the following patients. Similarly, if $\mathcal{I}_J^* < \mathcal{I}_J$, the patients enrolled by the time of the penultimate analysis are relatively downweighted. In addition to concerns regarding the efficacy of such an approach, there may be some ethical concerns with giving certain patients more weight than other patients.

Mehta & Pocock[17] borrow ideas from Cui, Hung, & Wang when constructing their adaptive rule and analysis procedure. They state that under the null hypothesis, $P(Z_2^* > b(z_1, \mathcal{I}_2^*)) = \alpha$, where

$$b(z_1, \mathcal{I}_2^*) := \frac{1}{\sqrt{\mathcal{I}_2^*}}\left(\left(z_\alpha\sqrt{\mathcal{I}_2} - z_1\sqrt{\mathcal{I}_1}\right)\sqrt{\frac{\tilde{\mathcal{I}}_2^*}{\tilde{\mathcal{I}}_2}} + z_1\sqrt{\mathcal{I}_1}\right).$$

Defining the "promising zone" as the set of possible values of $z_1$ such that $b(z_1, \mathcal{I}_2^*) \leq z_\alpha$, their procedure is to modify $\mathcal{I}_2$ only if $z_1$ is in the "promising zone", and to compare $Z_2^*$

to $z_\alpha$ at the final analysis. Because $P(Z_2^* \in (b(z_1, \mathcal{I}_2^*), z_\alpha)) > 0$, this analysis procedure is conservative. Notably, $Z_2^* > b(z_1, \mathcal{I}_2^*)$ if and only if $Z_2^{CHW} > z_\alpha$. In other words, Mehta & Pocock transform the rejection boundary to control the type I error when using the unadjusted test statistic, while Cui et al. transform the test statistic to control the type I error when using the unadjusted rejection boundary. These two approaches are mathematically equivalent.

### 1.3.3 Efficiency and Estimation with Adaptive Designs

It is reasonable to suspect that the relative flexibility of adaptive designs may result in efficiency gains over GSDs, and some have investigated this possibility. Tsiatis & Mehta[18] found that the adaptive designs they considered were less efficient than comparable GSDs. Jennison & Turnbull[12] suggested certain GSDs that can outperform adaptive designs based on Fisher's variance spending procedure. In a later publication, Jennison & Turnbull[19] found that any efficiency gains from prespecified adaptive designs are rather small, but adaptive designs that attempt to rescue underpowered studies result in efficiency losses. Levin, Emerson, & Emerson[20] reported that when holding the number of analyses constant, prespecified adaptive designs only lead to small efficiency gains compared to GSDs.

Estimation is another important aspect of adaptive clinical trials that has been investigated. Brannath, König, & Bauer[21] found that the class of flexible mean unbiased point estimates may result in large mean squared errors, and the usual maximum likelihood estimate does better with respect to this metric. They also found that with sample size modifications that are not too extreme, the median unbiased estimate performs well. Levin, Emerson, & Emerson[22] concluded that within the class of prespecified adaptive designs they considered, the bias adjusted mean performs best among all the point estimates they evaluated, and that the likelihood ratio ordering generally leads to lower p-values and narrower confidence intervals, on average, even when applied in the setting of fully adaptive designs.

## 1.4 Survival Analysis

Previously described methods are applicable to a variety of regression models, including those involving survival analysis. In this dissertation we are particularly interested in some issues arising in adaptive designs in the time-to-event setting.

There are multiple paradigms available to work under when analyzing right-censored data where the censoring is noninformative. The accelerated failure time model is a parametric model whose parameters are easily interpreted. For example, if the model assumptions hold, a simple analysis comparing two treatment groups would allow for inference on the ratio of mean (or median or any other quantile of) survival times between the two groups.

However, the proportional hazards model is more often assumed than the accelerated failure time model, and is used for the remainder of this dissertation. We note that the accelerated failure time model and the proportional hazards model are equivalent when the survival times follow a Weibull distribution. Furthermore, if the log hazard is approximately linear in log time, then an accelerated failure time model is well approximated by a Weibull distribution, which can be specified correctly by both the accelerated failure time model and the proportional hazards model.

Under the semi-parametric proportional hazards model, when comparing two treatment groups,

$$h_1(t) = h_0(t)e^{\beta X},$$

where $h_0(\cdot)$ and $h_1(\cdot)$ are the hazard functions for the control and experimental treatment groups, respectively, $X$ is an indicator variable for being in the experimental treatment group, and $\beta$ is an unknown but fixed parameter. The hazard ratio, $e^{\beta}$, is typically the summary measure of interest when using this model, and the null hypothesis $H_0 : e^{\beta} = e^{\beta_0}$ can be tested with the log-rank test or by a Cox proportional hazards regression model.

In a time-to-event setting, there are a number of variables that correspond to an individual

patient:

- $E_i$: Calendar time of accrual of patient $i$ into the RCT
- $T_i$: Amount of time from $E_i$ to the time of the event corresponding to the primary outcome for patient $i$
- $C_i$: Amount of time from $E_i$ to the time of loss to follow-up for patient $i$
- $X_i$: An indicator variable that patient $i$ is in the experimental treatment group

In such a setting, it is not always feasible to wait until all patients accrued into the RCT are either observed to experience an event or be lost to follow-up. As a result, it may be decided that the data will be analyzed at some time $\tau$, and for each patient the following information will be available:

- $Y_i \coloneqq \max(\min(T_i, C_i, t - E_i), 0)$, the amount of time enrolled in the study without experiencing the event, being lost to follow-up, or being censored administratively
- $D_i \coloneqq \mathbb{1}_{\{Y_i < \min(C_i, t - E_i)\}}$, an indicator of whether or not the patient was observed to experience the event
- $X_i$

Note that while in previous sections $T$ represents a general test statistic, here it represents a time to the event of primary interest. Moving forward, we allow for the context of the discussion to make clear what $T$ represents. Furthermore, for the remainder of this dissertation, $C_i$ is set to $\infty$ $\forall i$, so that any censoring is purely administrative.

In a group sequential setting, the data may be analyzed at times $\tau_1, \tau_2, \ldots, \tau_J$. For each patient, the values of $Y_i$, $D_i$, and $X_i$ may vary depending on the time of the analysis.

The partial score statistic for the Cox proportional hazards regression model is

$$\mathcal{U}(\beta) = \sum\nolimits_{i=1}^{n} \frac{n_1(Y_i) n_0(Y_i)}{n_0(Y_i) + n_1(Y_i) e^{\beta}} \left( \frac{D_i T_i}{n_1(Y_i)} - \frac{D_i \left(1 - T_i\right)}{n_0(Y_i)} e^{\beta} \right),$$

where the total sample size is $n$, and the numbers at risk in the experimental and control

treatment groups at time $t$ are $n_1(t)$ and $n_0(t)$, respectively.

The numerator of the log-rank test statistic is $\mathcal{U}(\beta)$ evaluated at the null value of $\beta = 0$, and is a weighted average of the difference in hazards between the two treatment groups:

$$\mathcal{U}(\beta)|_{\beta=0} = \sum_{i=1}^{n} \frac{n_1(Y_i)n_0(Y_i)}{n_0(Y_i) + n_1(Y_i)} \left( \frac{D_i T_i}{n_1(Y_i)} - \frac{D_i \left(1 - T_i\right)}{n_0(Y_i)} \right),$$

where the total sample size is $n$, and the numbers at risk in the experimental and control treatment groups at time $t$ are $n_1(t)$ and $n_0(t)$, respectively.

The denominator of the log-rank test statistic is the square root of the following estimator:

$$\hat{\mathrm{Var}}\Big(\mathcal{U}(\beta)|_{\beta=0}\Big) = \sum_{i=1}^{n} \frac{n_1(Y_i)n_0(Y_i)}{n_0(Y_i) + n_1(Y_i)} \left( 1 - \frac{D_i}{n_0(Y_i) + n_1(Y_i)} \right) \frac{D_i}{n_0(Y_i) + n_1(Y_i)}.$$

Under the strong null hypothesis ($h_1(t) = h_0(t)$), the log-rank test statistic $\frac{\mathcal{U}(\beta)|_{\beta=0}}{\sqrt{\hat{\mathrm{Var}}\left(\mathcal{U}(\beta)|_{\beta=0}\right)}} \longrightarrow_d \mathcal{N}(0,1)$.

When fitting a Cox proportional hazards regression model comparing two treatment groups, if the only covariate in the model is an indicator variable for the treatment group, then the log-rank test is the score test for this model, and is therefore asymptotically equivalent to the corresponding Wald-based test under the strong null hypothesis.

Once again assuming a 1:1 randomization scheme, the log-rank test statistic[23] has an asymptotic normal distribution in a fixed sample design, with $Z \dot{\sim} \mathcal{N}(\delta, 1)$, where $\delta = \frac{\sqrt{d}(\log(\lambda) - \log(\lambda_0))}{\sqrt{V}}$. $d$ is the number of events observed in the entire study group at the time of the analysis, $\lambda$ is the hazard ratio comparing the experimental treatment group to the control group, $\lambda_0$ is the value of $\lambda$ under the null hypothesis, and $V = 4$. The approximate statistical information is $\mathcal{I} = \frac{d}{V}$. We might design a fixed sample design such that a one-sided level $\alpha$ test of $H_0 : \log(\lambda) = \log(\lambda_0)$ would detect $H_a : \log(\lambda) \geq \log(\lambda_a)$ with power $\beta$. In that case we require that the number of events observed is $d = \frac{\left(z_{1-\alpha} + z_\beta\right)^2 V}{\left(\log(\lambda_a) - \log(\lambda_0)\right)^2}$, where $z_\alpha = \Phi^{-1}(\alpha)$ is the $\alpha$-th quantile of the standard normal distribution.

It has been shown that in a group sequential setting, the increments of the log-rank test

statistic are asymptotically uncorrelated[24]. Because the "independent increment structure" is present in the asymptotic sense, the usual procedures to determine appropriate stopping boundaries for GSDs carry over to the time-to-event setting directly, with the only difference being that the sample size now corresponds to the number of observed events.

## 1.5 Adaptive Methods in Survival Analysis

When both the proportional hazards assumption and the "independent increment structure" hold, methods to control the type I error rate in adaptive designs generally apply to the survival setting. When the proportional hazards assumption is violated, the censoring distribution can affect what is being estimated. However, the scope of this dissertation is restricted to scenarios where the proportional hazards assumption holds.

The CHW adjustment can still be used in this setting, with a slightly different approach. The log-rank test statistic is cumulative in nature, so it cannot be expressed directly as a weighted average of incremental test statistics. However, increments of the log-rank test statistic are asymptotically independent[24]. Given only $Z_{J-1}$ and $Z_J^*$, $\tilde{Z}_J^*$ can easily be solved for algebraically using the weighted average representation of $Z_J^*$, to determine the contribution of data collected after the $(J-1)^{st}$ analysis and refer to the algebraic expression of $\tilde{Z}_J^* = \left(Z_J^* - Z_{J-1}\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J^*}}\right)\sqrt{\frac{\mathcal{I}_J^*}{\tilde{\mathcal{I}}_J^*}}$ as the incremental test statistic corresponding to the data with statistical information $\mathcal{I}_J^*$, thus allowing for a direct application of the CHW adjustment. For computational reasons, it may be faster to plug this expression into the formula for the CHW adjustment and simplify to a weighted average of the interim and unadjusted final test statistics, to avoid the calculation of the contribution of the data collected after the penultimate analysis. The CHW adjusted statistic in this case is:

$$Z_J^{CHW} = Z_{J-1}\sqrt{\frac{\mathcal{I}_{J-1}}{\mathcal{I}_J}}\left(1 - \sqrt{\frac{\tilde{\mathcal{I}}_J}{\tilde{\mathcal{I}}_J^*}}\right) + Z_J^*\sqrt{\frac{\tilde{\mathcal{I}}_J}{\mathcal{I}_J}}\sqrt{\frac{\mathcal{I}_J^*}{\tilde{\mathcal{I}}_J^*}}.$$

### 1.5.1 Impact of Surrogate Data

A key point of the above approaches for immediately available outcomes is that $\tilde{Z}_1$ is sufficient for the treatment effect on the primary endpoint. More recently, special attention has been drawn to adaptive designs in the time-to-event setting. In a letter to the editor, Bauer & Posch[25] point out that solutions for adaptations in the immediately available outcomes setting do not necessarily translate well to the time-to-event setting, due to the presence of surrogate data. They describe a scenario where surrogate data on patients enrolled into the RCT by the time of the interim analysis can be used to perfectly predict the event times of these patients, thus allowing for perfect prediction of future values of the log-rank test statistic if no further patients are accrued into the RCT after the interim analysis. This would allow for the analysis time to be adapted to a time where the log-rank test statistic is predicted to exceed the rejection boundary, potentially allowing for a great increase in the type I error.

There are a few methods proposed in the literature that control the type I error even in the presence of patient surrogate data. Jenkins, Stone, & Jennison[26] combine two p-values calculated from data corresponding to patients accrued into the RCT before and after the time of adaptation. Conversely, Irle & Schäfer[27], as well as Magirr et al.[28] propose related solutions, where the rejection boundary at the final analysis is calculated by determining the asymptotically normal distribution of the final test statistic conditional on the data at the final analysis observed on patients accrued into the RCT by the interim analysis time. However, this last kind of approach has a disadvantage where the final boundary cannot be calculated until the time of the final analysis. In the following chapter, we develop a framework where the data available at a given interim analysis is partitioned into four distinct groups. Because it is beneficial to discuss these methods within the context of this framework, we postpone a more detailed discussion of these methods for the time being.

As in the time-to-event setting, surrogate data can be used inappropriately when modifying a RCT's sampling plan in other settings, such as those involving longitudinal data.

However, these settings are not investigated in this dissertation. As discussed in Section 1.3.1, investigation of such adaptive designs is outside of the scope of this dissertation, though we comment on generalizability to other adaptive goals in Chapter 6.

## 1.6 Dissertation Aims

In Chapter 2, we aim to develop a framework in the time-to-event setting where the data available at a given interim analysis is partitioned into four distinct groups. We further aim to use this framework to discuss the features and limitations of methods for type I error control in the time-to-event setting put forth by Jenkins, Stone, & Jennison[26], Irle & Schäfer[27], and Magirr et al.[28]. Because we refer to these three methods extensively in later chapters, their reference numbers are provided sporadically for the remainder of this dissertation.

Following that, in Chapter 3 we quantify the operating characteristics of clinical trials where investigators either fail to adjust for the use of surrogate data in their adaptation rule or adjust for it incorrectly.

In Chapter 4, we aim to investigate alternative approaches to controlling the type I error rate in the time-to-event setting. In particular, we are interested in methods that do not require partitioning of the data according to accrual before or after adaptation.

Finally, in Chapter 5 we compare adaptive designs that use adjustments that account for the use of surrogate data, to group sequential designs and efficient adaptive designs whose adaptive rules are not functions of surrogate data and therefore do not require adjusting for such surrogate data.

# Chapter 2

# Adaptive Designs in the Presence of Surrogate Data

## 2.1 Inadequacy of Certain Analysis Methods

Some analysis methods for adaptive clinical trials use formulas that incorporate values of sufficient statistics from the available data on primary outcomes. In the absence of surrogate data, these sufficient statistics capture all available information on the parameter of interest, so these methods adequately control the type I error rate.

However, if additional surrogate data is informative of future analyses but not reflected in the sufficient statistics, type I error control can fail. Bauer & Posch[25] describe such a situation in their letter to the editor when expressing concern regarding a proposed analysis procedure. In this scenario, a level $\alpha$ log-rank test is to be used to compare the hazards of two treatment groups, and surrogate data is both available for all patients recruited into the study and able to perfectly predict the event times of those patients. Because the surrogate data can be used to predict event times with perfect accuracy, the log-rank test statistic at the planned analysis time can also be predicted, even before the first event is observed.

Bauer & Posch then describe an adaptive procedure where if the log-rank test statistic

is predicted to be statistically significant, no further patients are to be recruited into the study. Otherwise, a sufficiently large number of additional patients are to be recruited so that the contribution of the initial patients to the test statistic is negligible at the time of the analysis. With this procedure, Bauer & Posch state that the type I error rate is nearly double that of nominal value $\alpha$. Though this is an extreme example, it illustrates the danger of failing to account for surrogate data that is informative of primary outcomes that have not yet been observed.

Modifications of this example can potentially lead to even greater levels of type I error inflation. If the timing of the analysis can be adapted, then the contributions to the log-rank test statistic might be known for all patients already recruited into the study. If at any of the forecastable event times, the log-rank test statistic is predicted to be statistically significant, recruitment can be stopped, and the analysis time can be changed to the earliest such time this occurs.

## 2.2 Data Available at Time of Analysis

To discuss issues with adaptive designs in the presence of surrogate data, it helps to characterize the data that is available at the time of the analysis. Specifically, it is useful to categorize study subjects according to what data is available for analysis and prediction, as well as baseline, treatment, and outcome characteristics for each patient.

Suppose that any potential adaptations on the time of the final analysis are restricted such that the adapted time is bounded above by $\tau^{max}$. One common scenario is when a dataset is "locked" at time $\tau$ to be analyzed at time $\tau^* \geq \tau$. For given times $\tau$ and $\tau^*$, subjects are grouped in the following manner:

- $K_1(0, \tau)$: Subjects accrued up until time $\tau$
- $K_2(\tau, \tau^*)$: Subjects accrued after time $\tau$ up until time $\tau*$
- $K_3(\tau^*, \tau^{max})$: Subjects accrued after time $\tau*$

For example, if subjects $1, 2, \ldots, N_1$ are enrolled at or before time $\tau$, subjects $N_1+1, N_1+2, \ldots, N_2$ are enrolled between times $\tau$ and $\tau^*$, and subjects $N_2+1, N_2+2, \ldots, N_3$ are enrolled after time $\tau^*$, then (recalling that $E_i$ is the calendar time of accrual of patient $i$ into the RCT)

$$\begin{aligned} K_1(0,\tau) &= \{i : E_i \in [0,\tau]\} \\ &= \{1, 2, \ldots, N_1\} \\ K_2(\tau,\tau^*) &= \{i : E_i \in (\tau,\tau^*]\} \\ &= \{N_1+1, N_1+2, \ldots, N_2\} \\ K_3(\tau^*,\tau^{max}) &= \{i : E_i \in (\tau^*,\tau^{max}]\} \\ &= \{N_2+1, N_2+2, \ldots, N_3\}. \end{aligned}$$

For a general group of patients $K = \{a, a+1, \ldots, a+k\}$, the following data is available at time $\tau$:

$$\mathcal{H}(K,\tau) \coloneqq \left\{ \{\boldsymbol{W}_i\}_{i=a}^{a+k}, \{\boldsymbol{X}_i\}_{i=a}^{a+k}, \{\boldsymbol{\mathcal{A}}_i(\tau), \boldsymbol{Y}_i(\tau)\}_{i=a}^{a+k} \right\},$$

where for patient $i$, $\boldsymbol{W}_i$ are the disease and baseline characteristics, $\boldsymbol{X}_i$ are the treatment characteristics, $\boldsymbol{A}_i(\tau)$ are the surrogate outcome data censored at time $\tau$, $\boldsymbol{\mathcal{A}}_i(\tau) \coloneqq \{\boldsymbol{A}_i(t), t \leq \tau\}$, and $\boldsymbol{Y}_i(\tau)$ are the primary outcome data censored at time $\tau$.

Using the above definitions, the dataset locked at time $\tau$ to be used for the analysis at time $\tau^*$ can be described as $\mathcal{H}(K_1(0,\tau),\tau)$. However, the data actually available to investigators at analysis time $\tau^*$ for use in any adaptation procedures to be carried out is the union of $\mathcal{H}(K_1(0,\tau),\tau^*)$ and $\mathcal{H}(K_2(\tau,\tau^*),\tau^*)$.

Elements of the primary outcome data available to investigators at time $\tau^*$ are divided in the following manner:

- $\mathcal{D}_1$: Exact event times known at $\tau$ (and $\tau^*$) that will be included in the analysis
- $\mathcal{D}_2$: Exact event times known at $\tau^*$ that will not be included in the analysis

- $\mathcal{D}_3$: Event times that are unknown but can be predicted with some (perhaps non-constant) level of imprecision at the time of the analysis
- $\mathcal{D}_4$: Event times from patients not yet accrued into the study that are therefore unknown and cannot be predicted with any meaningful level of precision

Defining $A[B]$ to mean "element $B$ of $A$", $\mathcal{D}_1$ consists all observed event times in $\mathcal{H}(K_1(0,\tau),\tau)[\{\boldsymbol{Y}_i(\tau)\}]$. Events observed in $\mathcal{H}(K_1(0,\tau),\tau^*)[\{\boldsymbol{Y}_i(\tau^*)\}]$ but not observed in $\mathcal{H}(K_1(0,\tau),\tau)[\{\boldsymbol{Y}_i(\tau)\}]$, as well as events observed in $\mathcal{H}(K_2(\tau,\tau^*),\tau^*)[\{\boldsymbol{Y}_i(\tau^*)\}]$, are categorized as belonging in $\mathcal{D}_2$. Depending on how accurate predictions from auxiliary variables $\boldsymbol{A}_i(\tau^*)$ are, event times that are unobserved but can be reasonably predicted from $\mathcal{H}(K_1(0,\tau),\tau^*)[\{\boldsymbol{A}_i(\tau^*),\boldsymbol{Y}_i(\tau^*)\}]$ and $\mathcal{H}(K_2(\tau,\tau^*),\tau^*)[\{\boldsymbol{A}_i(\tau^*),\boldsymbol{Y}_i(\tau^*)\}]$ are categorized into $\mathcal{D}_2$ or $\mathcal{D}_3$. Any events from subjects in $K_3(\tau^*,\tau^{max})$ belong to $\mathcal{D}_4$.

As described above, one example in how $\mathcal{D}_1$ and $\mathcal{D}_2$ differ is that regarding a data lock. For a given upcoming interim analysis, data is locked at a certain date, and despite the RCT procedures continuing, any additional data on primary outcomes that is collected will not be included in the upcoming analysis. If a study participant experiences the event of interest between the time of the data lock and the time of the analysis, information regarding this event would be considered to belong to $\mathcal{D}_2$, not $\mathcal{D}_1$.

Another example of $\mathcal{D}_2$ might be a scenario where exactly six months before a study participant will experience an event, he or she develops a marker such as disease progression to the next stage. In such a scenario, study coordinators can add 182 days to the date of participants developing this marker, and thus predict their event times in a very accurate manner. Perhaps it might be the case that developing this marker indicates an event time between 175 days and 195 days into the future, this data belongs to $\mathcal{D}_3$.

A marker might not be so predictive of a participant's event time. If instead the marker can take different values corresponding to a participant's disease progression on a continuous spectrum, a coordinator might be able to place this participant's event time in a certain time window with a high degree of confidence. Perhaps with low values of the marker, it would be reasonably certain that the participant's event time would be

between five and seven months from the present time. Conversely, high values of the marker would indicate an event time in between two and four days into the future. Such surrogate data would be said to belong to $\mathcal{D}_3$.

The setting of a colon adjuvant RCT might serve as a classic example of events that belong to $\mathcal{D}_3$. If the primary outcome is overall survival, surrogate data such as whether or not colon cancer has been observed to recur can be informative regarding the time of a given living patient's future time of death. The events of such patients would be said to belong in $\mathcal{D}_3$. Similarly, in a HIV vaccine RCT, CD4 counts and viral load measurements might be used to accurately predict the times of death in patients who are still alive. The events of these patients would also belong to $\mathcal{D}_3$.

At the time of the interim analysis, some participants have not yet been recruited into the study. Because of this, study coordinators have no data on them. The event times for these study participants cannot be predicted with enough precision to be worthwhile, and are then said to belong to $\mathcal{D}_4$.

The scenario Bauer & Posch had described was one where the dataset was both locked and analyzed at some interim analysis time $\tau \equiv \tau^*$. In their extreme example, all events from $K_1(0, \tau)$ that are observed by time $\tau$ belong to $\mathcal{D}_1$, with the rest belonging to $\mathcal{D}_2$. The concern they raise in their letter to the editor is that adaptive analysis procedures for immediately available outcomes data assume knowledge of only $\mathcal{D}_1$ at the time of adaptation. However, in time-to-event settings knowledge of $\mathcal{D}_2$ and $\mathcal{D}_3$ at the time of adaptation is unaccounted for with these methods, and type I error may not be controlled as a result.

The problem of adaptations in the presence of surrogate data can be described more generally. Data collected before and after a dataset lock can be summarized via incremental test statistics that are averaged in a weighted manner:

$$w_1 \tilde{Z}_1 + w_2 \tilde{Z}_2.$$

Adaptive methods generally assume that at the time of adaptation all event times can be categorized into $\mathcal{D}_1$ or $\mathcal{D}_4$, so $w_1$, $w_2$, and $\tilde{Z}_1$ are known but there is no information on $\tilde{Z}_2$. However, with auxiliary data on event times, some event times assumed to belong to $\mathcal{D}_4$ actually belong to $\mathcal{D}_2$ or $\mathcal{D}_3$, and therefore partial knowledge regarding $\tilde{Z}_2$ is known. This knowledge can be used to modify the amount of additional data to be collected after the time of adaptation to achieve optimal values of incremental test statistic $\tilde{Z}_2^*$, potentially leading to inflated type I error.

To aid in the discussion of methods in following sections, RCT examples for different diseases are presented below, assuming accrual periods with Uniform distributions and event times with Exponential distributions. Specifically, these examples aim to reflect typical scenarios in lung cancer trials, breast cancer trials, and HIV vaccine trials. Parameters used in these examples are roughly based on International Adjuvant Lung Cancer Trial Collaborative Group and others[29], van der Hage et al.[30], and rgp120 HIV Vaccine Study Group and others[31], respectively. All examples assume that each analysis is carried out exactly two months after the dataset for that analysis is locked, and that time is measured in years.

We note that the choice of the length of time between $\tau$ and $\tau^*$ affects the average number of events belonging to $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$. We do not investigate how the value of $\tau^* - \tau$ affects results in the remainder of this chapter and in Chapter 3. However, in Chapter 3, scenarios where surrogate data perfectly predict event times not yet observed give some insight as to how having more events in $\mathcal{D}_2$ affects the extent to which the metrics considered deviate from their nominal values.

### 2.2.1 Lung Cancer

A lung cancer RCT might recruit 1900 participants in a period of 5 years and 11 months, with survival times on both treatment arms following an Exponential(rate = 0.19) distribution. If the RCT had a GSD with analyses at 500 events and 1000 events, the following would reflect the average times and data available at the two analyses:

| Analysis | $\tau$ | $\tau^*$ | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ | $\mathcal{D}_4$ |
|---|---|---|---|---|---|---|
| 1 | 4.64 | 4.81 | 500.00 | 31.71 | 1012.17 | 356.12 |
| 2 | 7.17 | 7.33 | 1000.00 | 28.05 | 871.95 | 0.00 |

By the time of the interim analysis ($\tau^* = 4.81$), 1543.88 participants would be enrolled, on average. Among these patients, 531.71 would have experienced the event of primary interest, and 1012.17 would still be alive. Of the observed events, 31.71 would not be included in the analysis. By the time of the final analysis ($\tau^* = 7.33$), all 1900 patients would be enrolled, 1028.05 events would be observed, and 1000 events would be included in the analysis.

### 2.2.2 Breast Cancer

A breast cancer RCT might recruit 700 participants in a period of 8 years and 1 month, with survival times on both treatment arms following an Exponential(rate = 0.04) distribution. If the RCT had a GSD with analyses at 70 events and 140 events, the following would reflect the average times and data available at the two analyses:

| Analysis | $\tau$ | $\tau^*$ | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ | $\mathcal{D}_4$ |
|---|---|---|---|---|---|---|
| 1 | 6.64 | 6.81 | 70.00 | 3.40 | 515.93 | 110.67 |
| 2 | 9.73 | 9.90 | 140.00 | 3.72 | 556.28 | 0.00 |

By the time of the interim analysis ($\tau^* = 6.81$), 589.33 participants would be enrolled, on average. Among these patients, 73.40 would have experienced the event of primary interest, and 515.93 would still be alive. Of the observed events, 3.40 would not be included in the analysis. By the time of the final analysis ($\tau^* = 9.90$), all 700 patients would be enrolled, 143.72 events would be observed, and 140 events would be included in the analysis.

### 2.2.3 HIV

An HIV RCT might recruit 5400 participants in a period of 1 year and 4 months, with survival times on both treatment arms following an Exponential(rate = 0.02) distribution.

If the RCT had a GSD with analyses at 180 events and 360 events, the following would reflect the average times and data available at the two analyses:

| Analysis | $\tau$ | $\tau^*$ | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ | $\mathcal{D}_4$ |
|---|---|---|---|---|---|---|
| 1 | 2.36 | 2.53 | 180.00 | 17.37 | 5202.63 | 0.00 |
| 2 | 4.12 | 4.28 | 360.00 | 16.77 | 5023.23 | 0.00 |

By the time of the interim analysis ($\tau^* = 2.53$), all 5400 participants would be enrolled, on average, so that no additional patients would be enrolled after the interim analysis. Among the patients, 197.37 would have experienced the event of primary interest, and 5202.63 would still be alive. Of the observed events, 17.37 would not be included in the analysis. By the time of the final analysis ($\tau^* = 4.28$), 179.40 additional events would be observed, and 360 total events would be included in the analysis.

## 2.3 Analysis Methods with Time-to-Event Data

There have been some methods proposed in the literature that do not allow for the use of surrogate data to inflate the type I error rate: Jenkins, Stone, & Jennison[26], Irle & Schäfer[27], and Magirr et al.[28]

### 2.3.1 Jenkins, Stone, & Jennison

The method proposed Jenkins, Stone, & Jennison was described within the context of subgroup selection, but may be applied to other types of adaptive designs. The general principle is the following:

1. Patients accrued before the time of adaptation are to be followed up for a prespecified amount of time.

2. P-values $p_1$ and $p_2$ are calculated from data corresponding only from patients before and after the time of adaptation, respectively.

3. The p-values are combined so that $w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)$ is compared to some critical value $c$, where $w_1$ and $w_2$ are prespecified subject to $w_1^2 + w_2^2 = 1$.

Because $p_1$ and $p_2$ are from different groups of patients, no data is available at the time of adaptation regarding what $p_2$ will be, so events from $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ will not contribute to the calculation of $p_2$. Thus, there will be no opportunity to inflate the type I error with prespecified weights $w_1$ and $w_2$. For a one-sided level $\alpha = 0.025$ test, Jenkins, Stone, & Jennison recommended a critical value of $c = 1.96$, when there is no early stopping at the time of adaptation. For a more general GSD framework, usual stopping boundaries can be used according to weights $w_1$ and $w_2$.

### 2.3.2 Irle & Schäfer

The method proposed by Irle & Schäfer is one that preserves the conditional error at the time of adaptation, even in the presence of surrogate data. Their procedure is the following:

1. Begin with a conventional study design (such as a GSD), where the analysis times with respect to calendar time or the number of events are prespecified.

2. If the final number of events $M$ is extended to $M^*$, then instead of observing test statistic $S_M$ at the final analysis, $S_{M^*}$ will be observed.

3. At the time of the final analysis, the original rejection boundary $b$ is modified to $b^*$. $b^*$ is chosen to satisfy $P_{H_0}(S_M > b|S'_M) = P_{H_0}(S^*_M > b^*|S'_{M^*})$, where $S'_m$ is the test statistic calculated with data censored at the time of the $m^{th}$ event in the study population, using data corresponding to the subpopulation of patients accrued by the time of adaptation.

One disadvantage to such an approach is that because $S'_M$ must be calculated, the design modification is restricted to $M^* \geq M$. Another is that this method is conditioning on the actual data that will be observed in the future, which is essentially making the conservative assumption that events from $K_1(0,\tau)$ and $K_2(\tau,\tau^*)$ can all be categorized

into $\mathcal{D}_1$ or $\mathcal{D}_2$ at the time of adaptation. In addition to this, the fact that $b^*$ will not be known until the time of the final analysis is potentially of concern.

### 2.3.3 Magirr et al.

Magirr et al. note that the Irle & Schäfer and Jenkins, Stone, & Jennison methods actually use the same test statistics, and that data after the originally planned follow-up time for the group of patients accrued before the time of adaptation is effectively thrown away.

Their proposed method is similar to that of of Jenkins, Stone & Jennison, except that data on the first group of patients beyond the originally planned follow-up time is included in the analyses. Of course, allowing for the additional data for the first group of patients to contribute to the test statistic can lead to type I error inflation when comparing the test statistic to $\Phi(1-\alpha)$, due to the amount of additional data from the original group depending on interim data. Using Brownian motion results, Magirr et al. present an approach of calculating a rejection boundary $b^{**} > \Phi(1-\alpha)$ such that the type I error rate is at most $\alpha$. Because $\alpha$ is an upper bound this approach may be conservative, with the type I error rate potentially being considerably lower.

### 2.3.4 Disadvantages of Methods

All three methods have a few disadvantages. First, the trial must run for a minimal length of time beyond the time of adaptation, according to the prespecified amount of time patients accrued before adaptation must be observed. For the Magirr et al. approach, the second disadvantage is that the type I error rate may be below the nominal value. However, for the Jenkins, Stone, & Jennison and Irle & Schäfer approaches, the second disadvantage is instead that in the event that the study is to run longer than the amount of time for observation of the first group of patients, some data collected on this group will not used in the final analysis, leading to concerns regarding statistical inefficiency

and ethics. Consider a scenario where at the interim analysis, it is decided that the cumulative number of events to be observed at the final analysis will be increased by 50% of its original value. In the lung cancer example RCT discussed in Section 2.2.1, 1500 events would be observed. However, on average only 1140.71 events (76.05%) would be included in the analysis with this method, due to data from the first group of participants collected after the time of the original final data lock being discarded. Even if events from all 1900 enrolled participants were to be observed at the final analysis, only an average of 1253.28 events (65.96%) would be included in the analysis.

In the breast cancer example RCT discussed in Section 2.2.2, 210 events would be observed. However, on average only 152.63 events (72.68%) would be included in the analysis with this method. If study coordinators wanted to ensure that 210 events would be included in the analysis, 528.06 of the 700 events would need to be observed, on average. In such a scenario, only 39.77% of the observed events would be included in the analysis.

In the HIV vaccine example RCT discussed in Section 2.2.3, all participants would contribute to $p_1$ since they were all enrolled before the interim analysis. Therefore, this method could not be used in the HIV vaccine RCT. Even in the examples where this method can be used, controlling the type I error rate appears to lead to notable levels of inefficiency.

A third disadvantage for all three methods is that while the final analysis adjusts for data observed at the originally planned time of the final analysis or later, of patients accrued by the time of adaptation, only a subset of this data is known at the time the adaptation occurs. In Chapters 4 and 5, we examine the consequences of this last disadvantage.

# Chapter 3

# Quantification of Operating Characteristics in the Presence of Surrogate Data

This chapter aims to evaluate the impact of failing to adjust or incorrectly adjusting for $\mathcal{D}_2 \cup \mathcal{D}_3$ when the adaptive rule utilizes $\mathcal{D}_2 \cup \mathcal{D}_3$. Type I error and power are used as metrics when evaluating different adjustment approaches under a range of scenarios, as these are among the standard values used to compare different RCT designs. In addition to these, we also use the naive Bayes Factor as a metric. The use of this last metric is described and justified below, in Section 3.2.7.

## 3.1 Motivation

Though the fact that one can use surrogate data to inflate the type I error has been pointed out in the past[25], there remains a need to quantify this inflation. The extent to which type I error can be inflated is a function of many different factors, and this chapter aims to determine how each of a subset of these factors individually affects the inflation, and if some factors have synergistic effects with respect to this inflation.

The following aspects of the GSD used at the outset of the RCT are investigated:

- Time of the interim analysis, relative to the time of the initially planned final analysis.
- The error spending function used to calculate rejection boundaries at the interim and final analyses.
- Whether or not patient accrual into the RCT is stopped at the interim analysis, should the RCT continue to the final analysis. In the scenarios considered, this decision is not a function of the data observed.

The following aspects of the adaptive rule used at the interim analysis are investigated:

- The minimum value to which the final sample size $n_2^*$ can be modified.
- The maximum value to which the final sample size $n_2^*$ can be modified.
- The availability of data to be used for the adaptation ($\mathcal{D}_1$ alone versus $\mathcal{D}_1$ and $\mathcal{D}_2$ versus $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$).
- The prognostic ability of $\mathcal{D}_3$ for the patient event time, at the individual level.

The following aspects of the analysis procedures are investigated:

- Whether or not the CHW adjustment is used on the final observed test statistic.
- Whether or not the CHW adjustment is used on the predicted future test statistics at the time of adaptation.

The order in which these factors are discussed is chosen to facilitate understanding.

## 3.2 Notation

To evaluate the three aforementioned metrics via simulation, we choose settings that are typical of RCT designs in the time-to-event setting.

### 3.2.1 Original GSD

The original GSD has an interim analysis and a final analysis at cumulative numbers of observed primary events $n_1$ and $n_2$, respectively. Given $n_2$, $n_1$ is chosen so that $\frac{n_1}{n_2} \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$.

Efficacy boundaries are calculated using one of the three error spending functions mentioned in Chapter 1 (listed below), as well as $n_1$ and $n_2$. No futility boundaries are used.

- $\mathcal{E}_1(t) = 2 \times \left(1 - \Phi\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{t}}\right)\right)$
- $\mathcal{E}_2(t) = \alpha \times \log(1 + (e-1) \times t)$
- $\mathcal{E}_3(t) = \alpha \times t$

$\mathcal{E}_1(t)$ spends little error early into a RCT, while $\mathcal{E}_2(t)$ and $\mathcal{E}_3(t)$ are not as conservative early.

The $n_1^{st}$ observed event occurs at time $\tau$, measured in years. The interim analysis occurs at time $\tau^* \coloneqq \tau + \frac{2}{12}$, to account for the time interval between the locking of the dataset and the DSMB meeting. The dataset to be used for the primary analysis at the interim analysis is locked at time $\tau$. At time $\tau^* = \tau + \frac{2}{12}$, $n_2$ may be modified to some other value $n_2^*$.

### 3.2.2 Accrual

A maximum of $m$ patients are accrued into the RCT. Patients in both the experimental treatment and placebo groups have a uniform accrual time $E$ with distribution $\text{Unif}(0, a)$, for some fixed value $a$. At the time of adaptation $\tau^*$ it may be that accrual is stopped, as decided before the start of the RCT. If this is the case and if $\tau^* < a$, then patients whose accrual time is in $(\tau^*, a]$ are not enrolled into the RCT.

### 3.2.3 Time Until Primary Event

Once accrued into the RCT, the time until the primary event $T$ of a patient in the placebo [experimental] treatment group has an Exponential(rate = $\lambda_0$) [Exponential(rate = $\lambda_a$)] distribution. Under $H_0$, $\lambda_a$ is equal to $\lambda_0$ so that the hazard ratio is 1. Under $H_a$, $\lambda_a$ is chosen such that the original GSD has an 80% probability of rejecting $H_0$ using the log-rank test statistic with the GSD boundaries.

### 3.2.4 Time Until Surrogate Event

Suppose that $\varepsilon_i \overset{\text{indep}}{\sim} \mathcal{N}(0, \sigma^2)$, for $i = 1, 2, \ldots, m$, with $\sigma^2 \geq 0$. Once accrued into the RCT at time $E_i$, the time until the surrogate event $T_i^{surr}$ of patient $i$ is $T_i + \epsilon_i$, where

$$\epsilon_i \coloneqq \begin{cases} 0 & E_i + T_i \leq \tau^* \\ \varepsilon_i & E_i + T_i + \varepsilon_i \geq \tau^* \\ \tau^* - (E_i + T_i) + 0.0001 & \text{otherwise.} \end{cases}$$

This definition for $\epsilon_i$ relative to $\varepsilon_i$ is chosen to reflect the fact that at time $\tau^*$ the best prediction for $T_i$ is $T_i$ if $E_i + T_i \leq \tau^*$, and $T_i + \varepsilon_i$ otherwise. However, if $E_i + T_i + \varepsilon_i < \tau^* < E_i + T_i$, then $T_i + \varepsilon_i$ is very informative regarding $T_i$, but not useful as a substitute for $T_i$ for the purpose of predicting test statistics into the future. In such an instance, $\tau^* - E_i + 0.0001$ is used instead, as it is reasonably near the plausible value of $T_i$ closest to $T_i + \varepsilon_i$.

In most simulation settings, $\sigma^2 = 0$ so that $T_i^{surr}$ is always equal to $T_i$, but a few are considered where this is not the case.

### 3.2.5 Adaptive Procedure

The general adaptive procedure is to attempt to maximize the type I error conditional on the data observed by time $\tau^*$. The exact procedure depends on the data available at the time of adaptation.

$\mathcal{D}_1$ available

At time $\tau^*$, $Z_1$ is calculated based on the dataset locked at $\tau$, when the $n_1^{st}$ event was observed. If $Z_1 > b_1$, the RCT is stopped and $H_0$ is rejected. Otherwise, $n_2^* = n_2^*(Z_1)$ is chosen according to the observed value of $Z_1$. If $Z_1 > b_2$, then $n_2^* = (n_1 + 1) \vee n^{min}$, where $n^{min}$ is the minimum allowed value of $n_2^*$. Otherwise, $n_2^*$ is chosen to be the whole number that maximizes the probability of the test statistic being greater than $b_2$ given $Z_1$ under $H_0$, in a manner similar to that of Proschan & Hunsberger[11]. The value of $n_2^*$ is subject to the lower bounds of $n_1 + 1$ and $n^{min}$, as well as the upper bounds of $n^{max}$ and the total number of patients enrolled into the RCT, where $n^{max}$ is the maximum allowed value of $n_2^*$. Under most scenarios considered, $n^{max} = 4n_2$.

$\mathcal{D}_1$ and $\mathcal{D}_2$ available

At time $\tau^*$, $Z_1$ is calculated based on the dataset locked at $\tau$, when the $n_1^{st}$ event was observed, despite the fact that $n_{observed}$ events have been observed by time $\tau^*$. If $Z_1 > b_1$, the RCT is stopped and $H_0$ is rejected. Otherwise, the test statistic is calculated at every observed event time from the $(n^{min})^{st}$ observed event to the $n_{observed}^{th}$ observed event. If any of these test statistics are greater than $b_2$, the lowest number of events where this occurs is chosen to be $n_2^*$. Otherwise, $n_2^*$ is chosen to be the whole number that maximizes the probability of the test statistic being greater than $b_2$ given the test statistic calculated at $n_{observed}$ observed events under $H_0$, in a manner similar to that of Proschan & Hunsberger. The value of $n_2^*$ is subject to the lower bounds of $n_1 + 1$ and $n^{min}$, as well as the upper bounds of $n^{max}$ and the total number of patients enrolled into the RCT. Under most scenarios considered, $n^{max} = 4n_2$.

$\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ available

At time $\tau^*$, $Z_1$ is calculated based on the dataset locked at $\tau$, when the $n_1^{st}$ event was observed, despite the fact that $n_{enrolled}$ patients have been enrolled by time $\tau^*$. If $Z_1 > b_1$, the RCT is stopped and $H_0$ is rejected. Otherwise, the test statistic is calculated or predicted (with varying degrees of accuracy, depending on the prognostic ability of $\mathcal{D}_3$) from the $(n^{min})^{st}$ observed event to the $(n_{enrolled} \wedge n^{max})^{th}$ observed/predicted event. If any of these test statistics are greater than $b_2$, the lowest number of events where this occurs is chosen to be $n_2^*$. Otherwise, $n_2^*$ is chosen to be the whole number that maximizes the probability of the test statistic being greater than $b_2$ given the test statistic calculated at $(n_{enrolled} \wedge n^{max})^{th}$ observed/predicted events under $H_0$, in a manner similar to that of Proschan & Hunsberger. The value of $n_2^*$ is subject to the lower bounds of $n_1 + 1$ and $n^{min}$, as well as the upper bounds of $n^{max}$ and the total number of patients enrolled into the RCT. Under most scenarios considered, $n^{max} = 4n_2$.

### 3.2.6 CHW Adjustment

If the CHW adjustment is used on the predicted and/or observed test statistics, the number of events considered to have been observed at the time of adjustment is $n_1$, the number of events observed by the time of the data lock.

We investigate four combinations of the ways in which the CHW adjustment can possibly be applied, and they are listed below:

1. Adjustment not used on predicted or observed test statistics
2. Adjustment used on both predicted and observed test statistics
3. Adjustment used on the observed test statistic at the final analysis, but not on the predicted test statistics
4. Adjustment used on the predicted test statistics, but not on the observed test statistic at the final analysis

The first combination reflects a scenario where investigators are not aware of the existence of the CHW adjustment or do not bother to adjust for the fact that they modified their final sample size $n_2$ to $n_2^*$. The second combination can be a result of investigators putting forth a naive but good faith effort to control their type I error, not realizing that the CHW relies on the existence of the "independent increment structure". One plausible scenario that can result in the third combination is where investigators learn about needing to adjust for their adaptation for type I error control at some point after $\tau^*$, the time of adaptation. Of the four combinations, the last one is the least likely in a real world scenario. However, in an effort to determine how predicted statistic adjustments and observed statistic adjustments each contribute to type I error inflation, results reflecting this combination are also presented.

### 3.2.7 Metrics Considered

Type I error control is of utmost importance in a regulatory setting, so it is of interest for investigators to ensure that the type I error remains at the nominal level when making any adaptations during the course of the RCT. For this reason, the type I error rate is investigated across a range of scenarios.

Power is also an important metric to consider when designing a RCT. Because RCTs are so costly, they must be powerful enough to justify the cost of conducting the RCT in the first place.

Some settings considered in this chapter allow for both an inflation of the type I error rate and a change in power. This leads to some difficulty when comparing RCTdesigns that differ in both type I error rate and power. A metric that standardizes across the the type I error rate and power allows for the comparison of such RCT designs. The naive Bayes Factor, as described in Section 3.2.7.1 below, is one such standardization and is the third metric evaluated.

#### 3.2.7.1 Naive Bayes Factor

The Positive Predictive Value (PPV) is a function of the type I error, power, and prevalence. Specifically, PPV is defined in the following manner:

$$\text{PPV} \coloneqq \frac{\text{power} \times \text{prevalence}}{\text{power} \times \text{prevalence} + \text{type I error} \times (1 - \text{prevalence})}.$$

Using this definition, it is straightforward to find that the posterior odds, $\frac{\text{PPV}}{1-\text{PPV}}$, can be expressed in the following way:

$$\begin{array}{rcccl} \frac{\text{PPV}}{1-\text{PPV}} & = & \frac{\text{power}}{\text{type I error}} & \times & \frac{\text{prevalence}}{1-\text{prevalence}} \\ \text{posterior odds} & = & \text{naive Bayes Factor} & \times & \text{prior odds} \end{array}.$$

Within the context of a particular RCT design, this last equation has the following Bayesian interpretation: Under a simplistic binary prior on the binary decision space (either the treatment has no effect as under the null hypothesis or the treatment has an effect equal exactly to the alternative hypothesis), the naive Bayes Factor describes how prior beliefs of a beneficial treatment are modified to the posterior beliefs of the beneficial treatment. Therefore, a high value for the naive Bayes Factor increases the credibility of results from the RCT. We note that this Bayes Factor calculation is naive because the choice of a binary prior on the parameter space may not necessarily reflect a realistic setting where one might choose a prior that gives weight to a range of values for the parameter over some interval.

It is of interest to use the naive Bayes Factor as a metric along with power and the type I error, as it provides some measure of the loss of predictive value between designs varying in both size and power. For the remainder of this dissertation, the phrase "Bayes Factor" is used when referring to the naive Bayes Factor.

## 3.3 Results

Results with respect to the Type I Error, Power, and the Bayes Factor are presented. To understand clearly how different choices in the RCT design and methods of adaptation affect these metrics, results are first presented with respect to each choice.

Unless otherwise specified, throughout the following sections the default setting is the lung cancer example with an early interim analysis. Of the maximum 1900 patients to be accrued, under $H_0$ an average of 1064.244 patients are enrolled and an average of 274.541 events are observed by the time of adaptation, while under $H_a$ the average enrollment and number of events observed are 1104.230 and 273.703, respectively. $\mathcal{E}_1(\cdot)$ is used to calculate the stopping boundaries, and accrual is stopped after the time of the adaptation. In this default setting, $\mathcal{D}_1$ and $\mathcal{D}_2$ are known at the time of adaptation, and the CHW adjustment is not used for either the predicted test statistics or the observed test statistics.

Thick red lines in the following figures represent the nominal value of the metric being evaluated. For the type I error, power, and Bayes Factor figures, these values are 0.025, 0.8, and 32, respectively. In addition to this, thinner red lines are displayed on the type I error and power figures. The range of these lines reflect asymptotically valid Wald-based 95% prediction intervals of the simulated type I error and power levels, assuming the nominal values reflect the truth. For example, for a given scenario being evaluated, if the true type I error rate was 0.025, we would expect that 95% of the time the simulated type I error rates would lie within the prediction interval displayed in the type I error figure.

### 3.3.1 Lower Bound on $n_2^*$

Figure 3.1 displays the effect of the lower bound for $n_2^*$ on the type I error rate. Though many factors affect the shape of the type I error rate curve as a function of the lower

bound on $n_2^*$, the general trend is typically a decreasing one. The type I error rate is highest when $\tilde{n}_2^* := n_2^* - n_1$ is very close to 1.

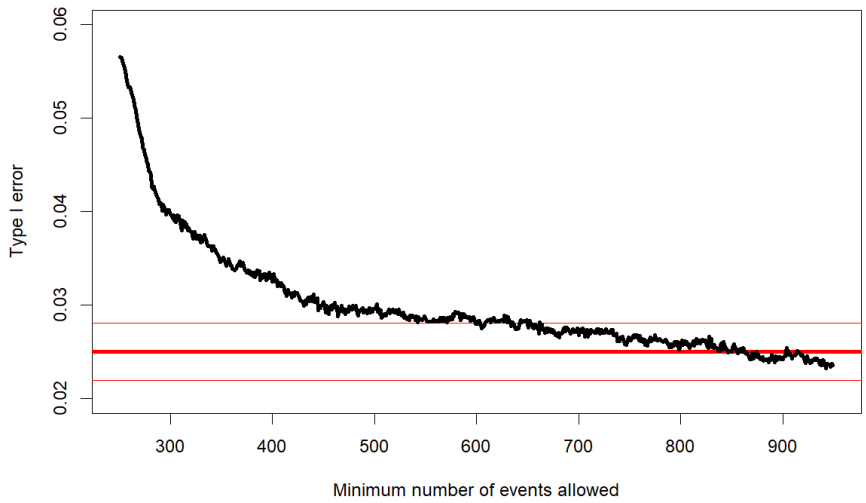



Figure 3.1: Example of type I error as a function of the lower bound on $n_2^*$. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

In Figure 3.1, the interim analysis occurs at 250 events. The type I error rate is highest when the lower bound on $n_2^*$ is between 251 and 300 events. The nominal type I error rate is 0.025, and in this figure the simulated rate drops sharply from 0.057 to levels close to the nominal value.

Under $H_a$ the trend can vary more widely depending on the scenario. It may be the case that as a function of the lower bound, the power is monotonically increasing, monotonically decreasing, or decreasing first and then increasing. However, when the lower bound on $n_2^*$ is very close to $n_2$, the power is typically close to the nominal level.

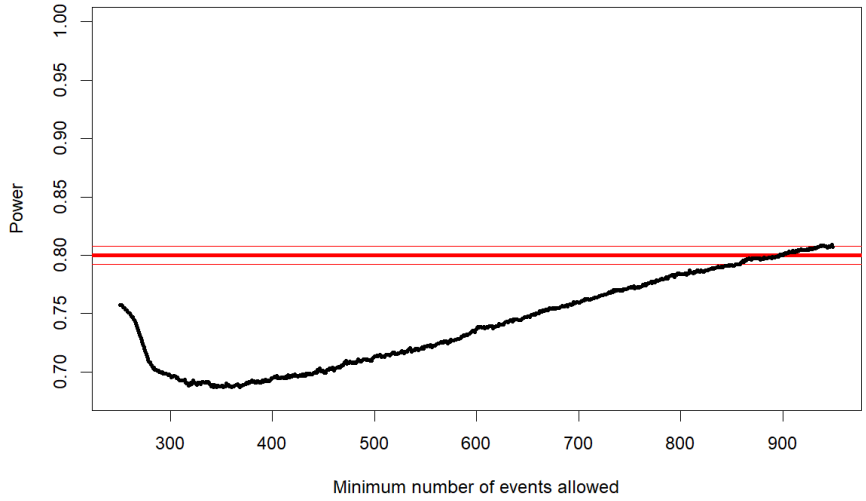



Figure 3.2: Example of power as a function of the lower bound on $n_2^*$. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

The setting in Figure 3.2 is the same as the one in Figure 3.1, except under $H_a$. In this particular example, the power decreases first, and then returns to levels close to the nominal value of 0.8.

The Bayes Factor as a function of the lower bound on $n_2^*$ is the most variable, as it completely depends on how the type I error and power behave as a function of the lower bound. In the same setting as in the previous two figures, the Bayes Factor generally increases to return to levels near the nominal value of 32, as can be seen in Figure 3.3.

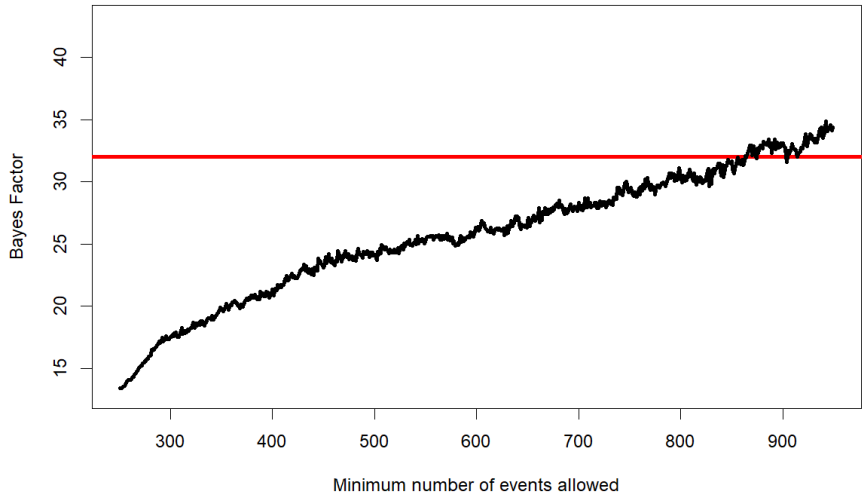



Figure 3.3: Example of the Bayes Factor as a function of the lower bound on $n_2^*$. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level.

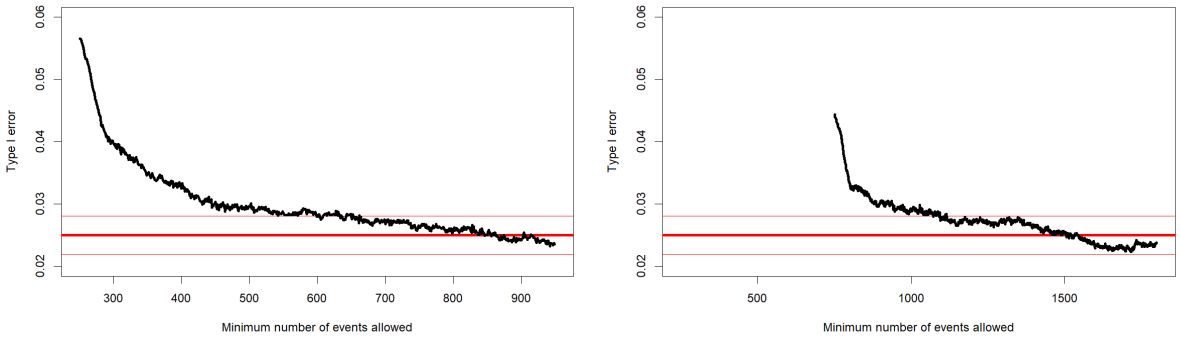
### 3.3.2 Timing of Interim Analysis

With an early interim analysis, the data is locked at the $250^{th}$ event. Under $H_0$, of the maximum 1900 patients to be accrued, the mean number of patients enrolled and the mean number of events observed at the time of adaptation are 1064.244 and 274.541, respectively, while under $H_a$ they are 1104.230 and 273.703, respectively.

Conversely, with a late interim analysis, the data is locked at the $750^{th}$ event. Under $H_0$, the mean number of patients enrolled and the mean number of events observed at the time of adaptation are 1898.444 and 785.634, respectively, while under $H_a$ they are 1899.995 and 782.864, respectively.

With a late interim analysis, the highest level of type I error inflation is not so severe. However, with a later analysis, the restriction on the lower bound of $n_2^*$ sometimes needs to be considerably greater than $n_2$ before the type I error is adequately controlled.

In Figure 3.4a, the maximum type I error rate is 0.057, whereas in Figure 3.4b it is only 0.044. With an early interim analysis, the type I error is reasonably close to the nominal

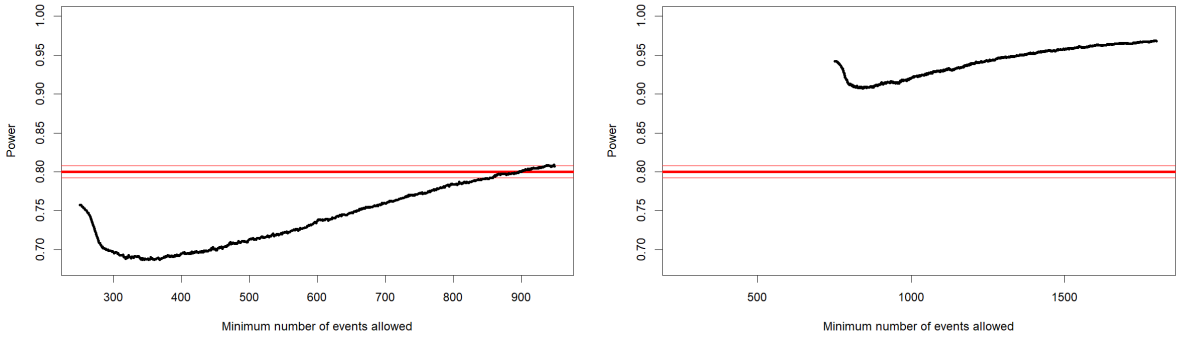



(a) Early interim analysis (b) Late interim analysis

Figure 3.4: Type I error as a function of the lower bound on $n_2^*$, depending on timing of interim analysis. Setting: Lung cancer RCT with conservative error spending early and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

rate if $n_2^*$ is no less than 700 events. However, with a late interim analysis, the type I error is not controlled unless $n_2^*$ has a lower bound of 1200 or greater.

The adaptive rule combined with a late interim analysis can result in levels of power well above the nominal level for a wide range of lower bounds on the value of $n_2^*$, as can be seen in Figure 3.5b.

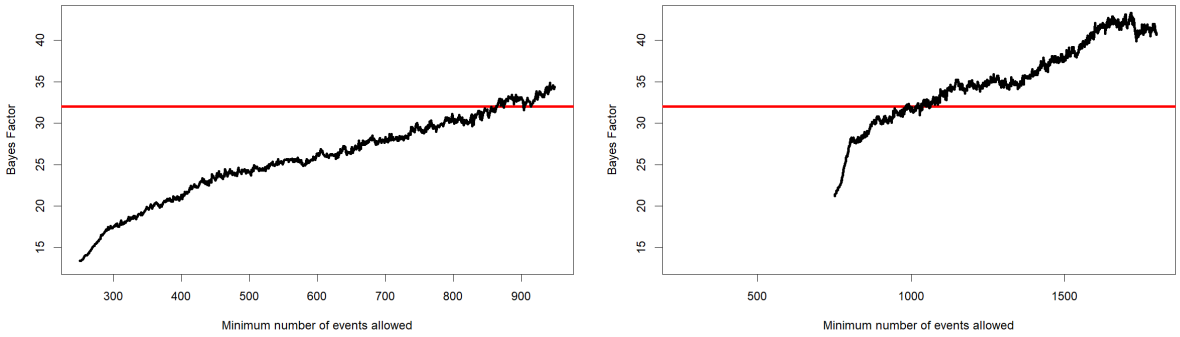



(a) Early interim analysis (b) Late interim analysis

Figure 3.5: Power as a function of the lower bound on $n_2^*$, depending on timing of interim analysis. Setting: Lung cancer RCT with conservative error spending early and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

With a late interim analysis, the type I error can be controlled with high enough values of the lower bound for $n_2^*$, and the power may be high enough with the adaptive rule

such that there are considerable gains with respect to the Bayes Factor. In the scenario considered thus far, a late interim analysis yields high values of the Bayes Factor for a range of the lower bound of $n_2^*$, as can be seen in Figure 3.6b.




(a) Early interim analysis

(b) Late interim analysis

Figure 3.6: Bayes Factor as a function of the lower bound on $n_2^*$, depending on timing of interim analysis. Setting: Lung cancer RCT with conservative error spending early and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level.

### 3.3.3 Knowledge of $\mathcal{D}_2$ Alone or with $\mathcal{D}_3$

Generally, the more information that is available with the adaptive rule considered, the higher type I error inflation is. Not surprisingly, the value of $\sigma^2$ is very influential, in cases where $\mathcal{D}_3$ is known at the time of adaptation.

In the same scenario being considered with an early interim analysis, the maximal type I error is 0.045 when only $\mathcal{D}_1$ is used, 0.057 when $\mathcal{D}_1$ and $\mathcal{D}_2$ are used, and up to 0.116 when $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are used, as shown in Figure 3.7. The additional knowledge of $\mathcal{D}_2$ has a noticeable but moderate effect on the type I error, but the type I error is significantly inflated with the knowledge of $\mathcal{D}_3$ when $\sigma^2 = 0$. When $\sigma^2 = 0.5$, the type I error rate with knowledge of $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ is comparable to that of when only $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. In this setting, the mean and variance of the survival times are $\frac{1}{0.19} \approx 5.263$ and $\frac{1}{0.19^2} \approx 27.701$, respectively, so in comparison the added noise with variance 0.5 is small in comparison. Even so, with just a bit of imprecision, the ability to use $\mathcal{D}_3$ to inflate the type I error is greatly diminished.

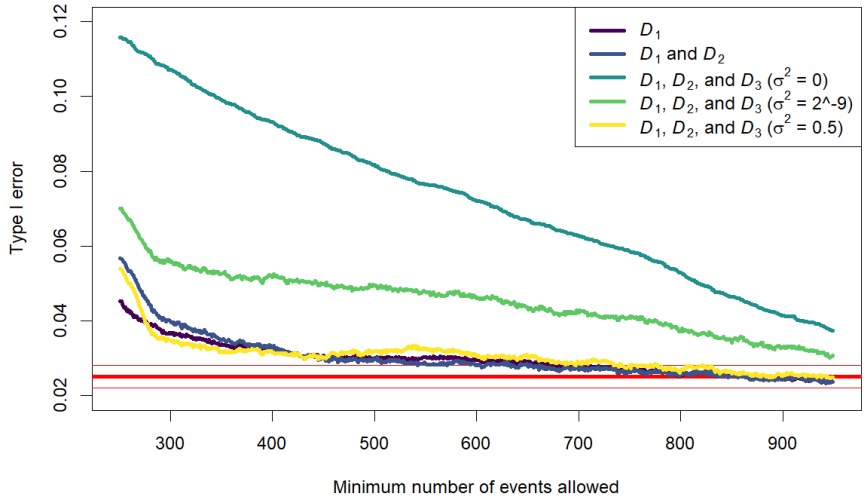



Figure 3.7: Type I error, depending on information available. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. The CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

The power curves are similar when comparing the use of $\mathcal{D}_1$ only with the use of $\mathcal{D}_1$ and $\mathcal{D}_2$. Interestingly, the use of $\mathcal{D}_1$ only results in higher power, but the difference in power shrinks as the lower bound on $n_2^*$ increases. The use of $\mathcal{D}_3$ significantly increases the power when $\sigma^2 = 0$, but is reduced when $\sigma^2 = 0.5$. In all four cases, increasing the lower bound on $n_2^*$ leads to the power approaching levels closer to the nominal level of 0.80.

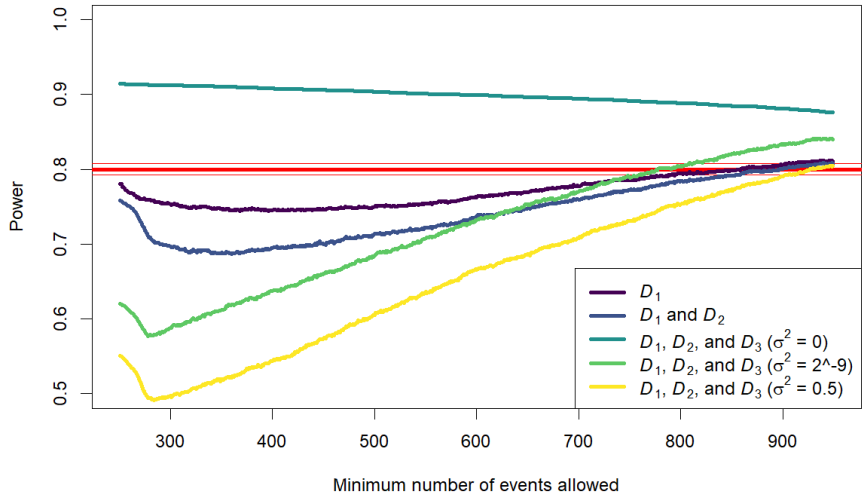



Figure 3.8: Power, depending on information available. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. The CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

More information used when adapting leads to lowered levels of the Bayes Factor, especially with small values for the lower bound on $n_2^*$. The increased power when using $\mathcal{D}_3$ is overshadowed by the type I error inflation when $\sigma^2 = 0$, resulting in values of the Bayes Factor lower than the nominal level, even with high values on the lower bound of $n_2^*$.

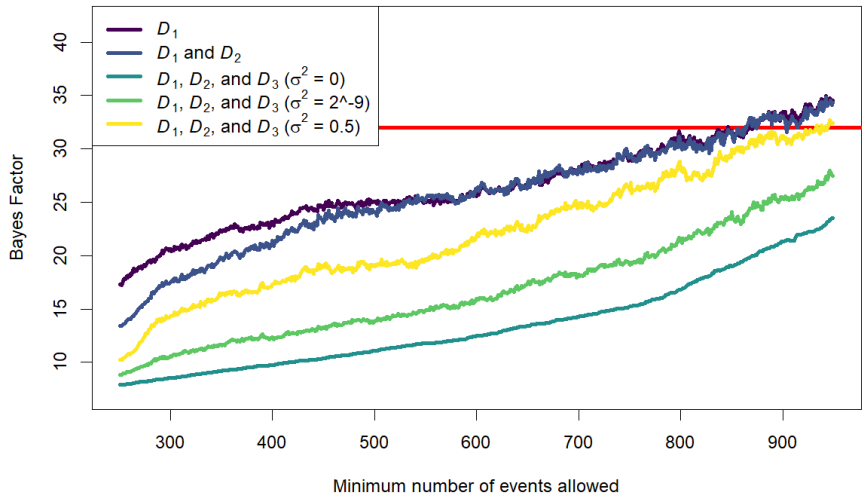



Figure 3.9: Bayes Factor, depending on information available. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. The CHW adjustment is not used. Thick red line is nominal level.

### 3.3.4 Error Spending Function

Error spending functions that are conservative early allow for greater type I error inflation. In the scenario considered in Figure 3.10, the maximal type I error is 0.057 when using $\mathcal{E}_1(\cdot)$, and 0.042 when using $\mathcal{E}_2(\cdot)$. However, with an early interim analysis, the type I error is controlled with a high enough lower bound on $n_2^*$.

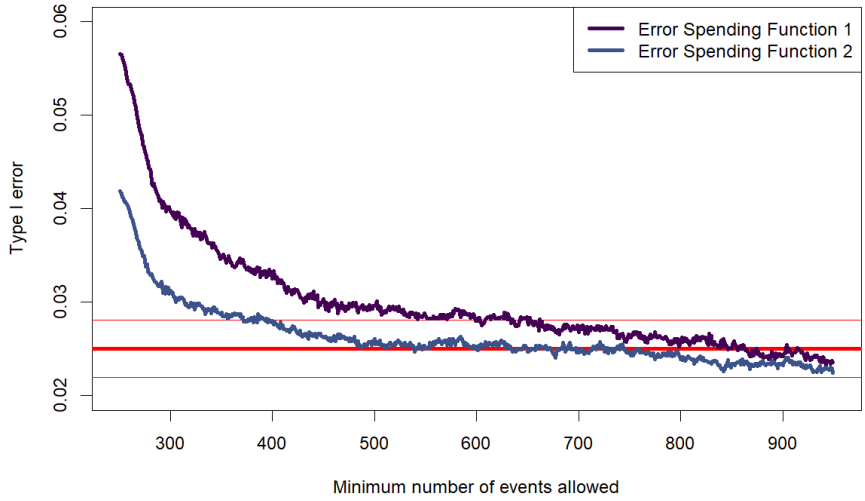



Figure 3.10: Type I error, depending on error spending function used. Setting: Lung cancer RCT with an early interim analysis and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

With an early interim analysis, the choice of the error spending function does not affect the power very much. As can be seen in Figure 3.11, the power curves are very similar.

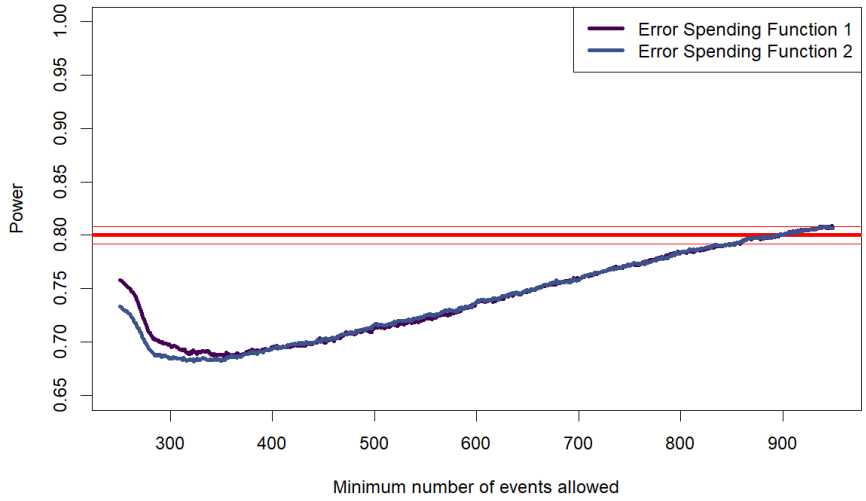



Figure 3.11: Power, depending on error spending function used. Setting: Lung cancer RCT with an early interim analysis and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

Because the use of an error spending function that is conservative early leads to higher levels of type I error inflation, levels of the Bayes Factor are decreased with the use of such a function. However, with a high enough lower bound on $n_2^*$, the Bayes Factor is close to its nominal value.

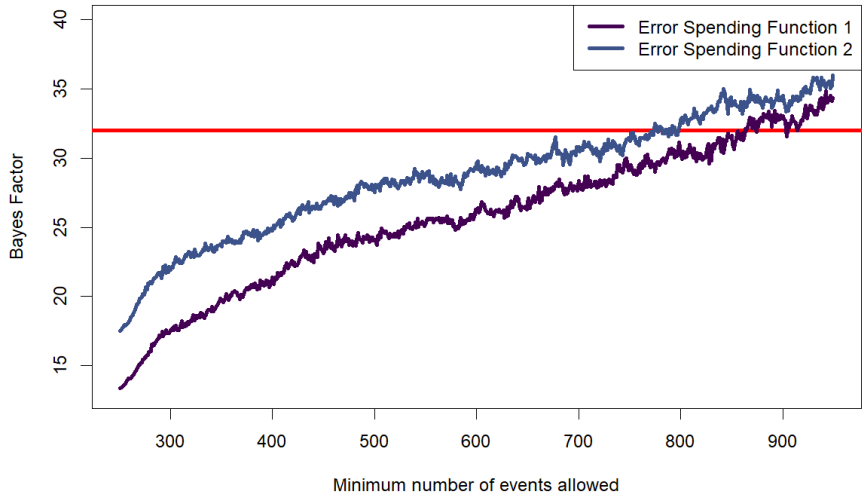



Figure 3.12: Bayes Factor, depending on error spending function used. Setting: Lung cancer RCT with an early interim analysis and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level.

### 3.3.5 Stopping Patient Accrual After Adaptation

Generally, if accrual is stopped after the time of adaptation, early portions of the survival curves and hence the predicted future values of the test statistic are more accurate. This is most obvious when $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are known. The following 3 figures are the results of such a scenario.

If accrual is continued after the adaptation, the maximal type I error is 0.058. Conversely, when the accrual is stopped, the maximal type I error is 0.116. With an early adaptation, the type I error is never adequately controlled with stopped accrual, even with high values for the lower bound of $n_2^*$.

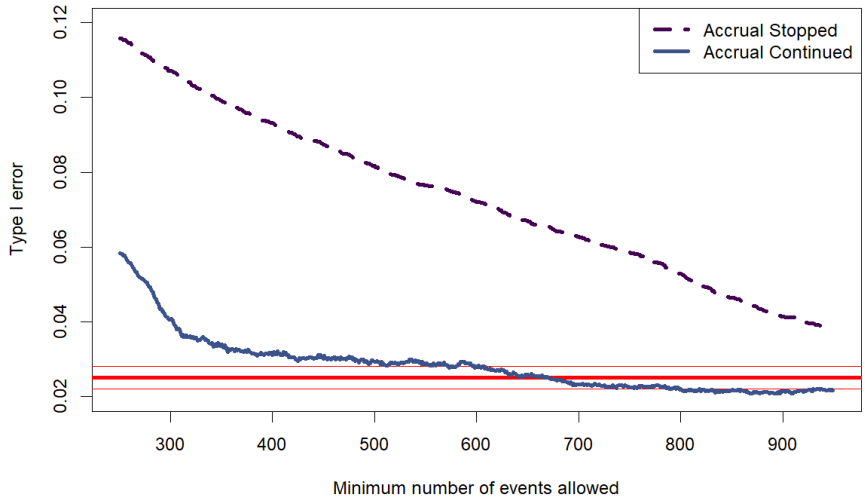



Figure 3.13: Type I error, depending on whether or not accrual continues after the time of adaptation. Setting: Lung cancer RCT with an early interim analysis and conservative error spending early. $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

Stopping accrual noticeably increases the power, as can be seen in Figure 3.14. With stopped accrual, the power is consistently above the nominal value, but decreases slightly as a function of the lower bound on $n_2^*$.

However, with continued accrual, the power has a noticeable decrease before increasing again. This decreasing trend is likely due to the fact that increasing the lower bound from

251 to about 300 observed events restricts adaptations so that investigators cannot adapt to times corresponding to the event times belonging to $\mathcal{D}_2$. The subsequent increasing trend can probably be attributed to the fact that under $H_a$, the expected value of the final test statistic is increasing in $n_2^*$.

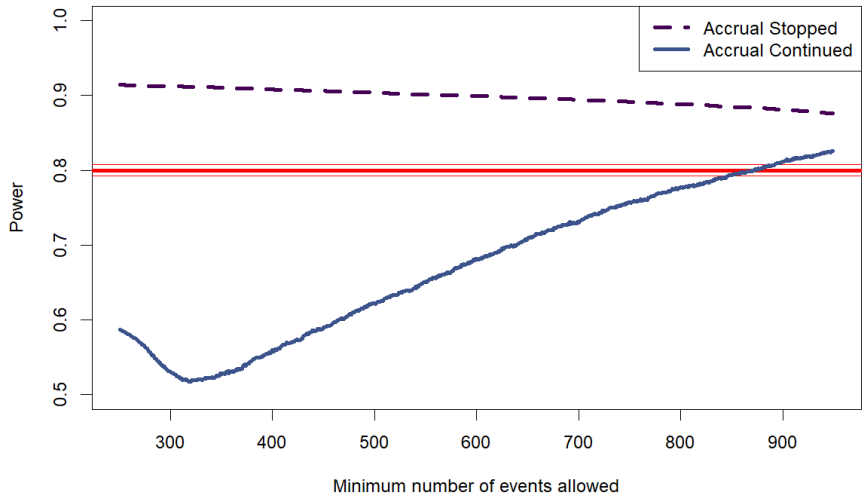



Figure 3.14: Power, depending on whether or not accrual continues after the time of adaptation. Setting: Lung cancer RCT with an early interim analysis and conservative error spending early. $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

Even with the increased power from stopping accrual, the type I error inflation drives the Bayes Factor downward. The Bayes Factor never achieves its nominal value when accrual is stopped.

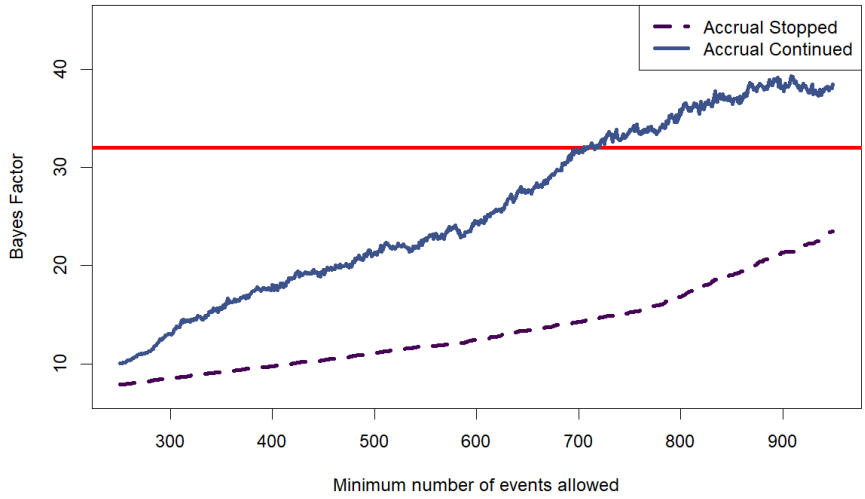



Figure 3.15: Bayes Factor, depending on whether or not accrual continues after the time of adaptation. Setting: Lung cancer RCT with an early interim analysis and conservative error spending early. $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are known, and the CHW adjustment is not used. Thick red line is nominal level.

### 3.3.6 Use of CHW Adjustment with Test Statistics

Here it is assumed that the CHW adjustment is used on the predicted test statistics at future events if and only if it is used on the observed test statistic at the final analysis.

With minimal restrictions on the value of $n_2^*$, using the CHW adjustment significantly increases the type I error. In the scenario considered in Figure 3.16, the maximal type I error is 0.057 when using unadjusted test statistics, but 0.121 when using CHW-adjusted test statistics. However, increasing the lower bound on $n_2^*$ sharply drops the type I error inflation when using CHW-adjusted test statistics, so that the type I error is actually higher when using unadjusted test statistics.

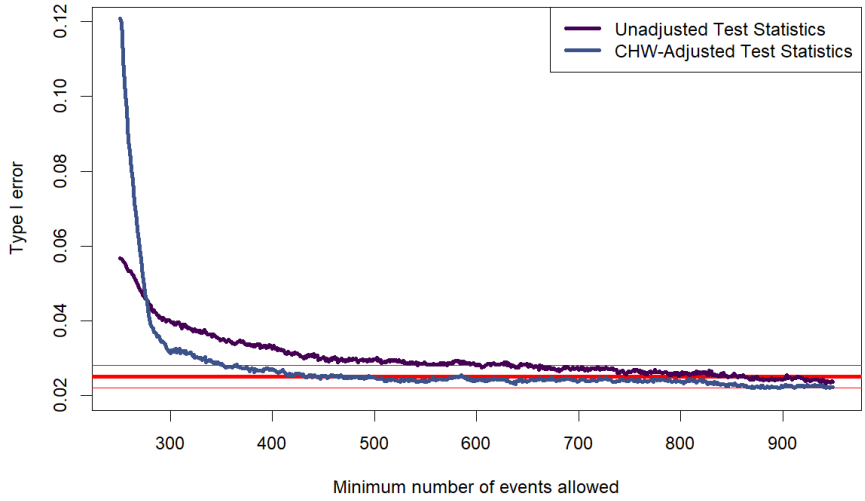



Figure 3.16: Type I error, depending on use of the CHW adjustment. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

The power curves are similar in both cases considered in Figure 3.17. However, the power is noticeably higher when using the CHW-adjusted test statistics if there are minimal restrictions on the value of $n_2^*$.

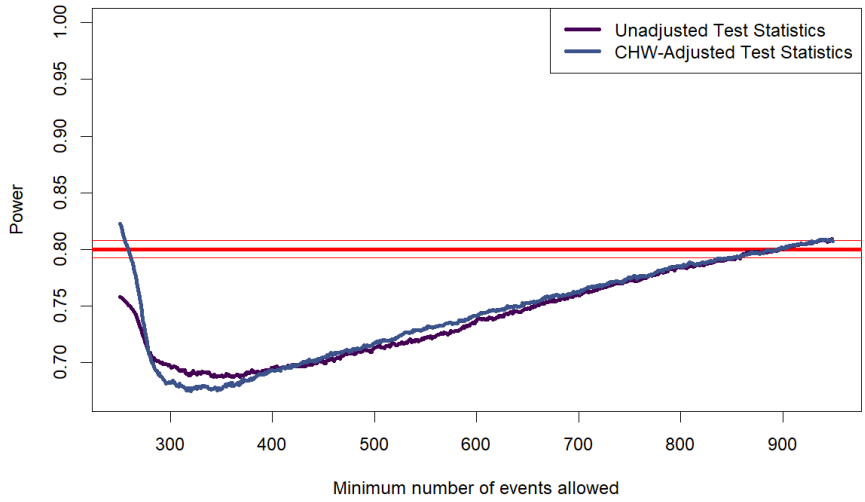



Figure 3.17: Power, depending on use of the CHW adjustment. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

With minimal restrictions on $n_2^*$, the Bayes Factor is lower when using CHW-adjusted statistics. Otherwise, the unadjusted test statistics lead to lower levels of the Bayes Factor.

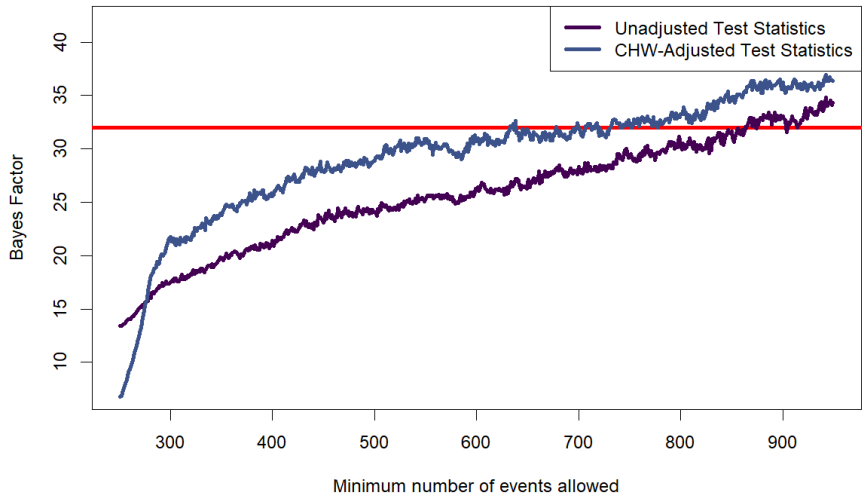



Figure 3.18: Bayes Factor, depending on use of the CHW adjustment. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. Thick red line is nominal level.

### 3.3.7 Concordance of Use of CHW Adjustment Between Predicted and Observed Test Statistics

The impact of the CHW adjustment comes from both using it on predicted test statistics at the time of adaptation and using it on the observed test statistic at the final analysis. To evaluate these individual effects, we examine all four combinations of CHW usage.

As can be seen in Figure 3.19, using the CHW adjustment on both or neither of predicted and observed test statistics leads to higher levels of type I error, compared to cases where it is used on one or the other, particularly when there are minimal restrictions on the value of $n_2^*$. However, the type I error levels in all four cases are similar when the lower bound is high enough.

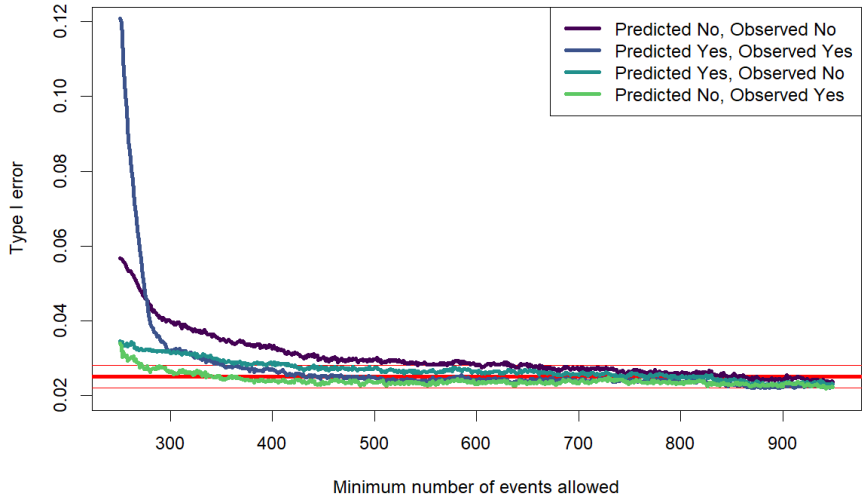



Figure 3.19: Type I error, depending on use of the CHW adjustment on predicted or observed test statistics. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

As with the type I error, the levels of power in the discordant cases are lower than the concordant cases when there are minimal restrictions on the value of $n_2^*$. However, with a high lower bound, the levels of power are similar across all four cases.

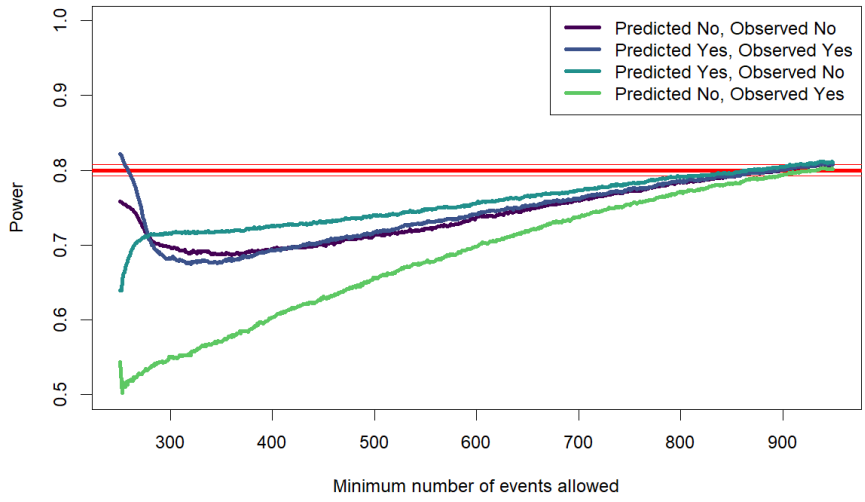



Figure 3.20: Power, depending on use of the CHW adjustment on predicted or observed test statistics. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

With minimal restrictions on the value of $n_2^*$, the discordant cases have slightly higher levels of the Bayes Factor. However, across most of the range of the lower bounds, the Bayes Factor levels are similar across the four cases.

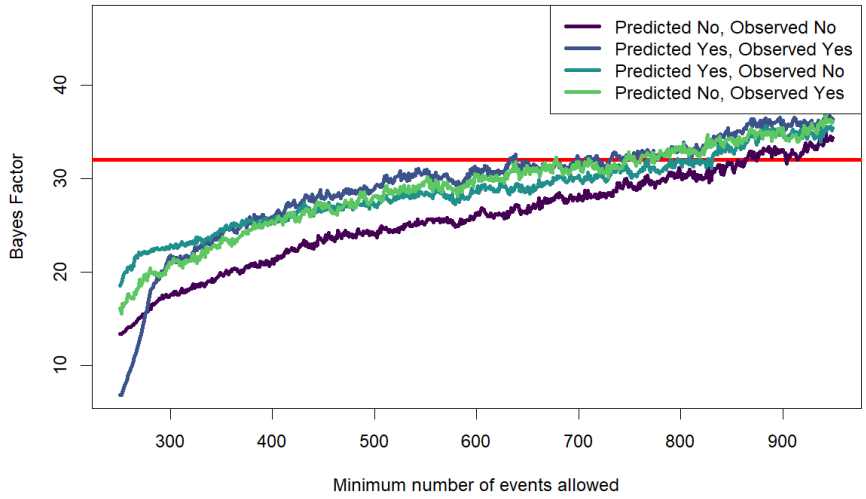



Figure 3.21: Bayes Factor, depending on use of the CHW adjustment on predicted or observed test statistics. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known. Thick red line is nominal level.

### 3.3.8 Upper Bound on $n_2^*$

Due to time constraints or limited availability of eligible patients, it is possible that it is unfeasible to increase the sample size at the time of adaptation. We thus consider a scenario where the adaptive rule has an upper bound for $n_2^*$ of $n_2$.

Compared to the case where $n_2^*$ is bounded above by the maximum number of patients accrued into the RCT, restricting $n_2^*$ to be less than or equal to $n_2$ slightly attenuates the type I error inflation. However, both cases have the same general trend with respect to the lower bound of $n_2^*$.

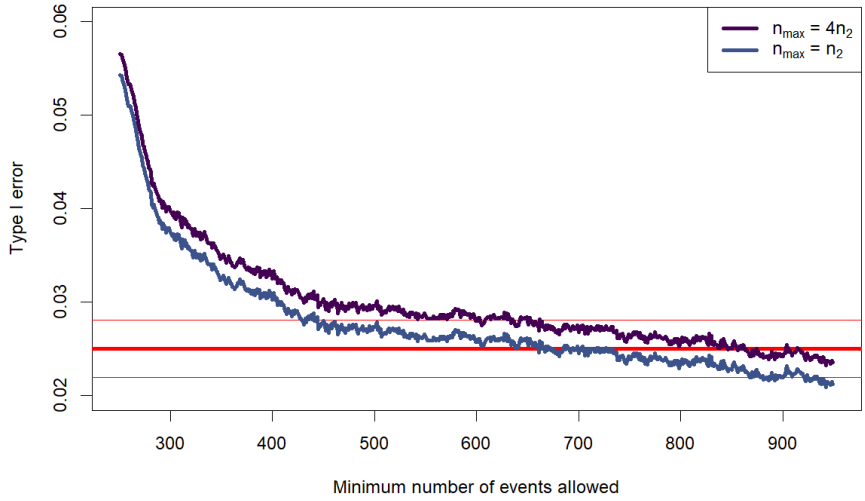



Figure 3.22: Type I error, depending on upper bound of $n_2^*$. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

As with the type I error, the power is slightly attenuated. However, while the type I error rates are very similar when there are minimal restrictions on the lower bound of $n_2^*$, there is a clear separation in the power curves throughout the entire range of the lower bounds of $n_2^*$.

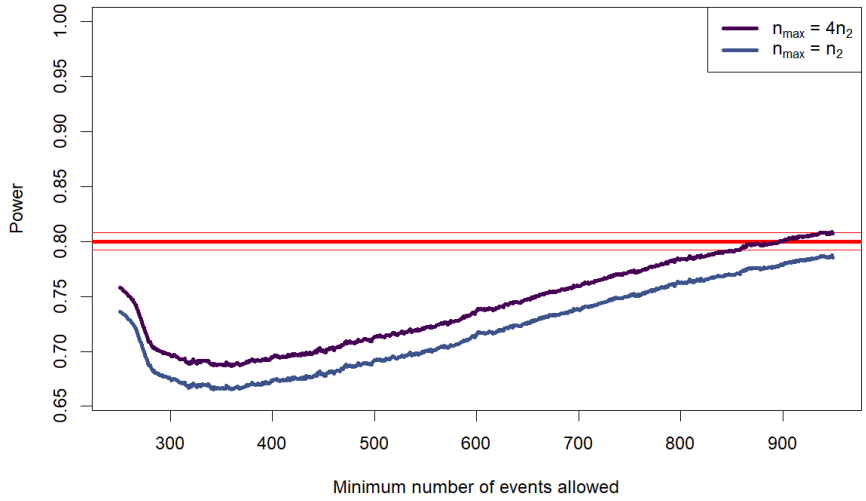



Figure 3.23: Power, depending on upper bound of $n_2^*$. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level, and thin red lines are 95% predicted range based on simulations.

The slightly lowered type I error rates when $n_2^*$ is at most $n_2$ contributes to the marginally improved Bayes Factor rates. The improvement is greater with higher values for the lower bound of $n_2^*$.

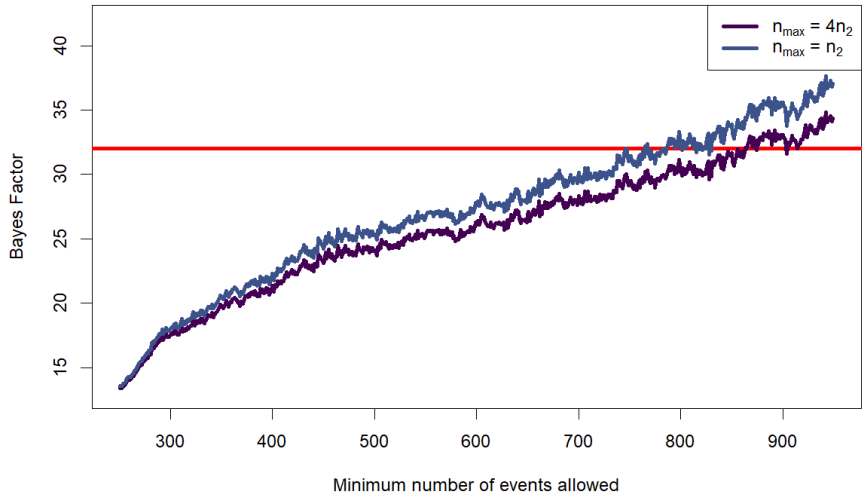



Figure 3.24: Bayes Factor, depending on upper bound of $n_2^*$. Setting: Lung cancer RCT with an early interim analysis, conservative error spending early, and stopped accrual. $\mathcal{D}_1$ and $\mathcal{D}_2$ are known, and the CHW adjustment is not used. Thick red line is nominal level.

## 3.4 Major Factors that Affect Type I Error

In a regulatory setting, strict type I error control is of utmost importance. The simulation results demonstrate that a variety of factors contribute to type I error inflation above the nominal desired level. These factors are discussed in more detail below.

### 3.4.1 Error Spending Function

Across all settings considered, the type I error is most inflated when the choice of the error spending function produces rejection boundaries that are conservative early. Because of the early conservatism, relatively little error is spent by the interim analysis, leading to the bulk of the error being spent at the final analysis. This leads to a rejection boundary that is closer to zero relative to a final rejection boundary that is produced by an error spending function that is not so conservative early. This lower final rejection boundary makes it more likely that one or more test statistics predicted into the future will be found to be statistically significant by chance under $H_0$.

### 3.4.2 Knowledge of $\mathcal{D}_2$ and $\mathcal{D}_3$

Having access to $\mathcal{D}_2$ alone or $\mathcal{D}_2$ and $\mathcal{D}_3$ allows for future values of the test statistic to be predicted for different values of $n_2^*$. These predictions allow for the possibility of choosing an optimal $n_2^*$ such that the test statistic will exceed the final rejection boundary with certainty or with high probability.

### 3.4.3 Concordance of CHW Adjustment Usage, Stopped Recruitment After Adaptation, and Accuracy of Predicted Event Times

Predicted values of the test statistic clearly are more accurate in reflecting what will be observed in the future when the CHW adjustment is used on the predicted test statistics if and only if the CHW adjustment is used on the final observed test statistic.

In a similar manner, the predicted values are more accurate if patient recruitment is stopped after the adaptation. This is because the test statistics predicted into the future are based on patients accrued thus far, and therefore reflect what will be observed in the future if no further patients are accrued.

Finally, predicted test statistics are useful to the extent that they accurately reflect what will be observed at the individual patient level. If the predicted event times are too inaccurate the predicted test statistics will not be accurate enough to be used in a way that can the inflate type I error.

### 3.4.4 Timing of Interim Analysis, CHW Adjustment, and Loose Restrictions on $n_2^*$

Type I error is consistently higher when $\frac{n_1}{n_2} = \frac{1}{4}$ compared to when $\frac{n_1}{n_2} = \frac{3}{4}$. This is partly due to the fact that an earlier interim analysis leads to less error being spent by said interim analysis, regardless of the error spending function that is used. However, an early interim analysis synergizes with the use of the CHW adjustment and loose restrictions on how low $n_2^*$ can be.

Recall that the form of the Z statistic after using the CHW adjustment is

$$Z_2^{CHW} = Z_1\sqrt{\frac{n_1}{n_2}} + \tilde{Z}_2^*\sqrt{\frac{\tilde{n}_2}{n_2}}.$$

When $\frac{n_1}{n_2} = \frac{1}{4}$, the bulk of the weight is given to the data observed after the adaptation, in summarized form $\tilde{Z}_2^*$. However, the path that $\tilde{Z}_2^*$ takes as a function of $\tilde{n}_2^*$ is very variable at low values of $\tilde{n}_2^*$. Therefore, when $\tilde{n}_2^*$ is allowed to be as low as 1, the probability that $Z_2^{CHW}$ will exceed the final rejection boundary at some point during its random walk under $H_0$ is moderate.

If test statistics can be predicted into the future with knowledge of $\mathcal{D}_2$ alone or knowledge of $\mathcal{D}_2$ and $\mathcal{D}_3$, the possibility of being able to predict a value of $Z_2^{CHW}$ that will exceed the final rejection boundary is greater with loose restrictions on how low $n_2^*$ can be, resulting in type I error rates as high as 0.205.

## 3.5 Major Factors that Affect Power

Investigators of expensive RCTs desire for the power to be close to the nominal level that was planned for or higher. Though the nature of the adaptive rule used in this section leads this rule to decrease the power of the RCT, the simulation results can be used to determine what factors might contribute to power attenuation. These factors are discussed in more detail below.

### 3.5.1 Loose Restrictions on Minimum Value of $n_2^*$

The adaptive rule that is used attempts to maximize the conditional type I error when none of the predicted test statistics exceed the rejection boundary at the final analysis. When attempting to maximize this conditional error, the rule often modifies the final sample size downward. This lowered sample size reduces the power of the final analysis. Higher values of the lower bound for $n_2^*$ prevent the final sample size from being so small that the power is reduced.

### 3.5.2 Late Interim Analysis

With a late interim analysis, patient accrual is done or almost done. Within the time between the data lock and the time of adaptation, a considerable number of additional events may be observed. The mean of the test statistic is increasing as a function of the number of events observed, so under $H_a$ is likely that the investigators can choose an optimal value for $n_2^*$ such that $H_0$ will likely be rejected at the final analysis.

### 3.5.3 Knowledge of $\mathcal{D}_2$ Alone or with Knowledge of $\mathcal{D}_3$

The more knowledge on events occurring after the data lock, the higher the number of predicted test statistics into the future that can be calculated. With more different values of $n_2^*$ at which the test statistic can be calculated, there is a higher chance of finding an optimal time to calculate the final test statistic to reject $H_0$.

### 3.5.4 Stopping Patient Accrual After Adaptation

As when under $H_0$, stopping patient accrual after the adaptation results in the predicted test statistics more accurately reflecting what will be observed in the future. As can be seen in Figure 3.14, for a given lower bound on the value of $n_2^*$, the power is higher when accrual is stopped, compared to when accrual is continued.

### 3.5.5 High Upper Bound on $n_2^*$

Under $H_a$, the mean of the final test statistic is an increasing function of $n_2^*$. Allowing $n_2^*$ to increase beyond $n_2$ allows for a final test statistic with a higher mean, increasing the RCT's power.

## 3.6 Conclusions

These simulation results demonstrate that it is crucial to correctly adjust for the surrogate information available at the time of adaptation, to ensure that at minimum the type I error rate is adequately controlled. Methods by Jenkins, Stone, & Jennison, Irle & Schäfer, and Magirr et al. allow for type I error control. However, as noted in Chapter 2, these approaches are either very restrictive with respect to what information can be used from patients accrued by the time of adaptation, or conservative. In the following chapter, we explore possible approaches to control the type I error in the presence of surrogate data without such limitations.

# Chapter 4

# Attempting to Control the Type I Error Rate

The results from Chapter 3 show that when surrogate data is used to adapt the final sample size, the type I error can be increased substantially if the final test statistic is not adjusted, and potentially more so if the wrong adjustment is used. In this chapter, we attempt to control the type I error with different approaches that do not require that data coming from patients accrued before and after the time of adaptation be summarized separately. Because type I error control and efficiency with respect to average sample size are the primary interest in this chapter, power and the Bayes Factor are not considered greatly.

## 4.1 Motivation

In the previous chapter, we demonstrated that when conducting adaptive RCTs, special care must be taken when surrogate data on outcomes are available on accrued patients who have not yet been observed to have an event. Failing to adjust for an adaptation can increase the type I error by substantial amounts, and incorrectly adjusting for the adaptation can increase it even further.

If an adaptive rule is a function of only $\mathcal{D}_1$ at the interim analysis, readily available analysis approaches such as the CHW adjustment control the type I error rate. However, such approaches that assume knowledge of only $\mathcal{D}_1$ are not appropriate when $\mathcal{D}_2$ and/or $\mathcal{D}_3$ are also used in the adaptive decision.

In instances where surrogate information may be used to adapt, a few approaches have been proposed in the literature that will control the type I error: Jenkins, Stone, & Jennison, Irle & Schäfer, and Magirr et al. However, the first and second approaches each require that data from the patients accrued into the RCT by the time of adaptation be censored at the time of the originally planned final analysis. The third method does not have such a restriction, but instead utilizes a final rejection boundary that is conservative, which can lead to type I error rates below the nominal value.

For each of these three approaches, the test at the final analysis is some mathematical equivalent of the test

$$w_1 Z_2^A + w_2 Z_2^B > b,$$

where $w_1$ and $w_2$ are prespecified weights such that $w_1^2 + w_2^2 = 1$, $Z_2^A$ and $Z_2^B$ are test statistics calculated using data from patients accrued into the RCT before and after the time of adaptation, respectively, and $b$ is some rejection boundary. In this chapter, we investigate whether it is possible to control the type I error while avoiding such a form of the test.

Settings 1 and 2 consider analysis procedures that take a similar form,

$$v_1 Z_2^{before} + v_2 Z_2^{after} > b,$$

where $v_1^2$ is an estimate of the information fraction available at the time of adaptation, $v_2^2 \coloneqq 1 - v_1^2$, and $Z_2^{before}$ and $Z_2^{after}$ are test statistics calculated from data collected before and after some time point $\tau$. In Setting 1, we explore the effect of choosing the information fraction to be different values that are not necessarily correct, and in Setting 2 we investigate as to whether or not the type I error is controlled when the information

fraction used in the adjustment accurately reflects the amount of statistical information available at the time of adaptation.

In Setting 3, we investigate whether the type I error can be controlled with an approach in which the conditional type I error at the interim analysis is controlled by conditioning on both the interim test statistic and predictions of test statistics into the future.

## 4.2 Setting 1

This setting considers a scenario where surrogate data may be used to predict test statistics into the future, but if test statistics can be predicted at certain sample sizes up to $n_c$, then the final sample size must be at least $n_c+1$. The results from Chapter 3 suggest that allowing for adaptations to $n_c$ or below can lead to type I error inflation, so the choice of a lower bound of $n_c + 1$ allows for the greatest flexibility while restricting adaptations to sample sizes where the test statistic cannot be predicted using surrogate data. We explore different ways available information can be used to adapt, as well as different ways of partitioning the data to apply the CHW adjustment.

### 4.2.1 Notation

In this chapter, we ultimately simulate test statistics rather than individual survival data, so we use a normal model to make our simulations more efficient.

A GSD is planned before the start of the RCT, such that the planned interim and final sample sizes are 500 and 1000, respectively. However, after an adaptation, the final sample size may differ from the originally planned value. O'Brien-Fleming boundaries are used, and the one-sided type I error and power are 0.025 and 0.80, respectively. Individual outcomes have a $\mathcal{N}(\theta, 1)$ distribution. Under $H_0$, $\theta = 0$, while under $H_a$, $\theta = \theta_a$, where $\theta_a$ is chosen so that the GSD has the appropriate level of power. At the final analysis, the efficacy boundary is $b$.

$Z_{[n_a:n_b]}$ is defined to be the $Z$ statistic calculated using the outcomes from patients $n_a$ through $n_b$, so that the interim test statistic $Z_1$ is equivalent to $Z_{[1:500]}$. At the interim analysis, test statistics $Z_{[1:501]}$ through $Z_{[1:750]}$ are able to be predicted, perhaps imperfectly. The final sample size can be decreased, or be increased up to a maximum of 2000.

The following is the general adaptation procedure, assuming $Z_1$ is in the continuation region of the GSD's interim analysis:

1. Determine the maximal sample size $n_c$ such that $Z_{[1:n_c]}$ is known or can be predicted for the purpose of adapting the final sample size. $n_c+1$ is the minimum final sample size, unless otherwise noted.
2. If $Z_{[1:n_c]} \leq 0$, modify the final sample size to 2000.
3. If $Z_{[1:n_c]} > b$, modify the final sample size to $n_c + 1$.
4. If $Z_{[1:n_c]} \in (0, b]$, modify the final sample size to the value that maximizes the conditional probability of the unadjusted final test statistic exceeding $b$, under $H_0$, in a manner similar to that of Proschan & Hunsberger.
5. Perform some CHW adjustment on the test statistic at the final analysis.

A number of different scenarios are considered, varying what can be used for adaptation, and where the data is partitioned for the CHW adjustment. They are listed below, noting that CHW($n_p$) indicates that the CHW adjustment is applied assuming that the first $n_p$ outcomes are observed by the time of the adaptation.

- Scenario 1: $Z_1$ is used to adapt $n_2$ to $n_2^* = n_2^*\big(Z_{[1:500]}\big)$, with CHW(500) and minimal final sample size $n_2^* \geq 501$. This scenario leads to a correct analysis, since the partitioning of the data results in the adaptation decision being independent of the second partition.
- Scenario 2: $Z_{[1:750]}$ is used to adapt $n_2$ to $n_2^* = n_2^*\big(Z_{[1:750]}\big)$, with CHW(500) and minimal final sample size $n_2^* \geq 501$. This scenario leads to an incorrect analysis,

since outcomes 501 through 750 are used to make the adaptation decision, and therefore the partitioning of the data results in the adaptation decision being a function of the second partition.

- Scenario 3: $Z_{[1:750]}$ is used to adapt $n_2$ to $n_2^* = n_2^*\big(Z_{[1:750]}\big)$, with CHW(750) and minimal final sample size $n_2^* \geq 751$. Though the interim analysis occurs at some time before the value of $n_p$ at which the data is partitioned, this scenario leads to a correct analysis, similar to Scenario 1. This scenario is considered to investigate the differences in operating characteristics between this scenario and Scenario 1.

- Scenario 4: $\hat{Z}_{[1:750]}$ is used to adapt $n_2$ to $n_2^* = n_2^*\Big(\hat{Z}_{[1:750]}\Big)$, with CHW(750) and minimal final sample size $n_2^* \geq 751$. As with Scenarios 1 and 3, this leads to a correct analysis. However, this scenario is included to investigate whether there are any costs associated with the predicted test statistics being inaccurate.

- Scenario 5: $Z_1$ is used to adapt $n_2$ to $n_2^* = n_2^*\big(Z_{[1:500]}\big)$, with CHW(750) and minimal final sample size $n_2^* \geq 751$. This scenario leads to a correct analysis as well. However, it is of interest to determine whether there are consequences to adjusting for data not used in the adaptation rule.

- Scenario 6: $Z_1$ is used to adapt $n_2$ to $n_2^* = n_2^*\big(Z_{[1:500]}\big)$, with CHW(500) and minimal final sample size $n_2^* \geq 751$. This scenario leads to a correct analysis. This scenario very similar to Scenario 1, but the adaptation decisions are the same as in Scenario 5.

In Scenario 4, noise is added to future individual outcomes, so that $\hat{Z}_{[1:750]}$ predicts $Z_{[1:750]}$ imperfectly. The individual pieces of noise are i.i.d. $\mathcal{N}(0, \sigma^2)$. Scenarios 4a through 4e correspond to $\sigma$ being 0.1, 0.5, 1, 5, and 10, respectively.

### 4.2.2 Results

Simulation results are presented in Table 4.1. For each adaptive design, a corresponding GSD is found, with the same interim analysis, same continuation region, and same type

I error rate and power. The average sample numbers (ASNs) are presented for both the adaptive designs and GSDs, under $H_0$ and $H_a$.

| Scenario | Type I Error | Power | ASN, $H_0$ | GSD ASN, $H_0$ | ASN, $H_a$ | GSD ASN, $H_a$ |
|---|---|---|---|---|---|---|
| 1 | 0.025 | 0.662 | 1843.587 | 715.421 | 880.312 | 671.151 |
| 2 | 0.036 | 0.888 | 1906.374 | 1152.314 | 922.510 | 1018.262 |
| 3 | 0.026 | 0.762 | 1906.374 | 894.758 | 922.510 | 813.635 |
| 4a | 0.025 | 0.763 | 1906.323 | 908.882 | 923.420 | 824.856 |
| 4b | 0.024 | 0.781 | 1905.678 | 963.144 | 950.088 | 867.967 |
| 4c | 0.026 | 0.802 | 1905.206 | 989.992 | 1005.704 | 889.297 |
| 4d | 0.024 | 0.850 | 1904.212 | 1156.833 | 1177.035 | 1021.852 |
| 4e | 0.025 | 0.859 | 1903.985 | 1173.235 | 1211.934 | 1034.884 |
| 5 | 0.025 | 0.768 | 1852.751 | 920.132 | 974.817 | 833.794 |
| 6 | 0.025 | 0.828 | 1852.751 | 1074.267 | 974.817 | 956.254 |

Table 4.1: Operating characteristics of the different scenarios. In addition, ASNs are provided for comparable GSDs.

Type I error is controlled across a majority of the scenarios considered. However, under $H_0$ the GSD with the same type I error, power, and interim analysis corresponding to each scenario is more efficient with respect to the ASN. In fact, in a majority of the scenarios, the GSD's ASN is less than half of the adaptive design's ASN, under $H_0$. Under $H_a$, the only scenario where the adaptive design's ASN is lower than that of the GSD is the one where the type I error is not adequately controlled. While the adaptive rule was chosen not to be efficient but to attempt to inflate the type I error, the results make it clear that if one is not allowed to adapt to a sample size where the test statistic can be predicted, attempting to inflate the type I error is a futile exercise if the CHW adjustment is applied in an adequate manner, and results in a very inefficient design.

Compared to Scenario 1, Scenario 3 is restricted to a later adaptation in that investigators are committed to giving the first 750 outcomes the same amount of weight at the final analysis as under the original RCT design. Because Scenario 3 has a stricter lower bound on the final sample size than Scenario 1, the ASN under both $H_0$ and $H_a$ is higher. As expected, the higher ASN under $H_a$ results in a higher level of power for Scenario 3. For each of these two scenarios, the corresponding GSD is more efficient with respect to the ASN.

Similar trends are observed when comparing Scenarios 1 and 6. These two scenarios are equivalent, except that in Scenario 1 the final sample size is bounded below by 501, whereas in Scenario 6 the lower bound is 751. The stricter lower bound in Scenario 6 leads to a higher ASN under $H_a$ and results in a higher level of power, compared to Scenario 1. Notably, the stricter lower bound of 751 slightly increases the power of the originally planned GSD, while the lax lower bound of 501 decreases the power considerably.

Comparing Scenarios 1 and 5 gives some insight regarding the cost of adjusting for data that is not used in the adaptation. While $Z_{[1:500]}$ is used in the adaptive rule for both scenarios, the CHW adjustment in Scenario 1 assumes that 500 outcomes have been observed by the time of the adaptation, while in Scenario 5 it assumes that 750 outcomes have been observed. The design in Scenario 5 has a noticeably higher level of power than Scenario 1. At a glance, it appears that adjusting for statistical information that is not used in the adaptive rule can actually increase the power of the design. However, it is not clear whether this is due to the different assumptions made when applying the CHW adjustment, or due to the different lower bounds in the final sample size between the two scenarios.

To address this lack of clarity, we compare Scenarios 5 and 6, as they have the same adaptation rule, but differ only in the application of the CHW adjustment. The power of Scenarios 5 and 6 under $H_a$ are 0.768 and 0.828, respectively. The lower power for Scenario 5 clarifies that contrary to what the previous comparison suggested, a penalty is paid when adjusting for statistical information that is not used in the adaptative rule.

Scenarios 4a through 4e demonstrate that the accuracy of the predicted test statistics does not affect one's ability to control the type I error, as long as the data is partitioned in an adequate manner. Across all of these scenarios, the simulated type I error is at or near the nominal level. Under $H_a$, the ASN of Scenario 4 is increasing in $\sigma$. This implies that adaptive rules based on imprecise surrogate data can result in inefficient designs.

Scenarios 5 and 6 are comparable because the adaptive rule is the same in both, and the analysis differs only in the choice of where the data is partitioned. The type I error is

controlled in both instances, and the ASNs are the same. However, partitioning the data at a higher value of the sample size results in lower power. This loss in power suggests that adjusting for statistical information that is not being used can result in non-negligible losses in statistical efficiency, as such an adjustment results in more extreme imbalances in the weights of the observations being used in the calculation of the final test statistic.

Scenario 2 is the only scenario in which the type I error is not adequately controlled. This is also the only scenario where $Z_{[1:n_1]}$ is used for the adaptation, the data is partitioned at $n_2$, and $n_1 > n_2$. This suggests that $n_1$ must not be so low relative to $n_2$. However, it is worth investigating whether or not $n_1$ needs to be at most $n_2$.

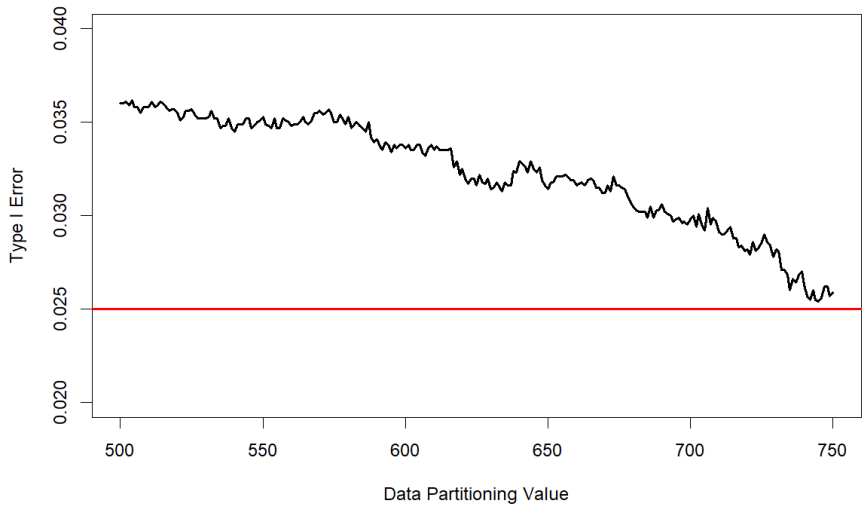



Figure 4.1: The effect of the choice of the data partitioning on the type I error.

Scenarios 2 and 3 represent two extremes with respect to the choice of data partitioning. We explore choices of data partitioning between the sample sizes of 500 and 750. Figure 4.1 displays the effect of the choice of data partitioning on the type I error, when $n_1 > n_2$. While higher values of the sample size at which the data is partitioned leads to lower type I error levels, the type I error is not controlled until the data is partitioned at values of the sample size of around 745 or greater.

In instances where future test statistics can be predicted exactly at sample sizes up to $n$, the data must be partitioned at $n$ or later, in order for the type I error to be adequately

controlled. The question remains as to whether this also holds for cases where the test statistic can be predicted imperfectly.

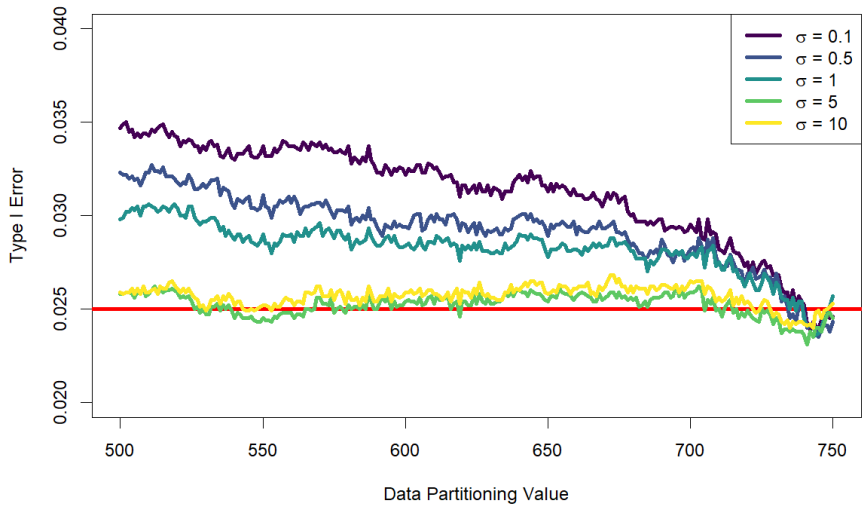



Figure 4.2: The effect of the choice of the data partitioning on the type I error, across a range of values of $\sigma$.

From Figure 4.2, it is clear that as long as there is some predictive power of the surrogate data for future test statistics up to sample size $n$, the data must be partitioned at $n$ or later. Only when the predictions are completely worthless is it acceptable to partition the data at sample sizes below $n$, as when $\sigma \geq 5$ in this example.

## 4.3 Setting 2

The results in Table 4.1 and Figures 4.1 and 4.2 pertaining to Scenarios 2, 3, and 4 in Section 4.2 suggest that partitioning the data incorrectly can lead to type I error inflation. This setting aims to explore this issue further. The Irle & Schäfer approach requires that the data be partitioned such that the weights used to weigh the test statistics reflect the statistical information being contributed to the overall test statistic from each of the partitions, under the original study design. In addition, the Irle & Schäfer approach requires that the partitioning of the data results in the incremental test statistics reflecting two independent groups of patients, such that no information on one of those groups has

been collected at the time of the adaptation.

In this setting, we aim to determine whether the type I error can be controlled by having the weights accurately reflect the information available at the time of the interim analysis, without the incremental test statistics reflecting two independent groups of patients.

### 4.3.1 Notation

As in Section 4.2, a GSD is planned before the start of the RCT, such that the interim and final sample sizes are $n_1 = 500$ and $n_2 = 1000$, respectively. O'Brien-Fleming boundaries are used, and the one-sided type I error and power are 0.025 and 0.80, respectively. Individual outcomes have a $\mathcal{N}(0, 1)$ distribution, and the variance is known to investigators. At the final analysis, the efficacy boundary is $b$.

At the interim analysis, in addition to the interim test statistic $Z_1$, investigators have perfect knowledge of a subset of the values of future outcomes to be observed, should the RCT continue to the final analysis. The probability of the value of future observation $i$ being available to investigators at the time of adaptation depends on the number $i$. For example, suppose that each of the future outcomes 1 through $n_{future}$ beyond the initial $n_1$ may be known to investigators at the interim analysis. The probability that future event $i$ is known is $p(i) = 1 - \left(\frac{i-1}{n_{future}+1}\right)^a$. The following figure displays this function for different values of $a$, when $n_{future} = 250$.

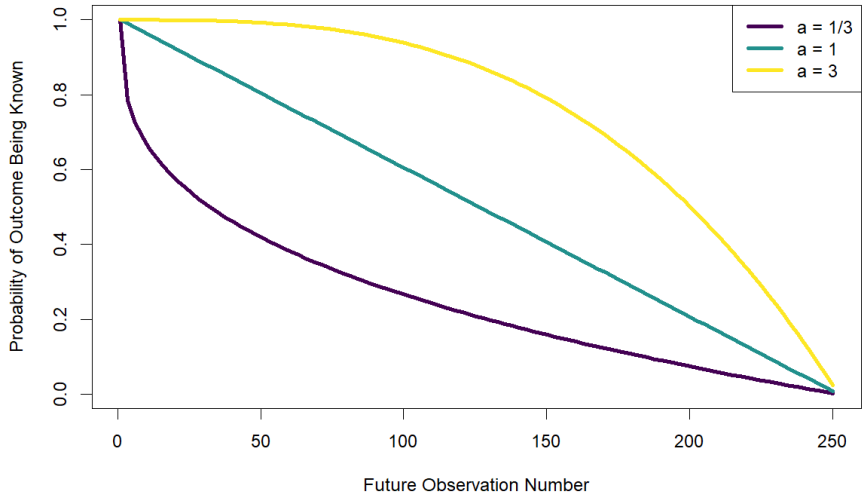



Figure 4.3: The probability of a future observation being known to the investigators at the interim analysis, as a function of the observation number.

The total number of outcomes known to investigators is $n_{total} \leq n_1 + n_{future}$. In this scenario, investigators are aware of which outcomes they know of, and of which outcomes they do not. For example, if future outcomes 1 and 2 are unknown but future outcome 3 is known, investigators can calculate a test statistic using the $n_1$ observed events and the third future event, and they may suspect that this statistic is correlated with the statistic that will be calculable immediately after observing the first three future events. In this way, investigators can estimate future test statistics at certain sample sizes, by using the observed outcomes and the subset of future outcomes that are known. However, the statistical information available as a fraction of the total information to be observed under the original GSD is $\frac{n_{total}}{n_2}$.

The adaptive rule considered is the following:

- Among all the estimates for future test statistics, determine the maximum value.
- If this value exceeds final efficacy boundary $b$, adapt to the outcome number corresponding to this estimated statistic.
- Otherwise, proceed with the original GSD.

At the final analysis, the data is partitioned according to the information fraction $\frac{n_{total}}{n_2}$,

and the CHW adjustment is applied with weights $\sqrt{\frac{n_{total}}{n_2}}$ and $\sqrt{1-\frac{n_{total}}{n_2}}$. This adjusted statistic is then compared to final efficacy boundary $b$, and $H_0$ is or is not rejected, accordingly.

In the simulation study of size 100,000, $a \in \left\{\frac{1}{3}, 1, 3\right\}$, and $n_{future} \in \{25, 50, 75, \ldots, 475\}$.

### 4.3.2 Results

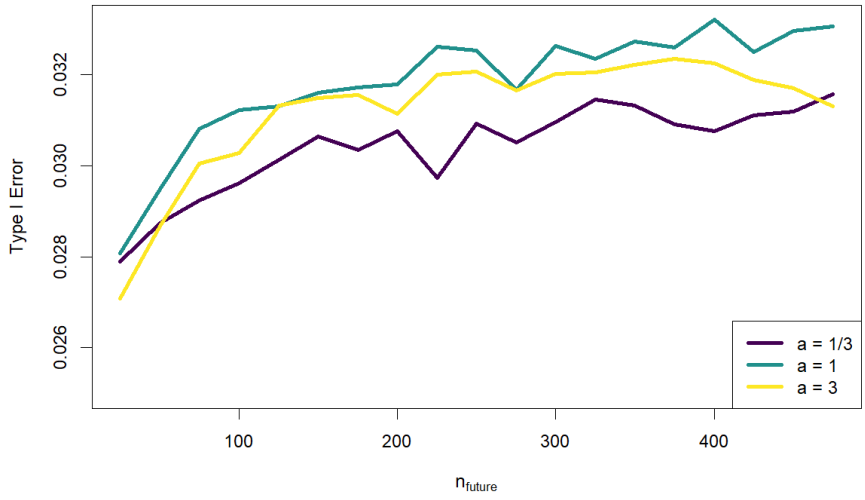



Figure 4.4: Type I error, as a function of $n_{future}$.

Figure 4.4 shows that the type I error can be slightly inflated, despite the fact that the weights used in the CHW adjustment reflect the information fraction that is available at the time of the adaptation. The type I error appears to be increasing in $n_{future}$, despite the fact that higher values of $n_{future}$ lead to a higher information fraction at $\frac{n_{total}}{n_2}$.

The information fraction at the interim analysis can shed some light into this trend. If $n_{future} = 475$ and $a = 3$, for example, the mean information fraction is 0.857, as can be seen in Figure 4.5. This means that of the 500 future observations, 357 of them are known by the investigator, on average. This high value leads a large number of future test statistics that can be estimated as well as accurate estimates, since only 143 future observations are unknown, on average.

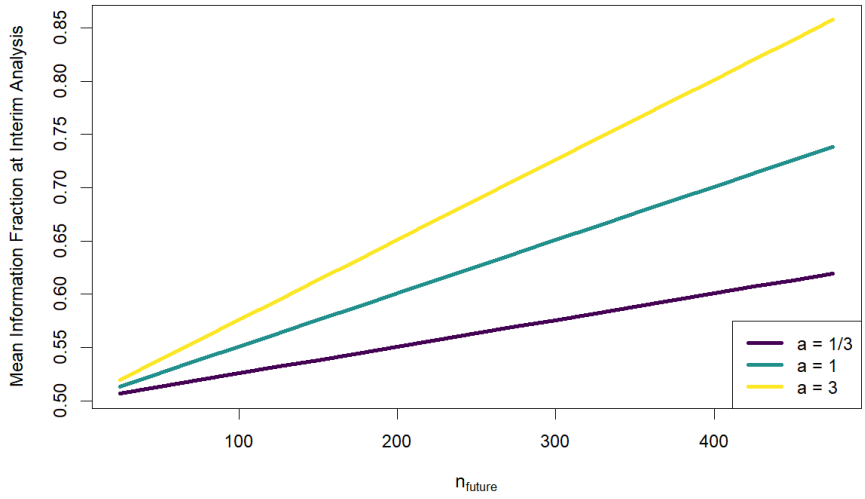



Figure 4.5: Information fraction, as a function of $n_{future}$.

Though the type I error inflation is slight in the results shown in Figure 4.4, this is enough to demonstrate that it is not sufficient to have the correct information fraction in the CHW adjustment. In order for type I error to be adequately controlled when using an adjustment such as the CHW adjustment, the data must be partitioned at the final analysis so that one of the partitions reflects patients whose data were not available to investigators at the time of the adaptation.

## 4.4 Setting 3

This setting considers the possibility of controlling the type I error by correctly specifying the joint distribution of observed test statistics and predicted future test statistics. By conditioning on the interim test statistic and the predicted future test statistics, the type I error may be adequately controlled. Of course, in a real world setting it may be unreasonable to expect for investigators to know the joint distribution. This section merely aims to determine whether the type I error can be controlled, under the strong assumption that the joint distribution can be correctly specified.

### 4.4.1 Notation

Similar to Section 4.2, a GSD is planned before the start of the RCT, such that the interim and final sample sizes are $n_1 = 500$ and $n_2 = 1000$, respectively. O'Brien-Fleming boundaries are used, and the one-sided type I error and power are 0.025 and 0.80, respectively. Individual outcomes have a $\mathcal{N}(\theta, 1)$ distribution, and the variance is known to investigators. Under $H_0$, $\theta = 0$, while under $H_a$, $\theta = \theta_a$, where $\theta_a$ is chosen so that the GSD has the appropriate level of power. At the final analysis, the efficacy boundary is $b$.

The observed test statistic at sample size $n$ is

$$Z_{[1:n]} \coloneqq \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}.$$

At the interim analysis, some future test statistics can be estimated imperfectly, with some future outcomes being estimable with some noise, through the use of surrogate data. That is, at the interim analysis, for fixed $m$ the following are known: $X_1, X_2, X_3, \ldots, X_{n_1}, X_{n_1+1} + \epsilon_{n_1+1}, X_{n_1+2} + \epsilon_{n_1+2}, \ldots, X_{n_1+m} + \epsilon_{n_1+m}$, where $\epsilon_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $j = n_1 + 1, n_1 + 2, \ldots, m$, for some known, fixed $\sigma^2 > 0$.

The estimated test statistic at sample size $n_1 + a$ is then

$$\hat{Z}_{(a)} \coloneqq \frac{\sum_{i=1}^{n_1} X_i + \sum_{j=n_1+1}^{n_1+a} (X_j + \epsilon_j)}{\sqrt{n + a\,(1 + \sigma^2)}},$$

defining as a special case $\hat{Z}_{(0)} \coloneqq Z_{[1:n_1]}$.

Suppose that the available surrogate data is only good for the purpose of calculating $\hat{\boldsymbol{Z}} \coloneqq \begin{pmatrix} \hat{Z}_{(1)} & \hat{Z}_{(2)} & \ldots & \hat{Z}_{(m)} \end{pmatrix}^{\top}$. Then if $n_2$ is adapted to $n_2^*$, one can attempt to control the type I error by choosing a new rejection boundary $b^*$ at the final analysis such that

$$P_{H_0}\left(Z_{[1:n_2]} > b \middle| Z_{[1:n_1]}, \hat{\boldsymbol{Z}}\right) = P_{H_0}\left(Z_{\left[1:n_2^*\right]} > b^* \middle| Z_{[1:n_1]}, \hat{\boldsymbol{Z}}\right).$$

Calculating these conditional probabilities requires the knowledge of the joint distribution of the observed and predicted test statistics. Under $H_0$, the observed and predicted test statistics are each standard normal. The joint distribution is then a multivariate normal distribution, and all that remains is to determine the values of the off-diagonal entries of the covariance matrix.

It is straightforward to show that

$$\mathrm{Cov}\left(\hat{Z}_{(a)}, \hat{Z}_{(b)}\right) = \frac{n_1 + (a \wedge b)\,(1+\sigma^2)}{\sqrt{(n_1 + a\,(1+\sigma^2))\,(n_1 + b\,(1+\sigma^2))}}.$$

Similarly, it is easily shown that

$$\mathrm{Cov}\left(Z_{[1:n]}, \hat{Z}_{(a)}\right) = \frac{n \wedge (n_1 + a)}{\sqrt{n\,(n_1 + a\,(1+\sigma^2))}}.$$

With these expressions, the joint distribution of $Z_{[1:n_1]}$, $Z_{[1:n]}$, and $\hat{\boldsymbol{Z}}$ is specified, and at the interim analysis conditional distributions $\left(Z_{[1:n_2]} \middle| Z_{[1:n_1]}, \hat{\boldsymbol{Z}}\right) \sim \mathcal{N}(\mu_1, \nu_1^2)$ and $\left(Z_{\left[1:n_2^*\right]} \middle| Z_{[1:n_1]}, \hat{\boldsymbol{Z}}\right) \sim \mathcal{N}(\mu_2, \nu_2^2)$ are known, and $b^*$ is easily found to be $\mu_2 + \frac{\nu_2}{\nu_1}\,(b - \mu_1)$.

Three adaptive rules are considered, and are described in more detail below:

- Adaptive Rule 1: Adapt $n_2$ to the sample size corresponding to the maximum value among the predicted test statistics in $\hat{\boldsymbol{Z}}$.
- Adaptive Rule 2: Restrict the minimum value of $n_2^*$ to $m+1$, and use $\hat{Z}_{(m)}$ to choose $n_2^*$ such that the conditional probability of the final test statistic exceeding $b$ under $H_0$ is maximized, in a manner similar to that of Proschan & Hunsberger.
- Adaptive Rule 3: If the maximum value among the predicted test statistics in $\hat{\boldsymbol{Z}}$ exceeds $b$, adapt $n_2$ to the sample size corresponding to this value. Otherwise, proceed with Adaptive Rule 2.

The values of $m$ and $\sigma$ considered in the simulation study are $m \in \{250, 1000\}$ and $\sigma \in \{0.1, 0.5, 1, 5, 10\}$. As in Section 4.2, type I error, power, and ASN are the metrics considered, and ASNs from comparable GSDs are compared to those from the adaptive

designs.

### 4.4.2 Results

As can be seen in Tables 4.2 and 4.3 below, the type I error is reasonably controlled across most scenarios considered. With $m = 250$, the power decreases when using Adaptive Rule 1, since in this specific scenario the final sample size is forced to be adapted downward, with a maximum possible value of 750.

Under $H_0$, usage of Adaptive Rules 2 and 3 leads to high ASNs, and the comparable GSDs seem to be remarkably more efficient with respect to ASNs. However, these adaptive rules are aiming to increase the type I error when using the original final efficacy boundary, and are not trying to be efficient with respect to the sample size.

Though type I error is adequately controlled in this scenario, and there are minimal restrictions on the range of $n_2^*$ in this setting, knowledge of the joint distribution of the observed and predicted test statistics is required. The results of this setting demonstrate that it is possible to condition on the values of the predicted test statistics. However, in a real-world scenario, proving that the joint distribution is correctly specified is difficult, and therefore this analysis approach is not practical.

When $m = 250$, the type I error is generally at the nominal level. Under most scenarios with $m = 250$, the corresponding GSD is more efficient than the adaptive design with respect to the ASN, as shown in Table 4.2. However, under a few scenarios, this is not necessarily the case. For example, when $\sigma = 0.5$, under $H_0$ the adaptive design is more efficient when using Adaptive Rule 1, and under $H_a$ the adaptive design is more efficient when using any of the three considered adaptive rules.

Under some scenarios, this approach can even be a bit conservative, as shown in Table 4.3. For $\sigma \in \{0.1, 0.5\}$, the type I error rate is noticeably below the nominal level of 0.025. The potential for conservatism, combined with the reliance on knowledge of the joint distribution of observed and predicted test statistics, suggests that this approach

is not a suitable substitute for an analysis that applies a Jenkins, Stone, & Jennison or Irle & Schäfer adjustment, especially because regulators or other critics might not believe that in the analysis the investigators have truly captured all available information from surrogate data.

| $\sigma$ | Rule | Type I Error | Power | ASN, $H_0$ | GSD ASN, $H_0$ | ASN, $H_a$ | GSD ASN, $H_a$ |
|---|---|---|---|---|---|---|---|
| | 1 | 0.024 | 0.596 | 618.989 | 623.075 | 637.047 | 597.783 |
| 0.1 | 2 | 0.022 | 0.776 | 1912.355 | 978.492 | 920.043 | 880.161 |
| | 3 | 0.022 | 0.732 | 1888.935 | 880.815 | 845.178 | 802.557 |
| | 1 | 0.026 | 0.649 | 618.429 | 683.952 | 630.111 | 646.150 |
| 0.5 | 2 | 0.024 | 0.824 | 1913.024 | 1076.254 | 945.952 | 957.833 |
| | 3 | 0.025 | 0.785 | 1888.856 | 959.976 | 861.777 | 865.450 |
| | 1 | 0.032 | 0.650 | 617.699 | 632.471 | 617.343 | 605.248 |
| 1 | 2 | 0.022 | 0.857 | 1913.664 | 1211.303 | 1001.033 | 1065.129 |
| | 3 | 0.029 | 0.808 | 1887.321 | 970.533 | 896.061 | 873.837 |
| | 1 | 0.027 | 0.548 | 618.419 | 529.813 | 582.327 | 523.686 |
| 5 | 2 | 0.022 | 0.894 | 1911.820 | 1359.300 | 1183.598 | 1182.713 |
| | 3 | 0.024 | 0.780 | 1885.530 | 960.747 | 999.611 | 866.062 |
| | 1 | 0.024 | 0.520 | 618.046 | 519.601 | 575.673 | 515.573 |
| 10 | 2 | 0.022 | 0.900 | 1911.732 | 1387.541 | 1218.542 | 1205.150 |
| | 3 | 0.024 | 0.769 | 1885.022 | 934.968 | 1018.225 | 845.582 |

Table 4.2: Operating characteristics when $m = 250$, for each of the three adaptive rules. In addition, ASNs are provided for comparable GSDs.

| $\sigma$ | Rule | Type I Error | Power | ASN, $H_0$ | GSD ASN, $H_0$ | ASN, $H_a$ | GSD ASN, $H_a$ |
|---|---|---|---|---|---|---|---|
| | 1 | 0.021 | 0.810 | 937.734 | 1081.051 | 1144.597 | 961.644 |
| 0.1 | 2 | 0.021 | 0.814 | 1978.967 | 1092.246 | 1320.106 | 970.539 |
| | 3 | 0.021 | 0.810 | 1895.218 | 1081.051 | 1167.687 | 961.644 |
| | 1 | 0.022 | 0.831 | 936.871 | 1126.361 | 1117.860 | 997.642 |
| 0.5 | 2 | 0.021 | 0.855 | 1978.595 | 1220.788 | 1336.357 | 1072.665 |
| | 3 | 0.022 | 0.837 | 1896.059 | 1144.900 | 1154.709 | 1012.372 |
| | 1 | 0.027 | 0.844 | 936.642 | 1096.913 | 1049.819 | 974.247 |
| 1 | 2 | 0.021 | 0.916 | 1978.656 | 1489.413 | 1376.505 | 1286.087 |
| | 3 | 0.025 | 0.875 | 1895.123 | 1232.078 | 1126.682 | 1081.635 |
| | 1 | 0.026 | 0.702 | 933.190 | 774.537 | 796.605 | 718.119 |
| 5 | 2 | 0.024 | 0.967 | 1978.056 | 1859.457 | 1564.482 | 1580.086 |
| | 3 | 0.025 | 0.840 | 1892.348 | 1110.568 | 1089.704 | 985.095 |
| | 1 | 0.024 | 0.648 | 933.978 | 703.086 | 746.372 | 661.351 |
| 10 | 2 | 0.024 | 0.968 | 1978.408 | 1872.911 | 1597.473 | 1590.775 |
| | 3 | 0.024 | 0.817 | 1892.918 | 1056.331 | 1091.967 | 942.004 |

Table 4.3: Operating characteristics when $m = 1000$, for each of the three adaptive rules. In addition, ASNs are provided for comparable GSDs.

# Chapter 5

# Comparing TTE Adaptive Designs to Efficient Adaptive Designs and GSDs

We have thus found that it is important to adequately adjust for surrogacy, but have not found an improvement that is sufficiently advantageous to counteract a lingering criticism that some surrogacy might not be modeled. We now address whether adaptation provides any advantage at all. It is not immediately clear that adaptive designs allowing for usage of surrogate data in their adaptive rules are advantageous to those that do not, or to GSDs. Furthermore, it is of interest to investigate how inefficient adaptive rules affect the efficiency of designs when adjusting for surrogate data. This chapter compares these different types of designs across a number of relevant metrics such as average sample size and average calendar time of the RCT. Because the type I error is adequately controlled in all of the designs considered in this chapter, the Bayes Factor is simply a scalar multiple of the power. Therefore, the Bayes Factor is not considered in this chapter.

## 5.1 Motivation

In Chapter 3, we demonstrated that when surrogate data is used to adapt the final sample size, the type I error can be increased substantially if the final test statistic is not adjusted, and potentially more so if the wrong adjustment is used. In Chapter 4, we showed that to control the type I error, it is not sufficient to use a CHW adjustment where the information fraction used accurately reflects the amount statistical information available at the time of the interim analysis. While the type I error is controlled by an approach in which at the interim analysis the conditional type I error is preserved by conditioning on both the interim test statistic and predicted future test statistics, this approach is conservative in some settings and requires knowledge of the joint distribution of the observed and predicted test statistics.

It appears that when surrogate data is used in the adaptive rule, one must use an adjustment such as those by Jenkins, Stone, & Jennison and Irle & Schäfer, where data coming from patients accrued by the time of the adaptation must be censored at the originally planned analysis time. However, it is not clear if such a restriction results in a loss of efficiency compared to designs whose adaptive rules are not functions of surrogate data available at the time of adaptation. In this chapter, we aim to compare adaptive designs that allow for the possibility that surrogate data was used in their adaptive rules to those that do not, and to GSDs. We compare these designs with metrics such as type I error, power, average sample size (ASN), and average calendar time.

In Section 5.2, we describe the contruction of an efficient design whose adaptive rule is a function of only the interim test statistic. Section 5.3 discusses how adjustments that keep the type I error at the nominal level have the potential to ignore observed events in the final analysis. In Section 5.4, we compare GSDs and adaptive designs that use an efficient adaptive rule, while in Section 5.5 we compare GSDs to adaptive designs using an inefficient adaptive rule coupled with the Irle & Schäfer adjustment.

## 5.2 Efficient Adaptive Designs

Levin et al.[20] describe a procedure to construct an efficient adaptive design with one interim analysis at a prespecified analysis time, and one final analysis at a time determined by the value of the interim test statistic. They begin with the most efficient GSD (with respect to average statistical information collected) that has the following properties:

- Symmetric design in the unified family, as described by Wang & Tsiatis[32]
- One interim analysis and one final analysis
- One-sided type I error $\alpha = 0.025$
- Power $\beta = 0.975$ under $H_a : \theta = \theta_a$

The resulting GSD is one with two analyses and degree of early conservatism $P = 0.542$. Because of the symmetry of the design, the average sample size is the same under the $H_0 : \theta = 0$ as under $H_a$. If the corresponding fixed sample design with the same type I error and power at $H_a$ has a sample size of $n$, this GSD has its interim analysis at $0.5n$, and the final analysis is at $1.18n$. At the interim analysis, the futility and efficacy boundaries on the $Z$ scale are 0.57 and 2.21, respectively. On this scale, the rejection boundary at the final analysis is 2.13.

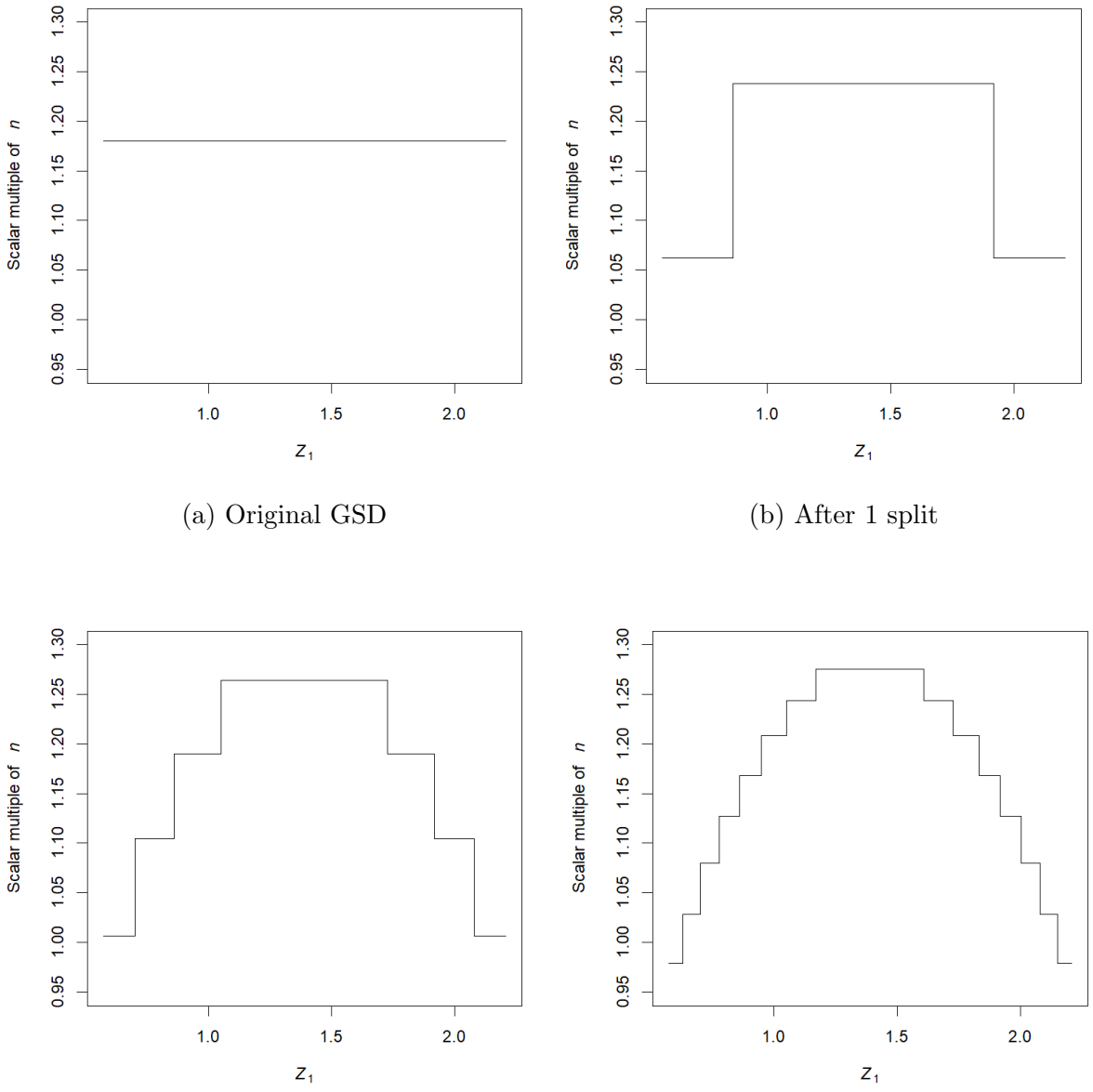
To construct an efficient adaptive design from this GSD, the continuation region at the interim analysis is partitioned into multiple continuation regions in an iterative manner, each corresponding to a new final sample size. The procedure to split a continuation region into two is the following:

1. On the sample mean scale, the initial continuation region $\mathcal{C}$ is of the form $\left(\frac{\theta_a}{2} - d_2, \frac{\theta_a}{2} - d_1\right] \cup \left[\frac{\theta_a}{2} + d_1, \frac{\theta_a}{2} + d_2\right)$, with $0 \leq d_1 < d_2$. The corresponding final sample size and final rejection boundary for $\mathcal{C}$ are $n_2^{\mathcal{C}}$ and $b_2^{\mathcal{C}}$, respectively.

2. $\mathcal{C}$ is partitioned into $\mathcal{C}_1$ and $\mathcal{C}_2$ so that $\mathcal{C}_1$ is of the form $\left(\frac{\theta_a}{2} - c, \frac{\theta_a}{2} - d_1\right] \cup \left[\frac{\theta_a}{2} + d_1, \frac{\theta_a}{2} + c\right)$ and $\mathcal{C}_2$ is of the form

$\left(\frac{\theta_a}{2} - d_2, \frac{\theta_a}{2} - c\right] \cup \left[\frac{\theta_a}{2} + c, \frac{\theta_a}{2} + d_2\right)$, where $c \in (d_1, d_2)$. The corresponding final sample sizes for $\mathcal{C}_1$ and $\mathcal{C}_2$ are $n_2^{\mathcal{C}_1}$ and $n_2^{\mathcal{C}_2}$, respectively, and the corresponding final rejection boundaries are $b_2^{\mathcal{C}_1}$ and $b_2^{\mathcal{C}_2}$, respectively. $c$, $n_2^{\mathcal{C}_1}$, $n_2^{\mathcal{C}_2}$, $b_2^{\mathcal{C}_1}$, and $b_2^{\mathcal{C}_2}$ are chosen such that the type I error and power are preserved, and so that the average sample size is minimized. These values are found via a grid search.

The results of the continuation region splitting process are displayed in graphical form below:




(a) Original GSD

(b) After 1 split

(c) After 2 additional splits

(d) After 4 additional splits

Figure 5.1: Adaptive rule progression, on the $Z$ scale

The two continuation regions in Figure 5.1b are each split into two new continuation regions, resulting in the adaptive rule presented in Figure 5.1c. Similarly, the four continuation regions in Figure 5.1c are each split into two new continuation regions, resulting in the adaptive rule presented in Figure 5.1d.

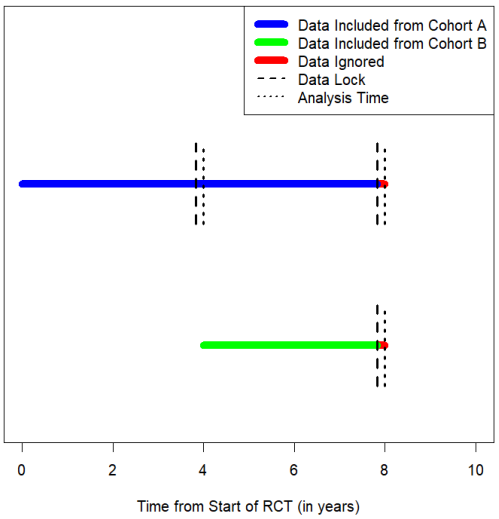
Under the finalized adaptive design, $Z_1$ is calculated at the interim analysis. If $Z_1$ is outside of the interval $(0.57, 2.21)$, the RCT is terminated, rejecting $H_0$ if appropriate. If $Z_1$ is in the aforementioned interval, the adaptive rule represented by Figure 5.1d is used to determine the sample size at the final analysis. The rejection boundary used at the final analysis depends on the final sample size chosen by the adaptive rule.

## 5.3 Adaptive Designs Accounting for Surrogate Data

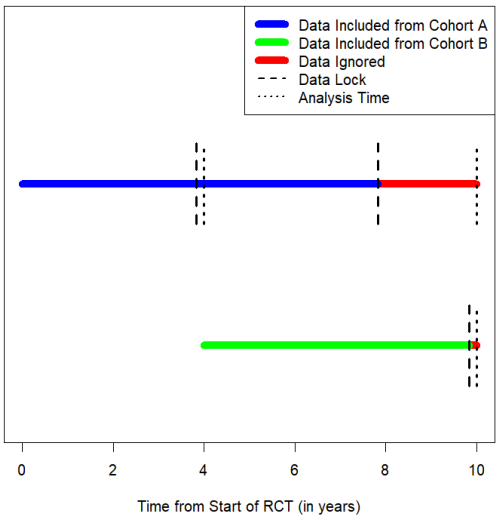
Because of the mathematical similarities between the approaches presented by Jenkins, Stone, & Jennison, Irle & Schäfer, and Magirr et al., it is sufficient to evaluate the performance of one of these approaches. The approach by Irle & Schäfer is chosen because certain weights can be chosen to make the Jenkins, Stone, & Jennison approach mathematically equivalent to it, and because it uses the same paradigm as the Magirr et al. approach.

Magirr et al. point out that when using the Irle & Schäfer adjustment, the cohort of patients recruited up until the moment of adaptation (Cohort A) must be censored at time $\tau_{orig}$ when the data was originally to be locked at the final analysis under the original design. If it is decided that data from the cohort of patients accrued after the moment of adaptation (Cohort B) will be censored at some time $\tau_{new} \neq \tau_{orig}$, it may be the case that some of the events observed by time $\tau_{orig} \vee \tau_{new}$ will be ignored in the final analysis.

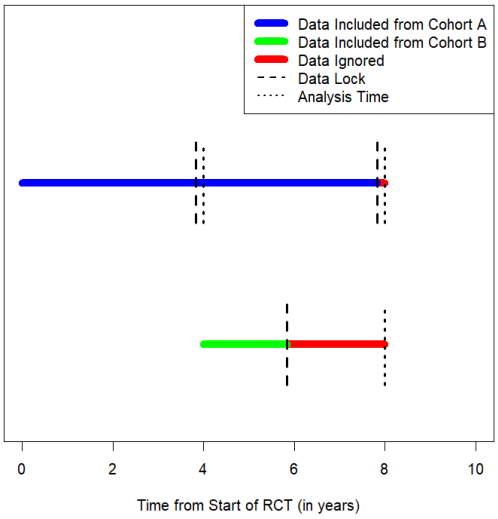
For example, if $\tau_{new} > \tau_{orig}$, any events from Cohort A occuring in the time interval $(\tau_{orig}, \tau_{new})$ are not included in the final analysis, as shown in Figure 5.2b. Conversely, if $\tau_{new} < \tau_{orig}$, any events from Cohort B occuring in the time interval $(\tau_{new}, \tau_{orig})$ are not included in the final analysis, as can be seen in Figure 5.2c.




(a) GSD




(b) Irle & Schäfer, with $\tau_{new} > \tau_{orig}$




(c) Irle & Schäfer, with $\tau_{new} < \tau_{orig}$

Figure 5.2: Illustrations of data to be included and ignored in the final analysis, under different scenarios.

Though Irle & Schäfer's approach controls type I error, the potential for some statistical information to be ignored can lead to loss of power, relative to an approach that wouldn't ignore this information, such as the adaptive design described in the previous section. Because this approach assumes a worst-case scenario of study investigators knowing the event times of patients accrued into the RCT by the time of the adaptation, any adaptive rule is not allowed to influence the time of censoring for data from these patients.

## 5.4 Comparing Designs with Efficient Adaptive Rules

The following designs are compared to each other:

1. The adaptive design described in Section 5.2
2. The efficient GSD used to create the adaptive design described in Section 5.2
3. A GSD that is modified using the adaptive rule described in Section 5.2, using the Irle & Schäfer adjustment at the time of the final analysis.

The original GSD in Design 3 has the same time of the original interim analysis and the same continuation region as Design 2. However, the time of the final analysis for this GSD is at $0.98n$, which is the lowest sample size allowed in the adaptive rule described in Section 5.2, given that the interim test statistic is in the continuation region. This choice is made to ensure that $\tau_{new} \geq \tau_{orig}$. The final rejection boundary of this GSD is chosen so that the type I error of this GSD is 0.025.

Note that with Design 3, the number of events to be included in the analysis follows the adaptive rule described in Section 5.2.

### 5.4.1 Rejection Boundary for Design 3

In the setting of Design 3, calculating the appropriate rejection boundary at the time of the final analysis is straightforward. Suppose that $Z_2$ and $b_2$ are the original full-study-population log-rank test statistic and rejection boundary at the final analysis under the original GSD. Recalling that a log-rank test statistic $Z$ consists of the partial score statistic $S$ and the variance calculation $V$ so that $Z \coloneqq \frac{S}{\sqrt{V}}$, and letting superscripts denote the cohort whose data is used for calculations, $b^*$ is chosen so that $P_{H_0}\left(Z_2 > b_2 \middle| S_2^A\right) = P_{H_0}\left(Z_{new}^B > b^* \middle| S_2^A\right)$, where $Z_{new}^B$ is calculated from data collected from Cohort B and censored at time $\tau_{new}$.

Note that

$$\begin{aligned} P(Z_2 > b_2) &= P\Big(S_2 > b_2\sqrt{V_2}\Big) \\ &= P\Big(S_2 - S_2^A > b_2\sqrt{V_2} - S_2^A\Big) \\ &\approx P\Big(S_2^B > b_2\sqrt{V_2} - S_2^A\Big) \\ &= P\left(Z_2^B > \frac{b_2\sqrt{V_2} - S_2^A}{\sqrt{V_2^B}}\right). \end{aligned}$$

The approximate equivalence holds because $S_2 - S_2^A$ and $S_2^B$ are asymptotically equivalent, as shown by Irle & Schäfer. Because under $H_0$, both $Z_2^B$ and $Z_{new}^B$ are asymptotically standard normal,

$$\begin{aligned} P_{H_0}\big(Z_2 > b_2 \big| S_2^A\big) &\approx P_{H_0}\left(Z_2^B > \frac{b_2\sqrt{V_2} - S_2^A}{\sqrt{V_2^B}}\middle| S_2^A\right) \\ &= P_{H_0}\left(Z_{new}^B > \frac{b_2\sqrt{V_2} - S_2^A}{\sqrt{V_2^B}}\middle| S_2^A\right), \end{aligned}$$

so that choosing $b^*$ to be $\frac{b_2\sqrt{V_2} - S_2^A}{\sqrt{V_2^B}}$ approximately controls the type I error.

With the use of certain weights, the approach by Jenkins, Stone, & Jennison is mathematically equivalent to that of Irle & Schäfer. This is discussed in more detail when adaptive designs with inefficient adaptive rules are considered.

### 5.4.2 Evaluation Metrics

The following metrics are used when comparing the different designs:

- Type I error
- Power
- Average calendar length of RCT
- Average number of events observed during the course of the RCT
- Average number of events ignored during the analyses

### 5.4.3 Scenarios Considered

The same set of accrual patterns and control group survival distributions for the lung cancer and breast cancer examples considered in Chapter 3 are used. Under the alternative hypothesis, the event rate in the treatment group is chosen so that under the GSD and the efficient adaptive design, the power is 97.5%.

In the HIV example used in Chapter 3, it is virtually guaranteed that patient accrual has stopped by the interim analysis. As a result, it is not possible to use the analysis approaches by Irle & Schäfer, Jenkins, Stone, & Jennison, and Magirr et al. Because of this, the HIV setting is not considered when evaluating the different designs.

### 5.4.4 Results

#### 5.4.4.1 Lung Cancer

| Metric | GSD | EAD | I&S |
|---|---|---|---|
| Type I error | 0.024 | 0.024 | 0.022 |
| Average RCT length, in years | 5.904 | 5.893 | 7.103 |
| Average Number of Events Observed | 728.577 | 725.467 | 817.713 |
| Average Number of Events Analyzed | 699.378 | 696.191 | 682.749 |
| Average Number of Events Ignored | 29.199 | 29.276 | 134.964 |

Table 5.1: Results for the lung cancer setting, under $H_0$.

The efficient adaptive design performs similarly to the GSD across all metrics. Compared to these two designs, the Irle & Schäfer design includes fewer observed events in the analyses. It has a notably longer average RCT length in calendar time, and both observes and ignores more events than the other designs.

The Irle & Schäfer design is slightly less powerful than the GSD and the efficient adaptive design. This is likely due to the fact that the Irle & Schäfer analysis approach requires that the data for Cohort A be censored at the originally planned time. As a result, it may be that the final number of events dictated by the adaptive rule is too high because

| Metric | GSD | EAD | I&S |
|---|---|---|---|
| Power | 0.975 | 0.975 | 0.965 |
| Average RCT length, in years | 6.286 | 6.266 | 7.506 |
| Average Number of Events Observed | 730.832 | 726.744 | 818.293 |
| Average Number of Events Analyzed | 703.466 | 699.312 | 675.922 |
| Average Number of Events Ignored | 27.366 | 27.432 | 142.371 |

Table 5.2: Results for the lung cancer setting, under $H_a$.

there either there are an insufficient number of patients in Cohort B or the amount of time required to observe this number of events is too much.

A decrease in power of less than 2% from the originally planned level of 97.5% may be small enough as to deem the use of the Irle & Schäfer adjustment acceptable. However, it is of interest to consider GSDs with the same operating characteristics. If one is committed to the same interim analysis with the same stopping boundaries, the above Irle & Schäfer design is compared to a corresponding two-analysis GSD.

| Metric | I&S | GSD |
|---|---|---|
| Power | 0.965 | 0.965 |
| Average RCT length, in years | 7.506 | 5.997 |
| Average Number of Events Observed | 818.293 | 695.189 |
| Average Number of Events Analyzed | 675.922 | 666.831 |
| Average Number of Events Ignored | 142.371 | 28.358 |

Table 5.3: Comparing Irle & Schäfer design to a similar GSD, under $H_a$.

Compared to the Irle & Schäfer design, the GSD with the same interim analysis and interim stopping boundaries, type I error, and power under $H_a$ is noticeably shorter in average calendar time. The average number of observed events is fewer, but this is mostly offset by the fact that fewer events are ignored in the analyses, on average.

It may instead be the case that the average number of events observed in the Irle & Schäfer design is deemed acceptable. The GSD in Table 5.2 is clearly superior to the Irle & Schäfer design with respect to power, calendar time, and number of events observed.

#### 5.4.4.2 Breast Cancer

| Metric | GSD | EAD | I&S |
|---|---|---|---|
| Type I Error | 0.025 | 0.025 | 0.026 |
| Average RCT length, in years | 11.391 | 11.359 | 11.840 |
| Average Number of Events Observed | 100.763 | 100.171 | 109.665 |
| Average Number of Events Analyzed | 98.161 | 97.579 | 97.536 |
| Average Number of Events Ignored | 2.602 | 2.592 | 12.129 |

Table 5.4: Results for the breast cancer setting, under $H_0$.

Similar trends are observed here as in the lung cancer setting. The GSD and efficient adaptive design are similar with respect to all metrics, and ignore fewer events in the analyses compared to the Irle & Schäfer design.

| Metric | GSD | EAD | I&S |
|---|---|---|---|
| Power | 0.973 | 0.974 | 0.967 |
| Average RCT length, in years | 13.137 | 13.100 | 13.563 |
| Average Number of Events Observed | 102.168 | 101.543 | 108.577 |
| Average Number of Events Analyzed | 99.946 | 99.319 | 96.928 |
| Average Number of Events Ignored | 2.223 | 2.223 | 11.650 |

Table 5.5: Results for the breast cancer setting, under $H_a$.

As in the lung cancer setting, the Irle & Schäfer design is slightly less powerful than the GSD and the efficient adaptive design. Compared to the Irle & Schäfer design, the GSD and the efficient adaptive design each ignore fewer observed events.

Similar procedures to those in Section 5.4.4.1 can be used to find a GSD corresponding to the Irle & Schäfer design, such that they are matched on power and compared on the other metrics.

| Metric | I&S | GSD |
|---|---|---|
| Power | 0.967 | 0.968 |
| Average RCT length, in years | 13.563 | 12.903 |
| Average Number of Events Observed | 108.577 | 98.445 |
| Average Number of Events Analyzed | 96.928 | 96.200 |
| Average Number of Events Ignored | 11.650 | 2.245 |

Table 5.6: Comparing the Irle & Schäfer design to a similar GSD, under $H_a$.

Fixing the timing of the interim analysis and the stopping rule, the Irle & Schäfer design is compared to a GSD with similar power. The GSD is shorter in calendar time and

includes a higher proportion of observed events in the analyses. Note that the GSD in Table 5.5 is more powerful, shorter in calendar time, and includes more observed events in the analyses, compared to Irle & Schäfer design in the same table.

### 5.4.5 Observations

In both the lung cancer and the breast cancer settings, the number of events included in the final analyses are consistently lower when using the Irle & Schäfer design, relative to the number of observed events. From Figure 5.2b, it is clear that this loss of efficiency is due to a number of events from Cohort A not being accounted for in the final analysis. This issue is not present with the GSD and the efficient adaptive design, as the only events not being counted in analyses are those occuring after the dataset has been locked.

The Irle & Schäfer design attempts to compensate for the fact that the censoring time for Cohort A is fixed for the purposes of the final analysis, by extending the follow-up time of Cohort B to observe more events. However, under a fixed accrual rate and with reasonable time constraints on the calendar length of the RCT, the simulation results demonstrate that it is not always possible to achieve the number of events included in the final analysis as it is when using the GSD or the efficient adaptive design.

Though the GSD and the efficient adaptive design appear to be superior to the design using the Irle & Schäfer adjustment, the extent to which they are superior is not very much. For example, in the breast cancer setting the loss in power when using the Irle & Schäfer adjustment is less than 2%. The Irle & Schäfer design has a noticeably longer average time for negligible recovery of the power loss, making this design less desirable.

## 5.5 Comparing Efficient Designs to Inefficient Designs

In the previous section, we demonstrated that when using an efficient adaptive rule, the Irle & Schäfer adjustment has costs with respect to calendar time, power, or both. However, when comparing this design to other efficient designs, the costs appear to be moderate. It may be the case that the use of inefficient adaptive rules may result in greater costs when using adjustments such as that by Irle & Schäfer.

This proposition is reasonable when considering the connection made by Magirr et al. between the approaches by Irle & Schäfer and Jenkins, Stone, & Jennison. Recall that with the Jenkins, Stone, & Jennison method, summary statistics are combined by prespecified weights $w_1$ and $w_2$ with the restriction $w_1^2 + w_2^2 = 1$, so that

$$Z_{JSS} \coloneqq w_1 Z_{\text{Cohort A}} + w_2 Z_{\text{Cohort B}}$$

is the final test statistic. If the original final sample size is $n_2$ and in the full study population, event $n_2$ occurs at time $\tau_{orig}$, then $Z_{\text{Cohort A}}$ is the log-rank test statistic calculated at time $\tau_{orig}$, using data from patients in Cohort A. The length of follow-up for Cohort B must be decided at the time of adaptation, but otherwise there are no restrictions regarding when log-rank test statistic $Z_{\text{Cohort B}}$ must be calculated.

Magirr et al. note that if $n_A$ events are observed in Cohort A by time $\tau_{orig}$, choosing $w_1 = \sqrt{\frac{n_A}{n_2}}$ and $w_2 = \sqrt{1 - \frac{n_A}{n_2}}$ is a special case where the Irle & Schäfer and Jenkins, Stone, & Jennison approaches are equivalent, in that $H_0$ will be rejected using one approach if and only if $H_0$ will be rejected using the other approach.

With these weights, if the data cutoff for Cohort B is also $\tau_{orig}$, then $Z_{JSS}$ is approximately equal to $Z_2$, the log-rank test statistic calculated in the full study population at time $\tau_{orig}$. In this instance, $Z_{JSS}$ is efficient, since $Z_2$ is the minimal sufficient statistic.

However, calculating $Z_{\text{Cohort B}}$ at any other time $\tau_{adapt} \neq \tau_{orig}$ results in events from Cohort A being upweighted or downweighted relative to events in Cohort B. Often, adaptive rules increase the final sample size beyond its originally planned value. In such instances, each individual event observed in Cohort A is given more weight than an observed event in Cohort B.

Each adaptive design in the previous section utilizes some variation of the adaptive rule associated with the efficient adaptive design. In this section, the efficient GSD has a final sample size of $1.18n$, where $n$ is the corresponding fixed sample design with the same type I error and power at $\theta_a$. The efficient adaptive rule allows for the sample size at the final analysis to range from $0.98n$ to $1.28n$. Because the endpoints of this range are relatively close to the value of $1.18n$, the statistic $Z_{JSS}$ does not weigh individual events from Cohort A substantially differently than individual events from Cohort B. However, adaptive rules that increase the sample size by considerable amounts can lead to individual events from Cohort B contributing notably less to $Z_{JSS}$ than those from Cohort A.

For example, consider a scenario that is similar to the lung cancer example, except that the interval over which patients are accrual is doubled, so that the distribution of a patient entering the RCT is uniformly distributed from the beginning of the RCT to 11 years and 10 months after its start. Though up to 1900 patients are accrued into the study, the final analysis of the original GSD occurs at $n_2 = 1000$. Depending on when the interim analysis occurs, the number of patients accrued by the time of adaptation varies, and weights $w_1$ and $w_2$ vary as a result.

| $n_1$ | Observed Events | $w_1^2$ | $w_2^2$ |
|---|---|---|---|
| 250 | 610.138 | 0.610 | 0.390 |
| 500 | 834.851 | 0.835 | 0.165 |
| 900 | 995.399 | 0.995 | 0.005 |

Table 5.7: Average number of observed events in Cohort A by $\tau_{orig}$, and average weights for $Z_{JSS}$, given $n_1$

Table 5.7 can be used to illustrate how $Z_{JSS}$ may weigh different events substantially differently, depending on the timing of the interim analysis relative to the original time

of the final analysis, as well as the adaptive rule.

If the interim analysis occurs at $n_1 = 500$, 834.851 events will be observed in Cohort A by $\tau_{orig}$, on average. Accordingly, $Z_{JSS}$ will roughly give $Z_{\text{Cohort A}}$ a weight of $\sqrt{0.835}$ and $Z_{\text{Cohort B}}$ a weight of $\sqrt{0.165}$, on average. In cases where the adaptive rule modifies the final sample size from 1000 to 2000, $Z_{\text{Cohort B}}$ is calculated using 1165.149 events, on average. Though Cohort B contributes more events to $Z_{JSS}$ than Cohort A, this cohort is treated as if it contributes less than 20% of the statistical information to the final test statistic.

Perhaps in an attempt to lessen the contrast of weights given to individual events in Cohorts A and B, the interim analysis is instead chosen to occur at $n_1 = 250$. With this interim analysis, 610.138 events are given a weight of $\sqrt{0.610}$ and $Z_{\text{Cohort B}}$ a weight of $\sqrt{0.390}$, on average. With the same doubling adaptive rule, $Z_{\text{Cohort B}}$ contributes 1389.862 events, on average, which is more than twice the number of events that Cohort A contributes to $Z_{JSS}$. However, Cohort B is treated as if it contributes less than 40% of the statistical information to the final test statistic, when computing $Z_{JSS}$.

The other extreme is to have the interim analysis very late into the RCT, perhaps at $n_1 = 900$. In such a scenario, a mean of 995.399 events are observed in Cohort A, and are given a weight of $\sqrt{0.995}$, on average. With the same doubling adaptive rule, 1004.601 events are given a weight of just $\sqrt{0.005}$.

Even in a lenient setting such as this one where the length of the accrual window is double of what might be typical, the statistic $Z_{JSS}$ is punishing when using adaptive rules that increase the final sample size by substantial amounts. GSDs and the efficient adaptive design described in the previous section do not suffer from the issue of giving different weights to individual events, as these designs use the minimal sufficient statistic at the final analysis. This suggests that the costs of using designs utilizing the Irle & Schäfer or Jenkins, Stone, & Jennison adjustments may be exacerbated by the use of inefficient adaptive rules.

In the following sections, the adaptive rule mentioned by Mehta & Pocock[17] is used

when modifying an originally planned GSD with power $\beta$ and final rejection boundary $b_2$. If the original analyses of the GSD are at $n_1$ and $n_2$ observed events, respectively, and if $n_{max}$ is the maximum number of events allowed by the adaptation, then the adaptive rule can be expressed in the following manner:

$$n_2^*(z_1) = \begin{cases} n_{max} & z_1 \in (k_1, k_2] \\ n_1 \left( \frac{1}{z_1^2} \left( \frac{b_2\sqrt{\frac{n_2}{n_1}} - z_1}{\sqrt{\frac{n_2}{n_1}}} + z_\beta \right)^2 + 1 \right) & z_1 \in (k_2, k_3] \\ n_2 & \text{otherwise,} \end{cases}$$

where

$$k_1 = b_2 \frac{\sqrt{\frac{n_{max}}{n_2}\left(\frac{n_2}{n_1} - 1\right)} - \sqrt{\frac{n_{max}}{n_2} \times \frac{n_2}{n_1} - 1}}{\sqrt{\frac{n_2}{n_1} - 1} - \sqrt{\frac{n_{max}}{n_2} \times \frac{n_2}{n_1} - 1}} \sqrt{\frac{n_2}{n_1}}$$

$$k_2 = \frac{b_2\sqrt{\frac{n_2}{n_1}} + z_\beta\sqrt{\frac{n_2}{n_1} - 1}}{1 + \sqrt{\left(\frac{n_2}{n_1} - 1\right)\left(\frac{n_{max}}{n_2} \times \frac{n_2}{n_1} - 1\right)}}$$

$$k_2 = \left(b_2 + z_\beta\sqrt{1 - \frac{n_1}{n_2}}\right)\sqrt{\frac{n_1}{n_2}}.$$

At the final analysis, $Z_{JSS}$ is compared to final rejection boundary $b_2$.

In the following sections, we investigate how adaptive designs utilizing the Jenkins, Stone, & Jennison adjustment compare to corresponding GSDs, when inefficient adaptive rules are used. For each resulting adaptive design, two GSDs are constructed for comparison. GSD 1 is constructed so that it has the same interim analysis, stopping boundaries, and type I error, with comparable levels of power under $H_a$. With this construction, the adaptive design and GSD 1 can be compared on metrics such as the average calendar length of the RCT, as well as the average number of events observed, included in the analysis, or ignored by the analysis.

Similarly, GSD 2 is constructed so that it has the same interim analysis, stopping boundaries, and type I error, with a comparable average number of events observed. The

adaptive design and GSD 2 can then be compared on metrics such as power and the average calendar length of the RCT, as well as the average number of events included in the analysis or ignored by the analysis.

### 5.5.1 Timing of Interim Analysis

Adaptive rules can increase the final sample size by substantial amounts. For now, $n_{max}$ is set to $2n_2$. Suppose that $n_2 = 500$, and $n_1 \in \{125, 250, 450\}$, so that the interim analysis occurs after observing 25%, 50%, or 90% of the originally planned number of events. As discussed in Section 5.5, the timing of the interim analysis affects weights $w_1$ and $w_2$ in the adjusted statistic $Z_{JSS}$.

The GSDs are symmetric, and the continuation region for each GSD is chosen to be efficient with respect to the average sample size under $H_0$ and $H_a$, given the timing of the interim analysis relative to the final analysis. The adaptive rule can only increase the GSD's power beyond the initial level of $\beta = 0.975$. The following results reflect key metrics of the adaptive design, as well as those of comparable GSDs.

| Metric | Adaptive Design | GSD 1 | GSD 2 |
|---|---|---|---|
| Power | 0.983 | 0.983 | 0.990 |
| Average RCT length, in years | 7.862 | 6.757 | 7.675 |
| Average Number of Events Observed | 430.610 | 338.830 | 430.570 |
| Average Number of Events Analyzed | 381.684 | 325.276 | 417.472 |
| Average Number of Events Ignored | 48.926 | 13.554 | 13.098 |

Table 5.8: Adaptive design versus GSDs under $H_a$, with $n_1 = 250$.

With the adaptive design, 55% of events included in the final analysis are observed in Cohort A, on average. However, when calculating $Z_{JSS}$ they are treated as if they contribute 87% of the statistical information. The GSD with the same interim analysis, stopping boundaries, and power at $H_a$ is over one year shorter, on average. The number of events required to be observed is substantially less for this GSD than for the adaptive design. Since the final test statistic for the GSD does not weigh events from Cohort A more heavily than those from Cohort B, it is more efficient.

A GSD with a similar average number of observed events is both shorter in calendar time and more powerful than the adaptive design. The increase in power likely comes from the use of the minimal sufficient statistic, as well as fewer ignored events at the final analysis.

| Metric | Adaptive Design | GSD 1 | GSD 2 |
|---|---|---|---|
| Power | 0.977 | 0.977 | 0.981 |
| Average RCT length, in years | 7.258 | 6.241 | 7.148 |
| Average Number of Events Observed | 401.197 | 309.775 | 401.260 |
| Average Number of Events Analyzed | 366.469 | 297.507 | 388.429 |
| Average Number of Events Ignored | 34.728 | 12.267 | 12.831 |

Table 5.9: Adaptive design versus GSDs under $H_a$, with $n_1 = 125$.

When the interim analysis is at $n_1 = 125$, 46% of the events in $Z_{JSS}$ are from Cohort A, but they are treated as if they contribute 66% of the information. The power increase from using the adaptive rule is negligible, considering that the power of the original GSD is 97.5%. The low increase in power may not justify the average calendar length of a RCT with this adaptive design, given that a comparable GSD with the same power has a calendar length that is one year shorter, on average. A comparable GSD with a similar number of average observed events is more powerful and slightly shorter in calendar time than the adaptive design.

| Metric | Adaptive Design | GSD 1 | GSD 2 |
|---|---|---|---|
| Power | 0.980 | 0.980 | 0.993 |
| Average RCT length, in years | 9.676 | 8.316 | 8.778 |
| Average Number of Events Observed | 516.360 | 471.638 | 516.350 |
| Average Number of Events Analyzed | 473.886 | 456.222 | 501.240 |
| Average Number of Events Ignored | 42.474 | 15.416 | 15.110 |

Table 5.10: Adaptive design versus GSDs under $H_a$, with $n_1 = 450$.

With an analysis so late relative to the original final analysis, the 70% of events coming from Cohort A are treated as 99.7% of the statistical information when calculating $Z_{JSS}$, on average. This inefficiency has the consequence of a comparable GSD being shorter in calendar time by over 1.3 years. With a similar number of observed events, a GSD would have a power of 99.3%, compared to the adaptive design's power of 98%. The fact that most patients accrued into the RCT belong to Cohort A means that Cohort B must be followed for a long amount of time to observe the required number of events, just to

contribute 0.3% of the statistical information to $Z_{JSS}$.

### 5.5.2 Maximum Allowed Sample Size Increase

| Metric | Adaptive Design | GSD 1 | GSD 2 |
|---|---|---|---|
| Power | 0.982 | 0.982 | 0.989 |
| Average RCT length, in years | 7.286 | 6.693 | 7.286 |
| Average Number of Events Observed | 392.556 | 332.641 | 392.579 |
| Average Number of Events Analyzed | 353.875 | 319.146 | 378.730 |
| Average Number of Events Ignored | 38.681 | 13.494 | 13.849 |

Table 5.11: Adaptive design versus GSDs under $H_a$, with $n_1 = 250$ and $n_{max} = 1.5n_2$.

When the maximum sample size increase is only 50% instead of 100%, differences in metrics between the adaptive design and the GSDs start to look less drastic. A comparable GSD with similar power is only 0.6 years shorter in calendar time, and a comparable GSD with a similar number of observed events has a power of 98.9%, compared to the adaptive design's power of 98.2%. Comparing this adaptive design to the one in Table 5.8, restricting $n_{max}$ to be $1.5n_2$ instead of $2n_2$ leads to a decrease in power of only 0.1%. The imbalance in weights in this setting is lessened, too. Here, 64% of the events come from Cohort A, while when $n_{max} = 2n_2$, only 55% do. In both scenarios, these events are treated as 87% of the statistical information.

| Metric | Adaptive Design | GSD 1 | GSD 2 |
|---|---|---|---|
| Power | 0.980 | 0.980 | 0.988 |
| Average RCT length, in years | 7.017 | 6.670 | 7.022 |
| Average Number of Events Observed | 365.095 | 330.401 | 365.188 |
| Average Number of Events Analyzed | 334.535 | 316.940 | 351.513 |
| Average Number of Events Ignored | 30.560 | 13.461 | 13.675 |

Table 5.12: Adaptive design versus GSDs under $H_a$, with $n_1 = 250$ and $n_{max} = 1.25n_2$.

Further restricting $n_{max}$ to be $1.25n_2$ lessens the inefficiency of the adaptive design. Now 73% of the events used in $Z_{JSS}$ come from Cohort A, and are treated as 87% of the statistical information. A comparable GSD with the same power is about a third of a year shorter in calendar time, on average. This adaptive design's power is nearly the same as when $n_{max} = 2n_2$.

### 5.5.3 Width of Original GSD's Continuation Region

All GSDs considered thus far have stopping boundaries that minimize the average sample size under $H_0$ and $H_a$. The designs in Table 5.8 have a continuation region of $(0.2324, 2.0964)$, on the $Z$ scale. The following table instead considers a continuation region of $(0.0000, 2.7897)$ on the same scale.

| Metric | Adaptive Design | GSD 1 | GSD 2 |
|---|---|---|---|
| Power | 0.984 | 0.984 | 0.997 |
| Average RCT length, in years | 8.760 | 7.580 | 8.552 |
| Average Number of Events Observed | 511.575 | 413.110 | 511.352 |
| Average Number of Events Analyzed | 456.307 | 398.676 | 496.276 |
| Average Number of Events Ignored | 55.268 | 14.434 | 15.076 |

Table 5.13: Adaptive design versus GSDs under $H_a$, with $n_1 = 250$, using inefficient stopping boundaries.

Having an inefficient continuation region appears to increase the inefficiency of the adaptive design. A comparable GSD with the same continuation region and power is about 1.2 years shorter in calendar time, on average. With the same average number of observed events, using a GSD instead increases the power from 98.4% to 99.7%, and decreases the average calendar time of the RCT slightly. The 66% of events that come from Cohort A are given a weight of 87%.

### 5.5.4 Observations

Across all scenarios considered, each adaptive design has a corresponding GSD that is more efficient, more powerful, or both. The fact that GSDs use the minimal sufficient statistic at the final analysis likely contributes to this trend, as well as the fact that GSDs ignore fewer observed events.

The largest driving factor in inefficiencies of the adaptive designs appears to be the maximal allowed sample size increase. It seems that the further an adaptive rule deviates from one corresponding to an efficient adaptive design, the more inefficient the analysis is.

## 5.6 Conclusions

There is a cost to adjusting for surrogate data to control the type I error. These costs are modest when the adaptive rule used is efficient. However, rules that increase the originally planned final sample size by substantial amounts can increase these costs.

When using the adjustment by Jenkins, Stone, & Jennison or the one by Irle & Schäfer, one must be mindful of the accrual pattern for the RCT, relative to the timing of the interim analysis. An accrual pattern that recruits most patients by the time of the adaptation results in events from Cohort A being weighed heavily relative to those from Cohort B, when the final sample size is increased substantially.

Furthermore, because the censoring time of data from Cohort A is fixed when using either of these adjustments, care must be taken to ensure that enough patients are accrued into Cohort B, and that the follow-up time for Cohort B is sufficiently long so that the number of events to be observed from this Cohort is achieved.

Designs that use the minimal sufficient statistic are more efficient than those that adjust for surrogate data, and do not require the same considerations regarding accrual and follow-up time of Cohort B. In many cases a simple efficient GSD can achieve similar operating characteristics to those of an adaptive design that adjusts for surrogate data.

# Chapter 6

# Overall Conclusions and Future Directions for Research

The results of this dissertation are of interest to regulators and investigators alike, and should be taken into consideration when planning time-to-event RCTs that may allow for surrogate data to be used to modify the sequential sampling plan. Moreover, these issues are not necessarily restricted to the time-to-event setting, as surrogate data can also be used to estimate future measurements in a longitudinal setting.

For example, in a setting involving patients with chronic obstructive pulmonary disease, the primary outcome of interest may be the change in distance walked over a period of six minutes, from the baseline measurement to the measurement taken two years after the patient has been randomized to a treatment arm. However, for patients who have been participating in the RCT for less than two years, measurements taken one year or 1.5 years after enrollment may be informative regarding the still-unobserved measurements at two years, and similar issues regarding inappropriate use of surrogate data may arise.

The data partioning framework developed in Chapter 2 allows for the problem to be stated succinctly: Most analysis approaches following a modification of the sequential sampling plan assume that at the time of the adaptation, events of patients that are participating or will eventually participate in the RCT belong to either $\mathcal{D}_1$ or $\mathcal{D}_4$, and the few approaches

that allow for events to belong to $\mathcal{D}_2$ or $\mathcal{D}_3$ are restrictive or conservative, and may even be inefficient. More generally, this framework is useful when discussing and investigating issues with inappropriate usage of surrogate data.

In the following sections, we summarize the results from Chapters 3, 4, and 5, and point to potential directions for future research.

## 6.1 Conclusions

The results in Chapter 3 allow for a number of factors that contribute to type I error inflation to be identified, in cases where investigators do not account for surrogate data or account for it incorrectly. Some choices that lead to increased type I error rates are knowledge of $\mathcal{D}_2$ alone or with knowledge of $\mathcal{D}_3$, an early interim analysis, stopping accrual after the time of adaptation, and loose restrictions on the modification of the sequential sampling plan. Contrary to what one might have assumed before seeing these results, conservative spending of error early and usage of the CHW adjustment at the final analysis do not adequately control the type I error, and in fact exacerbate the type I error inflation. In the lung cancer setting considered, all of the above choices combined lead the type I error to be inflated from the nominal value of 0.025 to levels as high as 0.205, an over-eight-fold increase.

The explorations in Chapter 4 suggest that estimating the amount of statistical information gathered by the time of the adaptation and adjusting for this amount using a CHW adjustment is not sufficient to control the type I error. It is possible to control the type I error if investigators know the joint distribution of observed and predicted test statistics, and preserve the conditional type I error at the time of adaptation by conditioning on the interim test statistic as well as all test statistics predicted into the future. However, real world settings generally require analysis approaches such as those proposed by Jenkins, Stone, & Jennison, Irle & Schäfer, and Magirr et al.

In Chapter 5, we find that the use of efficient adaptive rules leads to moderate inefficiencies

of adaptive designs that adjust for surrogate data, compared to GSDs or efficient adaptive designs that do not have to adjust for surrogate data. However, the use of inefficient adaptive rules leads to increased inefficiencies, compared to GSDs. Of particular concern is the observation that in a realistic setting considered, using a Jenkins, Stone, & Jennison or Irle & Schäfer analysis adjustment can result in over half of the observed events being given a weight of 0.5%. The inefficiencies of these analysis approaches are significant enough that the flexibility that adaptive designs provide come at a very high cost, and the use of GSDs or efficient adaptive designs, combined with some careful planning, is more appealing.

## 6.2 Future Directions for Research

In this dissertation, certain aspects of the data collection process and the underlying properties of the data to be collected were held constant. However, relaxing these restrictions may provide further insights into the issues this dissertation considered. For example, how the patient accrual pattern affects metrics such as type I error and power was not explored in any meaningful manner. Deviating from an accrual pattern that is Uniformly distributed or allowing for the post-adaptation accrual pattern to be a function of the surrogate data may lead to even greater levels of type I error inflation, when investigators do not adjust for surrogate data at the final analysis or adjust for it incorrectly. Similarly, allowing for the hazard rates of the different treatment arms to not be proportional may further inflate the type I error. Future research might explore these different scenarios.

As shown in Chapter 4, knowledge and use of the joint distribution of the observed and predicted test statistics can lead to the type I error being adequately controlled. However, it is unrealistic for investigators to know this joint distribution. A first step to a more realistic setting is one where this joint distribution can be adequately estimated at the time of the final analysis so that the type I error is still controlled, and future research may explore whether this is a feasible procedure.

It was observed in Chapter 5 that prespecifying weights for the Jenkins, Stone, & Jennison adjustment such that this approach is equivalent to the Irle & Schäfer adjustment can lead to great statistical inefficiencies. However, the weights do not necessarily have to be prespecified to these values. Given event rates and a range of possible values for the final sample size adaptive rule, it is possible that there are sensible accrual patterns and optimal prespecified weights for the Jenkins, Stone, & Jennison adjustment that minimize the inefficiencies of adaptive designs that account for surrogate data. Future research can explore this possibility.

# References


[1] Emanuel, Ezekiel J et al. "The costs of conducting clinical research". In: *Journal of Clinical Oncology* 21.22 (2003), pp. 4145–4150.

[2] Armitage, Peter, McPherson, CK, and Rowe, BC. "Repeated significance tests on accumulating data". In: *Journal of the Royal Statistical Society. Series A (General)* (1969), pp. 235–244.

[3] Emerson, Scott S, Kittelson, John M, and Gillen, Daniel L. "Frequentist evaluation of group sequential clinical trial designs". In: *Statistics in medicine* 26.28 (2007), pp. 5047–5080.

[4] Lewis, Roger J and Berry, Donald A. "Group sequential clinical trials: A classical evaluation of Bayesian decision-theoretic designs". In: *Journal of the American Statistical Association* 89.428 (1994), pp. 1528–1534.

[5] Emerson, Scott S, Kittelson, John M, and Gillen, Daniel L. "Bayesian evaluation of group sequential clinical trial designs". In: *UW Biostatistics Working Paper Series* (2005).

[6] Pocock, Stuart J. "Group sequential methods in the design and analysis of clinical trials". In: *Biometrika* (1977), pp. 191–199.

[7] O'Brien, Peter C and Fleming, Thomas R. "A multiple testing procedure for clinical trials". In: *Biometrics* (1979), pp. 549–556.

[8] Kittelson, John M and Emerson, Scott S. "A unifying family of group sequential test designs". In: *Biometrics* 55.3 (1999), pp. 874–882.

[9] Demets, David L and Lan, KK. "Interim analysis: the alpha spending function approach". In: *Statistics in medicine* 13.13-14 (1994), pp. 1341–1352.

[10] FDA. “Adaptive design clinical trials for drugs and biologics”. In: *Biotechnol Law Rep* 29.2 (2010), p. 173.

[11] Proschan, Michael A and Hunsberger, Sally A. “Designed extension of studies based on conditional power”. In: *Biometrics* (1995), pp. 1315–1324.

[12] Jennison, Christopher and Turnbull, Bruce W. “Mid-course sample size modification in clinical trials based on the observed treatment effect”. In: *Statistics in medicine* 22.6 (2003), pp. 971–993.

[13] Bauer, Peter and Kohne, K. “Evaluation of experiments with adaptive interim analyses”. In: *Biometrics* (1994), pp. 1029–1041.

[14] Fisher, Ronald Aylmer. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[15] Fisher, Lloyd D. “Self-designing clinical trials”. In: *Statistics in medicine* 17.14 (1998), pp. 1551–1562.

[16] Cui, Lu, Hung, HM, and Wang, Sue-Jane. “Modification of sample size in group sequential clinical trials”. In: *Biometrics* 55.3 (1999), pp. 853–857.

[17] Mehta, Cyrus R and Pocock, Stuart J. “Adaptive increase in sample size when interim results are promising: A practical guide with examples”. In: *Statistics in medicine* 30.28 (2011), pp. 3267–3284.

[18] Tsiatis, Anastasios A and Mehta, Cyrus. “On the inefficiency of the adaptive design for monitoring clinical trials”. In: *Biometrika* (2003), pp. 367–378.

[19] Jennison, Christopher and Turnbull, Bruce W. “Adaptive and nonadaptive group sequential tests”. In: *Biometrika* (2006), pp. 1–21.

[20] Levin, Gregory P, Emerson, Sarah C, and Emerson, Scott S. “Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation”. In: *Statistics in medicine* 32.8 (2013), pp. 1259–1275.

[21] Brannath, Werner, König, Franz, and Bauer, Peter. “Estimation in flexible two stage designs”. In: *Statistics in Medicine* 25.19 (2006), pp. 3366–3381.

[22] Levin, Gregory P, Emerson, Sarah C, and Emerson, Scott S. "An evaluation of inferential procedures for adaptive clinical trial designs with pre-specified rules for modifying the sample size". In: *Biometrics* 70.3 (2014), pp. 556–567.

[23] Schoenfeld, David. "The asymptotic properties of nonparametric tests for comparing survival distributions". In: *Biometrika* (1981), pp. 316–319.

[24] Tsiatis, Anastasios A. "Repeated significance testing for a general class of statistics used in censored survival analysis". In: *Journal of the American Statistical Association* 77.380 (1982), pp. 855–861.

[25] Bauer, P and Posch, M. "Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller, Statistics in Medicine 2001; 20: 3741–3751". In: *Statistics in Medicine* 23.8 (2004), pp. 1333–1334.

[26] Jenkins, Martin, Stone, Andrew, and Jennison, Christopher. "An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints". In: *Pharmaceutical statistics* 10.4 (2011), pp. 347–356.

[27] Irle, Sebastian and Schäfer, Helmut. "Interim design modifications in time-to-event studies". In: *Journal of the American Statistical Association* 107.497 (2012), pp. 341–348.

[28] Magirr, Dominic et al. "Sample size reassessment and hypothesis testing in adaptive survival trials". In: *PloS one* 11.2 (2016), e0146465.

[29] Group, International Adjuvant Lung Cancer Trial Collaborative et al. "Cisplatin-based adjuvant chemotherapy in patients with completely resected non–small-cell lung cancer". In: *N Engl J Med* 2004.350 (2004), pp. 351–360.

[30] Hage, Jos A van der et al. "Preoperative chemotherapy in primary operable breast cancer: results from the European Organization for Research and Treatment of Cancer trial 10902". In: *Journal of Clinical Oncology* 19.22 (2001), pp. 4224–4237.

[31] HIV Vaccine Study Group, rgp120 et al. "Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection". In: *Journal of Infectious Diseases* 191.5 (2005), pp. 654–665.

[32] Wang, Samuel K and Tsiatis, Anastasios A. “Approximately optimal one-parameter boundaries for group sequential trials”. In: *Biometrics* (1987), pp. 193–199.