Information Growth in Longitudinal Clinical Trials

Abigail B. Shoben

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

Doctor of Philosophy

University of Washington

2010

Program Authorized to Offer Degree: Biostatistics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Abigail B. Shoben

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____

Scott S. Emerson

Reading Committee:

_____

Scott S. Emerson

_____

Patrick J. Heagerty

_____

Kenneth M. Rice

Date: _____

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature_____

Date_____

University of Washington

Abstract

Information Growth in Longitudinal Clinical Trials

Abigail B. Shoben

Chair of the Supervisory Committee:
Professor Scott S. Emerson
Biostatistics

In group sequential trials, interim analyses are performed for ethical and financial considerations. In order to perform such analyses, estimates of the statistical information at the time of the interim analysis relative to the amount of statistical information at the end of the trial are needed. Longitudinal trials where the primary outcome is a change over time (slope) present special problems for estimates of the information growth. We consider potential difficulties due to (a) failing to estimate correctly the information in a longitudinal setting, (b) heteroscedasticity over the course of the trial, and (c) correlation of measurements on the same individual. In some longitudinal trials these issues result in the covariance of the interim and final statistics not having the assumed form of independent increments due to using an inefficient statistic. This is of a practical concern due to reliance of most sequential software on the independent increment assumption.

We demonstrate that dramatic misestimation of the information at interim analysis times can lead to inflated type I error rates and loss of power for a specific alternative. The amount of heteroscedasticity also impacts the information at interim analyses, which implies that the information growth over time will be different for different alternatives in the setting of a mean-variance relationship. We found that this relationship can cause difficulties in extreme mean-variance settings, but that adjusting for the observed true information growth at the end of the study is sufficient to maintain type I error rates in most realistic circumstances. With correlated data, we illustrate circumstances in which

the generalized estimating equation (GEE) method can lead to nonmontonic information growth in longitudinal trials. We describe cases in which such nonmontonicity is possible and discuss possible options for such settings.

We also demonstrate situations with heteroscedastic and correlated data that lead to violations of the independent increment assumption for the covariance of interim and final statistics in a group sequential trial. Specifically, we give circumstances in which the lack of independent increments may cause departures from the nominal type I error rate. We illustrate that in most common circumstances, existing group sequential methodology can be used despite departures from the independent increment assumption, and we give guidelines for when such departures may be problematic.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere gratitude to Professor Scott Emerson for his support, advice, and encouragement throughout my time at the University of Washington. Thanks are also due to Professors Tom Fleming, Patrick Heagerty, and Ken Rice for their helpful comments and suggestions as part of my dissertation committee. Professor Chris Li was also very helpful in his capacity as Graduate School Representative. Special thanks to the Department of Biostatistics for its support and to the faculty, staff, and students for their dedication and camaraderie.

DEDICATION

To my family.

Chapter 1

INTRODUCTION

The purpose of this introduction is to provide a brief overview of existing methodology in sequential clinical trials and to provide motivation for this work in the specific case of longitudinal clinical trials. Further details on existing methodology are found in background chapters on group sequential trials and longitudinal data.

1.1 Why Clinical Trials?

Evidence-based medicine requires that treatment options be evaluated in the most rigorous manner possible. Usually this approach requires a series of clinical trials, in different phases. Phase 1 tests for toxicity in human subjects. Phase 2 investigates dose-response and proof of concept sufficient to continue to a phase 3 study. Phase 3 tests for efficacy of a treatment, usually with a randomized design. Randomization allows for an unconfounded comparison of two or more treatment groups. Assuming that the trial is well-planned and well-executed, this randomization provides the strongest evidence with which to compare the groups and ultimately decide whether a treatment should be approved.

Ethical and validity concerns govern regulatory agencies that decide if a treatment should be approved. These regulatory agencies in most countries, including the United States, require that all statistical analyses for the trials done on humans are prespecified to ensure that correct statistical inference can be made. For instance, a decision to compare geometric mean weight loss rather than arithmetic mean weight loss prior to data collection is perfectly acceptable depending on the scientific circumstance, but deciding to switch to the geometric mean after observing some data is not. The prespecified statistical analysis plan must ensure that the overall type I error rate of the study is maintained regardless of what happens in the trial.

1.2 Conduct of Clinical Trials

Concepts of time in a sequential clinical trial are of particular importance in the setting of longitudinal trials. In actual calendar time, a clinical trial has several stages, which may overlap. There is a stage before actual recruitment of patients for the study, during which planning for the study and training of the study coordinators may occur. Once the study is ready to begin, recruitment of patients for the study and randomization take place. The time during which new patients are being actively recruited and randomized for the study is known as the accrual period. This period could be very brief if participants can be recruited from everyone with a prevalent disease, such as testing pain medication in arthritis. In such a case, many patients could be recruited and started on the randomized treatment quickly. In contrast, accrual could be lengthy, as might be the case if the trial is being conducted in a rare disease and only incident cases can be included. Accrual patterns can also be influenced by factors that could change over the course of the study. For example, in a multi-center study, all sites may not start recruiting at the same calendar time, so accrual could be slow at the start of the trial and then increase. The rate of successful recruitment of eligible patients may also change over time. For example, study coordinators may become better at identifying patients who are eligible for the study and/or ensuring that patients have all the information to make an informed decision about whether to participate.

Measuring the outcomes of patients enrolled in the trial is the next stage, and this stage may be overlapping with the accrual portion, as outcomes for some patients may be (and likely are) measured while other patients are being accrued. There may be a large amount of overlap between the accrual and the measuring of the outcomes, or most outcome measurements may be obtained after accrual is finished, such as in long-term survival studies. Once the last outcome has been measured, the clinical trial may have an end stage in which patients are followed for additional adverse events, but this time period is not relevant for our purposes. We are concerned with the total calendar time from the start of accrual to when the last outcome measurement is obtained.

For the purpose of this dissertation, we will be interested in longitudinal clinical trials in which the primary interest is in the rate of change over time (a slope). For the sake of

clarity, we use "calendar time" as defined above (starting at the time the first participant is accrued and ending when the last measurement has been made). We use "study time" to denote the time from randomization on each participant.

1.3 Monitoring a Clinical Trial

Clinical trials are conducted in human volunteers and thus ethical concerns are of paramount importance. These ethical concerns motivate the use of interim analyses and sequential monitoring of data from a clinical trial in order to prevent undue harm to test subjects or to expedite delivery of effective treatments. Additionally, interim analyses can halt trials once clinically important results can be ruled out, which can save monetary and scientific resources.

Special statistical techniques are needed to conduct interim analyses of a trial because test statistics from the interim analyses are based on varying amounts of information and are correlated with each other. Group sequential methods to account for this correlation have been developed exactly with the following assumptions (Whitehead, 1997; Jennison and Turnbull, 2000):

- The scientific interest lies in the mean of the data or the mean of a transformation of the data.

- The primary outcome is measured once (and only once) on each individual in the study. This assumption requires that outcome measurements will be available on participants after a fixed amount of time from randomization (e.g. cholesterol measurement 3 months after starting treatment) and that additional measurements on individuals already measured will not be made if the study is continued.

- The contributions of measured outcomes to the test statistic are independent of each other.

- The outcome measurements have a constant variance (homoscedastic).

- The variance of the outcome measurements is known.

- The (possibly transformed) outcome measurements are normally distributed.

Existing statistical methods for group sequential trials have been developed and explored under assumptions all of the above conditions hold. Further, in some cases, it has been shown that some of these assumptions can be relaxed and that the existing methodology will still work well, even though the assumptions are not met exactly. Specifically, existing techniques permit relaxation of the following assumptions:

- **Normally Distributed Data:** Provided the sample size of a trial is large enough at each interim analysis for the Central Limit Theorem to hold, the assumption of normally distributed data is satisfied by the asymptotic normality of the statistic, even without normally distributed outcome data.

- **Known Variance of the Outcome Measurements:** If the variance can be consistently estimated from the data, using the estimate from the data can work as an approximate value in the existing methods given a large enough sample size.

- **Constant Variance of the Outcome Measurements:** This assumption can be relaxed in two circumstances. (1) If there is a predictor-variance relationship (e.g. measurements are becoming more variable as study time increases, or there is a difference in the variance by treatment group), but the design is completely balanced at all interim analyses on the predictor that is associated with the changing variance, then there is no problem due to the heteroscedasticity. (2) If the non-constant variance is due to a mean-variance relationship but the sample size is sufficiently large, changes in the variance due to a changing alternative hypothesis are slight due to the large sample size in which case the assumption of homoscedasticity is reasonable. Further, the type I error rate will be maintained under the strong null regardless of sample size with an assumed mean-variance relationship.

- **Independent Data:** Existing group sequential methods have been shown to work with dependent data, provided they are analyzed with the efficient statistic for the

situation. In such circumstances, the dependent data are combined into a statistic in such a way (through weighting) to attain properties of independent data.

- **Single Outcomes:** One common case in which outcomes are measured repeatedly in time but that existing group sequential methodology can be used is survival data. The use of proportional hazards in survival models incorporates a repeated outcome in that individuals in the study may experience events at different times; however, if the proportional hazards assumption holds, existing techniques may be used.

Each of these assumptions is important to conduct interim analyses correctly. In order to perform interim analyses in a clinical trial using the existing statistical techniques, the variability of the statistic must be estimated at each interim analysis. The variability of the statistic is generally the inverse of Fisher's Information; hence we refer to the inverse of the variance of the statistic as the "information." In general, the variability of the statistic will decrease (the information will increase) as more outcome measurements are obtained. The rate at which statistical information increases over the course of a trial (the "information growth") will depend on both the accrual pattern and the amount of additional statistical information obtained from newly measured outcomes. Existing statistical techniques for sequential clinical trials rely on being able to estimate the fraction of total information present at each interim analysis time.

1.4 Our Focus and Motivation

The information growth in longitudinal clinical trials is the primary focus of this dissertation. In particular, we are interested in the ways that heteroscedastic and correlated data affect the information in a sequential clinical trial.

These issues can be motivated by considering a clinical trial designed to study a rate of change in longitudinal Poisson counts, such as monitoring the rate of decline of skin lesions over time. Such a trial would violate the assumptions needed for standard sequential monitoring of a longitudinal trial. Specifically, correctly measuring the information growth in this scenario may be difficult due to: (a) repeated measurements, (b) correlation of

the measured outcomes, and (c) heteroscedasticity due to a mean-variance relationship. Further, some accepted statististical techniques may not be the efficient statistic in this setting and thus (d) the use of the inefficient statistic may cause additional difficulties. For the purposes of this work, we are interested in the population-level effect of a treatment and thus we will be focused on methods that allow for marginal inference on the slope parameter. In particular, we explore the use of least squares slope (an inefficient statistic) with correlated data.

In order to explore these issues fully and to examine the consequences of violating each assumption in turn, we will focus on a linear model of data with a possible mean-variance relationship and possible correlation between measurements on the same subject. Letting i denote the randomized treatment group, we are interested in a simple regression model such that for a particular study time (after randomization) x, the mean in treatment i is given by:

$$\mu_{ix} = \beta_{0i} + \beta_{1i}x \tag{1.1}$$

For our purposes, we will assume that at the planned end of the study, all individuals in all treatment groups will have been observed at identical study times. Thus the vector of study observation times at the end of the study, $\boldsymbol{x}$, is constant across treatment groups and individuals. Thus, the vector mean of observations at all study times $\boldsymbol{x}$ is given by:

$$\begin{aligned}\boldsymbol{\mu_i} &= \beta_{0i}\mathbf{1} + \beta_{1i}\boldsymbol{x} \\ \boldsymbol{\mu_i} &= \boldsymbol{X}\beta_i\end{aligned}$$

where $\boldsymbol{X}$ is the design matrix that in our case combines the vectors $\mathbf{1}$ and $\boldsymbol{x}$.

Then, letting j denote a specific individual, we assume that the measured vector of outcomes at study times $\boldsymbol{x}$, $\boldsymbol{Y_{ij}}$, is distributed as:

$$\boldsymbol{Y_{ij}} \sim (\boldsymbol{\mu_i}, \sigma_i^2 V(\boldsymbol{\mu_i})) \tag{1.2}$$

The matrix $V(\boldsymbol{\mu_i})$ will allow for both correlation between measurements on the same individual and for a mean-variance relationship due to the dependence on the mean vector

$\boldsymbol{\mu_i}$, though we will also consider a predictor-variance relationship with this model. For purposes of simulation, we will assume that these Y_{ij} are multivariate normal.

We will first explore the case of independent data, as might be the case if a clinical trial was conducted to study tumor growth in rats. In such a case, randomized rats may be given cancer at a specific starting time but then the outcome measurements of tumor size are obtained at different study times on different rats, as the rat must be sacrificed to measure the tumor size.

Within the setting of independent data, we first explore consequences of violating the assumption of being able to estimate consistently a constant variance in the setting of model misspecification, where the estimate of the variance may not be constant over the course of the study (chapter 4). We then explore consequences of violating the assumption of homoscedasticity with heteroscedasticity due to a predictor-variance relationship (chapter 5) and due to a known mean-variance relationship (chapter 6). In both settings we examine the impact of proceeding naively with existing group sequential methods and provide recommendations for design and inference of trials in which heteroscedasticity may be a concern.

After exploring independent linear data, we explore correlated linear data to examine consequences of violating the assumption of independent data in chapter 7. As before, we examine the impact of proceeding naively with existing methods and provide recommendations for design and inference in trials with longitudinal correlated data. In this setting we explore consequences of a known mean-variance relationship as well.

Finally, we provide recommendations for future studies and evaluate our recommendations with a specific example in chapter 8.

There are important issues with longitudinal clinical trials that are beyond the scope of this work. Specifically, although we consider the case of model misspecification to examine consequences of inconsistently estimated variance, we do not consider further cases in which the model is not correctly specified (either the mean or the mean-variance relationship). If the mean model is not correctly specified (i.e. the data are not perfectly linear), then each of the interim analyses would be estimating a linear contrast over a different period of study time. Estimates from interim analyses would thus be estimating different scientific

quantities and there are important scientific considerations about interim analyses in this setting that are beyond the scope of this work. If the effect does differ over time, such as can happen with early and late effects in survival data, this has important implications both for the scientific and statistical considerations of the trial. The situation in which the effect is different early in the trial compared to later with survival data has been investigated by Gillen and Emerson (2005) and Hanley (2005). If the model is correctly specified and the data are linear, then the contrast estimated at every interim analysis is the same slope that would be estimated at the final analysis. This situation is the one we consider for this work.

Further, although the scope of this work does not include nonlinear models, as might be used in the case of longitudinal Poisson count data, the results about information growth would likely generalize on the transformed scale to such a setting. Additionally, longitudinal studies in the real world may have observed observation times for each individual that are random and not fixed at the same time for every individual, including possible missing data at some expected measurement times. Theses issues are also beyond the scope of this dissertation.

Chapter 2

BACKGROUND – GROUP SEQUENTIAL TRIALS

The need for interim analyses in clinical trials is driven by ethical and financial concerns. We want to be able to stop trials early for both positive and negative reasons. If the new treatment is highly advantageous, we want to stop the trial early to expedite delivery of the effective treatment to eligible patients, including those in the control arm of the trial. If the new treatment is suggestive of harm rather than benefit, we want to stop the trial early to prevent harm to patients on the treatment arm of the trial. Further, if the new treatment is simply ineffective – there is no scientifically important difference between it and the control – i.e. there are ethical and financial gains in stopping these trials early as well. If such futile trials are stopped early, then money and resources can be reallocated to researching additional treatments.

These ethical reasons for stopping a trial early govern the conduct of an interim analysis of a clinical trial. We describe the process in the context of a one-sided comparison for a new treatment to placebo (or standard of care). At each interim analysis, a summary measurement (the test statistic) of the data is calculated. This test statistic is then compared to predetermined critical values (boundaries) for this interim analysis. If the test statistic is large (indicating a highly effective treatment) the trial is stopped for efficacy. If the test statistic is small (indicating an ineffective treatment) the trial is stopped for futility. If the test statistic is between the two critical values (indicating neither large efficacy nor clear futility), the trial is continued to the next analysis.

These interim analyses require additional techniques to account for the correlation between the test statistics at each interim analysis. At the first interim analysis, a "standard" statistical analysis could be performed, and the test would be expected to have correct type I error and power, however if subsequent tests were also completed using standard techniques, the cumulative effect of repeated significance tests would lead to inflated type I

error, as quantified by Armitage et al. (1969).

Possible solutions to this problem were hinted at by Armitage et al. (1969), who proposed performing repeated significance tests at lower values that would allow the overall level of the test to remain fixed. Pocock (1977) and O'Brien and Fleming (1979) each subsequently examined other stopping boundaries for clinical trials that would also protect against inflated type I error. Many other group sequential stopping boundary designs have been suggested; these designs vary in terms of the likelihood of early stopping, which in turn has effects on average sample size and power.

We describe here the existing methodology for group sequential trials that relies on the assumptions discussed in the introduction, specifically normally distributed test statistics, known variance, constant variance, independent data, and single measurements of the primary outcome. Note that in a randomized study it will be sufficient to consider the one-sample model as the results can be generalized to a two-sample case when there is no confounding between the groups. The primary consideration is the joint distribution of the test statistic at all interim analyses.

2.1 Notation

In a sequential design, analyses occur at times t_j, where j indicates the number of the analysis to be performed ($j = 1, ..., J$). These analysis times, t_j can be specified and considered in either calendar time or statistical information time. The final analysis, at time t_J occurs at the end of the study. The statistic of interest at each of the analysis times will be subscripted to indicate the number of the analysis for this statistic.

2.1.1 Parameter Scales

For a generic trial, let Y_i be the response for the ith sample unit and let σ^2 be the variability of each sampling unit, and let N_j indicate the number of units obtained prior to the jth analysis time. Let μ be the unknown population treatment response and let μ_0 be the value of this parameter under the null hypothesis. There are several choices of scale for a statistic

computed at the $j$ith analysis:

$$\begin{aligned}
\text{Partial Sum} \qquad & S_j = \sum_{i=1}^{N_j} Y_i \\
\text{Sample Mean} \qquad & \bar{X}_j = \frac{S_j}{N_j} \\
\text{Z-Statistic} \qquad & Z_j = \sqrt{N_j}\frac{(\bar{Y}_j - \mu_0)}{\sigma} \\
\text{fixed sample P value} \qquad & P_j = 1 - \Phi(Z_j) = 1 - \int_{-\infty}^{Z_j} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du
\end{aligned}$$

It is possible to transform statistics on one scale to another with knowledge of N_j, σ, and μ_0. We note that because these are all monotonic transformations, such transformations preserve order so we are free to choose scales that are computationally and/or scientifically convenient as needed.

2.1.2 Boundaries

The boundaries at an interim analysis time are critical values used for testing if a trial should be stopped early. For the sake of convenience, we consider a scientific test of a greater alternative, such that larger values of the test statistic would be considered a more favorable scientific result. In this setting, the upper boundary corresponds to rejecting the null hypothesis of $H_0 : \mu \leq \mu_0$. Following the notation of Kittelson and Emerson (1999), this boundary is denoted by the letter d, with d_j indicating the value of this boundary at analysis time t_j. If the test statistic, i.e. S_j, at this analysis is greater than d_j, the trial is stopped for efficacy of the treatment (due to rejecting the null hypothesis).

As noted earlier, it may be prudent for financial and ethical reasons to stop early not only for efficacy, but also for futility. Thus, in a setting in which only a one-sided alternative is of interest, a second boundary could still be added to allow for stopping early for futility. In the example of a one-sided test of a greater alternative, this futility boundary would be the lower boundary, denoted by the letter a. If the test statistic, S_j is less than a_j, the trial is stopped early for futility, rejecting the possibility that $\mu \geq \mu_a$.

The continuation region of a trial (values for which the trial is not stopped at the interim analysis) is denoted C_j. In the case of the efficacy and futility for a one-sided hypothesis,

this continuation region is just (a_j, d_j). Note that the restriction that $a_J = d_J$ guarantees stopping at the final analysis.

A more detailed description of the choice of boundaries that are commonly used in clinical trials and the implications of such choices is deferred to section 2.3. For now, we consider boundaries on the partial sum scale only, and suppress the notation indicating the dependence of the boundary on the choice of scale (e.g. $C_{Sj} = C_j$).

2.2 Sampling Density and Independent Increments in Sequential Analyses

In order to maintain the type I error rate in a group sequential clinical trial, we need to ensure that

$$P(\bigcup_{j=1}^{J} S_j > d_j | \mu = \mu_0) = \alpha$$

To calculate this probability (as well as other quantities of interest), we need to know the sampling density of the test statistic.

In a fixed sample test ($J = 1$), the sampling density of the test statistic is given by standard statistical theory; under sufficient conditions for the central limit theorem (which we assume throughout this dissertation), a test statistic will have a normal density. For ease of notation, we consider the partial sum statistic and note that for a fixed sample test, its density is:

$$f(s; \mu) = \frac{1}{\sqrt{n}\sigma} \phi \left(\frac{s - n\mu}{\sqrt{n}\sigma} \right) \tag{2.1}$$

where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, the density for the standard normal distribution. In this case, $P(S_J > d_J | \mu = \mu_0)$ is given by integrating the density in equation 2.1 from d_J to ∞ when $\mu = \mu_0$.

In sequential analyses, the test statistic is now two dimensional, as the sampling density will depend on both the value of the statistic (S on the partial sum scale) and the analysis time at which this statistic was observed (denoted by M). The density of the test statistic at the first analysis time will be identical to that in the fixed sample setting (equation 2.1), and this fact will give the density of the test statistic if the trial is stopped at the first analysis. The density of the statistic at the second and other later analyses will be more

complicated as it will depend on the joint density of the test statistics and the boundaries used at the prior analyses.

$$P(\bigcup_{j=1}^{J} S_j > d_j | \mu = \mu_0) = \int_B f(s_1, s_2, ..., s_J | \mu_0) \; d\boldsymbol{s} \tag{2.2}$$

where B indicates that the density should be integrated over the stopping region for each statistic (analogous to integrating from d_1 to ∞ in the fixed sample case) and that the integration should take place only at later analysis times for the continuation regions of all prior analyses. For a single boundary (no stopping for futility), this restriction would correspond to $B = \{S_1 > d_1 \cup (S_1 \leq d_1 \;\&\; S_2 > d_2) \cup (S_1 \leq d_q \;\&\; S_2 \leq d_2 \;\&\; S_3 > d_3) \cup ...\}$.

Assuming that the central limit theorem conditions hold, the partial sums at each interim analysis, S_j will be asymptotically distributed multivariate normal with covariance matrix Σ. This general density is computationally quite challenging. However, if the contribution of the new data accrued between the time of the current and the previous analysis is independent of the previously acquired data, these data are said to have an independent increment structure and the above formula simplifies into a more manageable form. Fortunately, independent increments will be true in many situations, including the case in which all outcome measurements are independent of each other.

On the partial sum scale, independent increments implies that $S_k - S_j = S_{k-j}$ is independent of S_j. This implication leads to, for $k > j$:

$$\begin{aligned} Cov(S_j, S_k) =& Cov(S_j, S_j + S_{k-j}) \\ =& Cov(S_j, S_j) + Cov(S_j, S_{k-j}) \\ =& Var(S_j) + 0 \end{aligned}$$

On the sample mean scale, it leads to:

$$\begin{aligned} Cov(\bar{X}_j, \bar{X}_k) =& Cov(\frac{1}{N_j}S_j, \frac{1}{N_k}S_k) \\ =& \frac{1}{N_j N_k} Cov(S_j, S_k) \\ =& \frac{1}{N_k} Var\left(\frac{S_j}{N_j}\right) \\ =& Var(\bar{X}_k) \end{aligned}$$

In the group sequential setting, independent increments lead to much easier numerical integration of the sampling density in equation 2.2. It allows the integrations to be done using approximations to the standard normal cumulative distribution function (Armitage et al., 1969). This fact leads to their use in all standard statistical packages for group sequential designs. Fortunately, most common statistics based on the mean of independent data lead to independent increments.

Assuming independent increments, the density of the test statistic at a specific analysis times is formally recursively defined as follows.

$$p(m, s; \mu) = \begin{cases} f(m, s; \mu) & x \notin C_{sm} \\ 0 & \text{else} \end{cases}$$

where $f(m, s; \mu)$ is defined as:

$$\begin{aligned} f(1, s; \mu) =& \frac{1}{\sqrt{n_1}\sigma}\phi\left(\frac{s - n_1\mu}{\sqrt{n_1}\sigma}\right) \\ f(j, s; \mu) =& \int_{C_{s(j-1)}} \frac{1}{\sqrt{n_k}\sigma}\phi\left(\frac{s - n_k\mu}{\sqrt{n_k}\sigma}\right) f(k-1, u; \mu)\ du \quad j = 2, ..., m \end{aligned} \tag{2.3}$$

where again $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, the density for the standard normal distribution, and n_j is the size of the group accrued between successive analysis times (so $n_j = N_j - N_{j-1}$). We note that these densities are expressed most easily on the partial sum scale, but that conversions between the partial sum scale and other scales can be made as needed.

In a more general setting, the effect of interest is a parameter from an efficient score statistic. In these scenarios, the efficient score function, evaluated at θ_0, $U(\theta_0)$, is asymptotically normally distributed with mean $(\theta - \theta_0)I(\theta)$ and variance $I(\theta)$. From this setting,

we can implement all of the above with $\mu = (\theta - \theta_0)$, $\sigma^2 = 1$, and $N = I(\theta)$ (Whitehead, 1997; Emerson, 2000). Scharfstein et al. (1997) and Jennison and Turnbull (1997) noted that using the efficient statistic will always lead to independent increments. Scharfstein et al. noted further that inefficient statistics do not necessarily preclude independent increments. They observed that the Mantel-Haenszel estimate of the log hazard ratio in a two-sample proportional hazards survival model has an independent increment structure but is not efficient.

The use of the statistical information as the variance leads to "information based" monitoring of clinical trials, where the fractional amount of information present at an interim analysis relative to the final amount, $\Pi_j = I_j(\theta)/I_J(\theta)$, is paramount. This approach means that if the amount of relative information at each interim analysis can be correctly specified, and if the other assumptions hold, then existing group sequential methods will work. If the information growth is not correctly specified, these procedures may not lead to valid inference.

In many clinical trials, estimating the information growth for a given sequence of analysis times is straightforward. For example, in a setting with a single measurement per person and in which the parameter of interest is a function of the population mean with no mean-variance relationship, the information growth is proportional to the number of measurements: $\Pi_j = N_j/N_J$. In a study of survival outcomes using proportional hazards, the information growth is proportional to the number of events: $\Pi_j = D_j/D_J$.

2.3 Stopping Rules

The sampling density described above is directly influenced by the choice of boundaries. Indeed, the defining features of a particular group sequential design are the stopping rules to be implemented at each interim analysis. These stopping rules dictate whether or not a trial should continue to the next analysis, and therefore scientific and statistical considerations must be taken into account when designing a group sequential trial. Stopping a trial eliminates future data on potential adverse events and further statistical precision with which to estimate the treatment effect. For an efficacious drug, the estimated treatment effect is of critical importance to the scientific community, and thus the estimated treatment

effect at early stopping times must be carefully considered.

We noted earlier that the upper boundary is denoted by the letter d and the lower boundary by the letter a. For a one-sided alternative, only one of these boundaries is needed, though both may be used to allow early stopping for both efficacy and futility.

In some circumstances, clinical trials are designed to test for a two-sided alternative, such as a case in which two treatments are commonly used and the trial is used to determine if one is superior to the other. In this scenario, as in a fixed sample test, there will be two critical values at each analysis; the upper boundary corresponding to rejecting the null hypothesis of $H_0 : \mu \leq \mu_0$, and a lower boundary corresponding to rejecting the null hypothesis of $H_0 : \mu \geq \mu_0$. The upper boundary is similar to the one described above, and the lower boundary is denoted by the letter a. At analysis time t_j the trial is stopped if the test statistic, S_j, is less than a_j and the null is rejected in favor of the lower alternative.

For completeness, we note that it is possible to allow for stopping early for futility in a two-sided test as well. Here, intermediate boundaries b and c (with $a_j \leq b_j \leq c_j \leq d_j$ provide stopping if $S_j \geq b_j$ and $S_j \leq c_j$. These boundaries correspond to rejecting $\mu \leq \mu_b$ and $\mu \geq \mu_c$, respectively. If a trial stops because $b_j \leq S_j \leq c_j$, the conclusion is that $\mu_b \leq \mu \leq \mu_c$ because it has rejected the hypotheses that $\mu \leq \mu_b$ and $\mu \geq \mu_c$, which is interpreted as approximate equivalence in this two-sided test setting.

To generalize, stopping boundaries can be described by the continuation set, $C_j = (a_j, b_j] \cup [c_j, d_j)$, with $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$. If the test statistic is contained in the continuation set, C_j, the trial continues to the next analysis time, t_{j+1}. By defining C_J as the empty set the trial is assured of having no more than J analyses. Each boundary can be regarded as rejecting a specific one-sided hypothesis:

$$
\begin{aligned}
a: &\quad \mu \geq \mu_a \\
b: &\quad \mu \leq \mu_b \\
c: &\quad \mu \geq \mu_c \\
d: &\quad \mu \leq \mu_d
\end{aligned}
$$

In our standard example of a one-sided test of a greater alternative, $\mu_d = \mu_0$, as crossing this upper boundary corresponds to rejecting the null hypothesis. Crossing the futility boundary a corresponds to rejecting the hypothesis that $\mu \geq \mu_{alt}$ where μ_{alt} is a scientifically important effect.

Boundary scales can vary (as with test statistics) and it is straightforward to change between various possible scales. As noted earlier, at the design stage of a clinical trial, potential boundaries should be considered on a scientifically relevant scale to ensure that the conclusions that would be drawn from stopping at each interim analysis are of scientific importance. From a statistical perspective, the boundaries chosen should have correct type I error under the null and should have the desired power for a specific alternative.

Given the constraint of type I and type II errors at the nominal level and the timing of interim analysis, there still are many possibilities for appropriate stopping boundaries. Restricting possible boundaries to be monotonic on the sample mean scale (i.e. $d_{j-1} \leq d_j \forall j$ seems sensible, however this criterion still leaves many possibilities. As long as the boundaries fullfill the basic type I and type II error constraint (and are monotonic) boundaries can differ greatly in terms of general operating characteristics for a clinical trial. Among all designs that have 97.5% power for a specific alternative, there can be differences in the maximal possible sample size (N_J) and the average sample number (ASN) for specific values (such as under the null and the specified alternative). These trade offs in general stem from the timing of interim analyses and the amount of so-called "conservatism" at early analyses. A higher probability of stopping at early analyses (lack of "conservatism") such as is true when comparing a Pocock (1977) design to an O'Brien-Fleming (1979) design, generally leads to a higher maximal sample size but lower ASN for several alternatives. The lower ASN means that the design is more efficient for that specific alternative.

One way to describe specific design boundaries is the amount of error being "spent" at each interim analysis. This approach was initially proposed as a type I error spending function by Lan and DeMets (1983). For a design with a null hypothesis that $\mu \leq \mu_0$, the

amount of type I error spent at each interim analysis is given by:

$$\alpha_j = P(S_j \geq d_j, \bigcap_{k=1}^{j-1} S_k \in C_k | \mu = \mu_0)$$

On this error spending scale, differences in conservatism at early analyses are easily seen as differences in the amount of type I error spent at early analyses. Designs with a higher amount of type I error spent (less conservatism) are willing to declare efficacy for a lower estimate of treatment effect at an early analysis than those with less type I error spent (greater conservatism). How much early conservatism makes sense depends on the scientific context of the study.

Pampallona et al. (1995) extend this concept of error spending to designs with early stopping for futility as well. In this case, the futility boundary is rejecting the alternative of interest (e.g. the one with 97.5% power) and thus at each interim analysis such a boundary leads to a certain amount of type II error. As with spending the type I error, futility boundaries can differ by how much type II error is spent at each interim analysis relative to the final. The amount of early conservatism for a futility boundary should be decided by scientific context.

2.3.1 Families of Designs

Pocock (1977) used stopping boundaries to account for multiple analyses by performing a level α' fixed sample test at each of the J analyses, where appropriate values for α' maintain the type I error. This approach leads to boundaries that are constant on the normalized Z-statistic scale. The O'Brien-Fleming (1979) design similarly accounts for the multiple analyses, but does so by maintaining a constant threshold on the standardized partial sum scale. Wang and Tsiatis (1987) extended these designs to a more general one-parameter family. Using their notation, the two-sided boundary rejects if the standardized partial sum crosses a boundary such that for equally spaced analyses:

$$|S_j| \geq \Gamma(\alpha, K, \Delta) j^{\Delta}; \quad j = 1, ..., K$$

The $\Gamma(\alpha, K, \Delta)$ value is found by numerical search and Δ is a user-specified parameter to control the shape of the boundary at interim analyses. These designs can be thought of as

moving smoothly from O'Brien-Fleming to Pocock in terms of the differences between the two in "conservatism" at early analysis times.

Kittelson and Emerson (1999) generalized these results into a unified family of designs which further allow for a greater variety of choices in boundary shape and hypotheses. They note that the boundaries on the sample mean (MLE) scale for common group sequential tests can be written in the form:

$$a_j = \mu_a - G_a f_a(\Pi_j)$$
$$b_j = \mu_b + G_b f_b(\Pi_j)$$
$$c_j = \mu_c - G_c f_c(\Pi_j)$$
$$d_j = \mu_d + G_d f_d(\Pi_j)$$

In these equations, the function f_* defines a boundary shape function for the amount of early conservatism as a function of the fractional amount of information ($\Pi_j = \frac{I(\theta_j)}{I(\theta_J)}$). The values of G must be found by a computer search such that the boundaries formed provide a correct level α test of the null hypothesis and a specified power level for a specific alternative hypothesis. The specification of the alternative hypotheses follows as in the previous discussion. For the purposes of this work, we are mostly considering cases of a one-sided hypothesis with either stopping only for the alternative (a one boundary design) or for either the alternative or the null (a two boundary design). Such designs mean that we will be mostly interested in the a_j and d_j boundaries only. However, there should be no difficultly in extending the results to a more general case.

The boundary shape function will specify the amount of "conservatism" at early analyses and in turn determine the amount of efficiency of the sequential design. The Pocock design for a two-sided test uses $f(\Pi_j) = \Pi_j^{-0.5}$ for the a_j and d_j boundaries. The O'Brien-Fleming design for a two-sided test similarly uses $f(\Pi_j) = \Pi_j^{-1}$. The Wang and Tsiatis one-parameter family uses $f(\Pi_j) = \Pi_j^{-P}$, where P is a user-specified parameter. Another proposed boundary – the triangular test of Whitehead and Stratton (1983) – uses $f(\Pi_j) = 1 + \Pi_j^{-1}$. The

unified family approach allows for all of these designs and additional flexibility by using:

$$f_*(\Pi_j) = A_* + \Pi_j^{-P_*}(1 - \Pi_j)^{R_*} \tag{2.4}$$

Values A, P, and R are user-specified shape parameters. Setting $R = A = 0$ gives the Wang and Tsiatis one-parameter family, with $P = 0.5$ (Pocock) and $P = 1$ (O'Brien-Fleming) as special cases. In general, the value P generally dictates the level of conservatism at early analysis times, and the value R dictates the level of non-conservatism at early analysis times, though there are relationships between the values of A, P, and R that make more general statements difficult.

In general, we will consider the Pocock and O'Brien-Fleming boundaries as examples of boundaries with less and more conservatism at early analysis times. In general, because they are more likely to stop earlier, Pocock designs tend to be more efficient in terms of average sample number than O'Brien-Fleming designs. However, for a fixed maximal sample size, O'Brien-Fleming boundaries will be more powerful than Pocock due to the early conservatism.

2.4 Post-Trial Inference

After a group sequential design reaches a conclusion, i.e. after a stopping boundary has been crossed, final inference is desired. As with most statistical tests, we desire a point estimate (on some scientifically relevant scale) and confidence intervals. Here again, the changes in the sampling density due to the sequential design cause some revision from traditional statistical inference. The maximum likelihood estimate at the time of stopping is not unbiased for the truth, due to the sequential design. Various methods exist to account for this bias: some methods such as the Rao-Blackwell adjusted unbiased estimate (RBUE) correct for the bias completely and thus give an unbiased estimate; others, such as the median unbiased estimate (MUE) and the bias adjusted (BAM) mean Whitehead (1986)

correct for some of the bias in an effort to minimize mean squared error.

$$RBUE\ (\ddot{\mu}) \qquad \ddot{\mu} = E\left(\frac{S_1}{N_1}|(M,S) = (m,s)\right)$$

$$MUE\ (\tilde{\mu}) \qquad P\big[(M,S) > (m,s); \tilde{\mu}\big] = 0.5$$

$$BAM\ (\check{\mu}) \qquad E\left(\frac{S}{N_M}; \mu = \check{\mu}\right) = \frac{s}{N_m}$$

All three statistics rely on the correct version of the sampling density that accounts for the interim analyses. The MUE and BAM also are dependent on the value of δ and thus will behave differently under a mean-variance relationship. Additionally, the MUE relies on an ordering of space as well, as it relies on defining the probability of the two-dimensional statistic (M, S) being greater than a specific value (in this case, the observed value (m, s)). Here different authors have explored different orderings of the sample space. Tsiatis et al. (1984) investigated an analysis time ordering, in which statistics that were observed earlier and caused the trial to stop are always more extreme than those observed later. Under this analysis time ordering and comparing only statistics that caused the trial to stop, $(M = 1,\ S/N_M = -5) < (M = 3,\ S/N_M = -10)$ and $(M = 1,\ S/N_M = 5) > (M = 3,\ S/N_M = 10)$. Emerson and Fleming (1990) explored ordering based on the sample mean instead, such that more extreme observed values of the sample mean are always considered more extreme, regardless of when the trial was stopped. Under the sample mean ordering, $(M = 3,\ S/N_M = -10) < (M = 1,\ S/N_M = -5)$ and $(M = 3,\ S/N_M = 10) > (M = 1,\ S/N_M = 5)$ regardless of stopping boudaries. A third ordering, the likelihood ratio ordering, was suggested by Chang and O'Brien (1986). Under this ordering, statistics are considered more extreme if they lead to a more extreme value of the likelihood ratio statistic:

$$\frac{p(M_1, S_1|\mu = \hat{\mu}_1)}{p(M_1, S_1|\mu_0} > \frac{p(M_2, S_2|\mu = \hat{\mu}_2)}{p(M_2, S_2|\mu_0)}$$

where $\hat{\mu}$ is the maximum likelihood estimate for μ given (M, S).

The construction of 95% confidence intervals also relies on the correct sampling density from a group sequential design and the choice of ordering. We first note that with a group sequential sampling density, a 95% confidence "interval" need not actually be a true interval, and thus is a 95% confidence set. This confidence set is defined as all values of the parameter

θ for which the observed statistic (m, s) would not be "unusual", specifically the probability of observing a statistic less the observed value if the truth were θ is between $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$.

$$CI\{(M, S) = (m, s)\} = \left\{\theta : \frac{\alpha}{2} \leq P((M, S) \leq (m, s)|\theta) \leq 1 - \frac{\alpha}{2}\right\} \tag{2.5}$$

This definition holds for all scenarios, however it is particularly useful to think of 95% confidence sets in this way with group sequential designs because it shows how they can be constructed. For each true value of the parameter of interest, θ, its inclusion in the confidence set is determined by a so-called "inverted hypothesis test" in which the possible value θ is included in the set if the observed value $\hat{\theta}$ can not be ruled out with 95% confidence if the true value were θ. Here again, we note that the $\leq$ criterion must be defined for a sequential trial and is determined by how the outcome space is ordered.

In the clinical trials setting, it is possible that confidence sets may not be true intervals, a fact that is unsettling scientifically. Confidence intervals would be guaranteed if there were stochastic ordering of the sample space. In a stochastically ordered space, X is stochastically less than Y if $P(c < X) < P(c < Y)$ for all c. Stochastic ordering was proven for the case of a fixed variance in clinical trials for the analysis time and sample mean (Emerson and Fleming, 1990). However, it has never been proven for the case of likelihood ratio ordering.

The specific setting of longitudinal group sequential trials, is covered in the following chapter, after we present a brief background of longitudinal analysis.

Chapter 3

BACKGROUND – LONGITUDINAL DATA AND LONGITUDINAL TRIALS

In fixed sample settings, longitudinal data present challenges not faced in a single measurement setting. Many of these challenges, such as correlated data, are amplified in the group sequential framework. Additional challenges posed by repeated measurements on individuals during the duration of the study are unique to sequential designs.

We consider the case in which the scientific quantity of interest is a change over time. In our model, this quantity is the slope parameter β_1 from a regression equation $E(Y|X) = \beta_0 + \beta_1 x$, where x represents study time from randomization.

3.1 Least Squares Regression

Throughout we assume that we are interested in the linear contrast over time, which leads to the model:

$$E(Y|X = x) = \beta_0 + \beta_1 x, \tag{3.1}$$

where β_1 is the parameter of interest. We will let V denote the true covariance matrix of the observations Y. We let X denote the design matrix, as in standard linear model notation.

We first consider the case of independent, homoscedastic data, so that $V = \sigma^2 I$. Then, assuming $X^T X$ is nonsingular, using standard methods for linear models gives:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{3.2}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{3.3}$$

From this equation, we know that the variance of the estimate for the parameter of interest is

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{nVar(x)}, \tag{3.4}$$

where σ^2 is the variance of the residuals, n is the total number of observations and x is the study time from randomization.

Ordinary least squares regression can be modified to accommodate heteroscedastic and correlated data if the covariance structure of the observations is known (or assumed to be known). We will use "weighted least squares" (WLS) regression to denote the case in which independence is still assumed, but the diagonal elements of the covariance matrix need not be constant - the situation of heteroscedastic data. We will use "generalized least squares" (GLS) to denote the case in which neither independence nor homoscedasticity is assumed. In both cases, for an assumed covariance matrix, W, the equations from OLS are modified to estimate $\hat{\beta}$ with different weights for the observations.

$$\hat{\beta} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$$
$$Var(\hat{\beta}) = (X^T W^{-1} X)^{-1} X^T W^{-1} V W^{-1} X (X^T W^{-1} X)^{-1}$$

All choices of W will yield unbiased estimates for $\hat{\beta}$, but they will differ in the standard error of the estimate. The Gauss-Markov theorem shows that among all linear unbiased estimators, the estimator from WLS with known, correct weights ($W^{-1} = V^{-1}$) is the most efficient and thus it is the best linear unbiased estimator (BLUE). If the known weights are used, the equation for the variance of the estimates simplifies to:

$$Var(\hat{\beta}) = (X^T W^{-1} X)^{-1}$$

We note that although this is the classic linear model, we are truly interested in the linear contrast over time, even if the data are not perfectly linear. Such a contrast is of scientific interest, and furthermore, in the clinical trial setting, analyses must be prespecified to satisfy regulatory authorities. Therefore, if a linear model was specified *a priori*, the model will be used even if the true data are not perfectly linear.

3.2 Generalized Estimating Equations

Liang and Zeger (1986) proposed Generalized Estimating Equations (GEE) as a alternative method to random effects models for estimating slopes with correlated longitudinal data.

Unlike random effects models, GEE estimates the marginal effect; it gives the population average effect rather than the average from an individual "typical" in the population.

GEE is implemented with estimating equations and accounts for possible correlation in the observations with a so-called "working" covariance matrix, W. As before, denote the true covariance matrix as V. Standard errors for the estimates are then obtained using the Eicker-Huber-White sandwich method (Liang and Zeger, 1986) . For a linear model, and assuming $Var(Y) = \sigma^2 V$, this method leads to:

$$\begin{aligned}\hat{\beta} =& (X^T W^{-1} X)^{-1} X^T W^{-1} Y \\ Var(\hat{\beta}) =& \sigma^2 (X^T W X)^{-1} X^T W^{-1} V W^{-1} X (X^T W X)^{-1}\end{aligned}$$

GEE is nearly always consistent for the true parameters β, regardless of the choice of working covariance (departures will be discussed later). Assuming adequate sample size, the sandwich estimates for the standard errors, when turned into confidence intervals, produce appropriate coverage probabilities, regardless of the choice of working covariance matrix. However, this choice does impact the asymptotic efficiency of the estimate. If the working covariance matrix is of the same form (and is asymptotically consistent for) the true form of the covariance matrix ($W \rightarrow_p V$), GEE will be asymptotically efficient. This result is analogous to the case of WLS using weights that are asymptotically consistent for V^{-1}.

In all cases except working independence with a linear link, the GEE equations must be solved iteratively. First, an estimate of $\hat{\beta}$ is made (often using working independence). Then the estimated $\hat{\beta}$ is used to generate an estimate for $\hat{\rho}$, where $\hat{\rho}$ is the vector of parameters needed to estimate the working covariance, $W(\hat{\rho})$. After an estimate for $W(\hat{\rho})$ is obtained, this new working covariance matrix is used to generate another estimate of $\hat{\beta}$. This process continues until updates no longer produce changes in the parameters (up to a certain tolerance).

Various methods exist for estimating the parameter ρ in working covariance matrices. The most common is that proposed by Liang and Zeger (1986), which simply uses a method of moments estimator for the parameter. A scale parameter ϕ (in the case of the linear model this scale parameter is the inverse of the constant error variance, $\frac{1}{\sigma^2}$) is estimated first, using Pearson residuals. In the case of the linear model, these residuals are given by $\hat{r}_{ij} = y_{ij} - \hat{y}_{ij}$.

For K individuals with n_i measurements per person, ϕ is estimated by:

$$\hat{\phi}^{-1} = \left(\sum_{i=1}^{K} \sum_{j=1} n_i \hat{r}_{ij}^2 \right) / (N - p)$$

For the linear model, this estimate is identical to the estimate of $\hat{\sigma}^2$ from S^2. With an estimate for $\hat{\phi}$, the parameter(s) of the working covariance matrix can be estimated. Following Liang and Zeger, for an exchangeable structure where $corr(y_{ij}, y_{ij'}) = \rho$ for all $j \neq j'$, ρ can be estimated as follows:

$$\hat{\rho} = \frac{\phi \sum_{i=1}^{K} \sum_{j>j'} \hat{r}_{ij} \hat{r}_{ij'}}{\sum_{i=1}^{K} \frac{n_i(n_i - 1)}{2} - p}$$

For an autoregressive with order one (AR(1)) structure, $corr(y_{ij}, y_{ij'}) = \rho^{|j-j'|}$. In this circumstance, ρ can be estimated by regression with a log-link and no intercept term:

$$\begin{aligned} log(E(\hat{r}_{ij}\hat{r}_{ij'})) &= |j - j'| * \gamma \\ log(\rho^{|j-j'|}) &= |j - j'| * \gamma \\ \rho &= e^{\gamma} \end{aligned}$$

Additional research on the GEE has pointed out some limitations with consistency and efficiency of the method. Efficiency concerns with GEE have been studied by several authors. Zhao et al. (1992) illustrated cases in which when the correlation is high, using working independence can lead to significant losses in efficiency. Other authors have noted situations in which using working independence does not appear to lead to substantial loses in efficiency Lipsitz et al. (1994). Fitzmaurice (1995) and Mancl and Leroux (1996) demonstrated cases in which covariate variation within clusters can lead to efficiency losses when using working independence, even with only moderate correlation. Wang and Carey (2003) noted that differences in the estimation techniques for the parameter ρ in the working covariance model can impact the relative efficiency of such designs. They note that using working AR(1) with a "Gaussian estimation" procedure leads to improved efficiency over a moment estimator similar to that proposed by Liang and Zeger. Finally, some authors have

suggested using an "unstructured" working covariance in which all correlation parameters are estimated from the data (Gange and DeMets, 1996).

As previously noted, Pepe and Anderson (1994) showed that although GEE is always consistent when using working independence, in certain circumstances with time-varying covariates, GEE need not be consistent when using other working covariance matrices. Crowder (1995) gave an example with correlated binary data in which using working AR(1) with truly exchangeable data led to inconsistent estimation of the true parameters. In the setting of this dissertation with the study times fixed by design, consistency will hold regardless of the choice of working covariance matrix. However, this condition will not necessarily be true in all longitudinal trials, and the potential inconsistency may be motivation for using working independence even when it may be inefficient.

3.3 Information Growth in Longitudinal Trials

In sequential clinical trials with longitudinal data, special consideration must be given to the information growth of such trials. There are several potential issues with estimating the information growth in a longitudinal trial. The first is that the information for the slope parameter increases dramatically as the spread of measurements in the predictor space increases, and the estimated information growth must account for this component of the information as well as the increasing total number of measurements. Secondly, the choice of how to handle the possible correlation of the data can lead to nonmonotonic information growth. Finally, a mean-variance relationship can lead to a different final total information under different alternatives, making the estimation of the information growth difficult.

Consider first the case in which all of the data are independent. Here there are two potential issues with the information growth, which we consider fully as an introduction to potential issues that will be encountered later with heteroscedastic and correlated data. The first is that, as noted in equation 3.3, the variance of the estimate of interest is given by $\frac{\sigma^2}{nVar(x)}$. Thus, the true fraction of information is $\pi_j = \frac{I(\hat{\beta}_j)}{I(\hat{\beta}_J)} = \frac{n_j Var_j(x)}{n_J Var_J(x)}$. Wu and Lan (1992) noted that using a "typical" estimate of the information growth as a ratio of the sample sizes, $\frac{n_j}{n_J}$, overestimates the true information in this setting, because $\frac{Var_j(x)}{Var_J(x)}$ must be less than or equal to 1, so $\frac{n_j}{n_J} \geq (\frac{n_j}{n_J})(\frac{Var_j(x)}{Var_J(x)})$. We will illustrate consequences of this

overestimation in chapter 4.

The GEE method has been previously studied in group sequential designs by Wei et al. (1990) and Lee et al. (1996), among others. Wei et al. studied the use of GEE with repeated measurements in a trial, for example with repeated cholesterol measurements on individuals enrolled in a trial. Unlike our setting, they were not interested in the change in cholesterol over time; rather they were interested in accounting for the correlation on measurements within an individual to evaluate a difference in mean cholesterol between treatment groups.

Lee et al. studied the case of using GEE to estimate a change over time in a group sequential design. They showed that the independent increment structure is true if the working covariance matrix is consistent for the true covariance matrix. This result is consistent with later theory that showed that the use of an efficient statistic must lead to an independent increment structure, as noted previously (Jennison and Turnbull, 1997; Scharfstein et al., 1997). Lee et al. further speculate that using a working covariance matrix that is not consistent for the truth but does converge to some (incorrect) matrix will lead to "nearly" independent increments.

Longitudinal data with correlated observations have the more general problem of nuisance parameters due to the need to estimate correlation (and variance). The problem of estimating nuisance parameters has been studied for group sequential trials by Burington and Emerson (2003). They noted that imprecision of the estimated nuisance parameters can lead to error spending boundaries that do not reflect the true known proportionate information available at each analysis, while boundaries constrained on other scales will not necessarily adhere to the desired boundary shape function. Several authors have further conjectured that the imprecision inherent in estimating within group variances or baseline event rates at the earliest of interim analyses might lead to a spurious appearance of nonmonotonic information growth during the monitoring of a study (Scharfstein et al., 1997; Burington and Emerson, 2003). They speculate that such situations are probably rare in practice, due to the relatively large increments of information typically accrued between successive analyses: the monotonic increase in available data is expected to overwhelm the potential nonmonotonicity in the estimates of the nuisance parameters across the analyses.

We explore the impact of assuming homoscedasticity with heteroscedastic data in chap-

ters 5 and 6. In chapter 7, we explore using working independence with correlated data, both with and without heteroscedasticity.

Chapter 4

INDEPENDENT DATA WITH NO HETEROSCEDASTICITY

This chapter illustrates potential problems with standard group sequential trials when the information growth is potentially misspecified due to each observation not contributing equally to the information about the statistic. This problem was first noted by Wu and Lan (1992) with longitudinal data. We use the case of longitudinal, independent, homoscedastic data to illustrate consequences of misspecified information growth. We then explore the potential problems if the constant variance cannot be correctly estimated due to model misspecification. If the data are not exactly linear, the model-based standard errors will not consistently estimate the constant variance, and we briefly consider the consequence of this situation in a group sequential setting.

4.1 Homoscedastic, Linear Data

As mentioned previously, the true information growth of independent, homoscedastic, linear data is easily estimated and can be planned for in a sequential clinical trial. For illustrative purposes, we compare the true information growth in this setting, $\pi_j = \frac{\sigma^2/n_j Var_j(x)}{\sigma^2/n_J Var_J(x)}$, to the naive estimate, $\pi_j = \frac{n_j}{n_J}$. As the difference in the two estimates depends on the difference in the variation of the predictor space at the interim compared to final analysis times ($\frac{Var_j(x)}{Var_J(x)}$), we expect that designs in which accrual is short relative to follow-up will see the greatest difference in the true information growth compared to the naive estimate. In designs where accrual is long relative to follow-up, the two estimates should be similar. In an extreme case of very slow accrual relative to follow-up, the change in the variance of the predictor space would be minimal with each added measurement and nearly all of the increase in information would be due simply to the increasing number of subjects.

For an example, consider the case of 10 measurements made over time, with one measurement at baseline and one at each of 9 months thereafter. Figure 4.1 shows the true

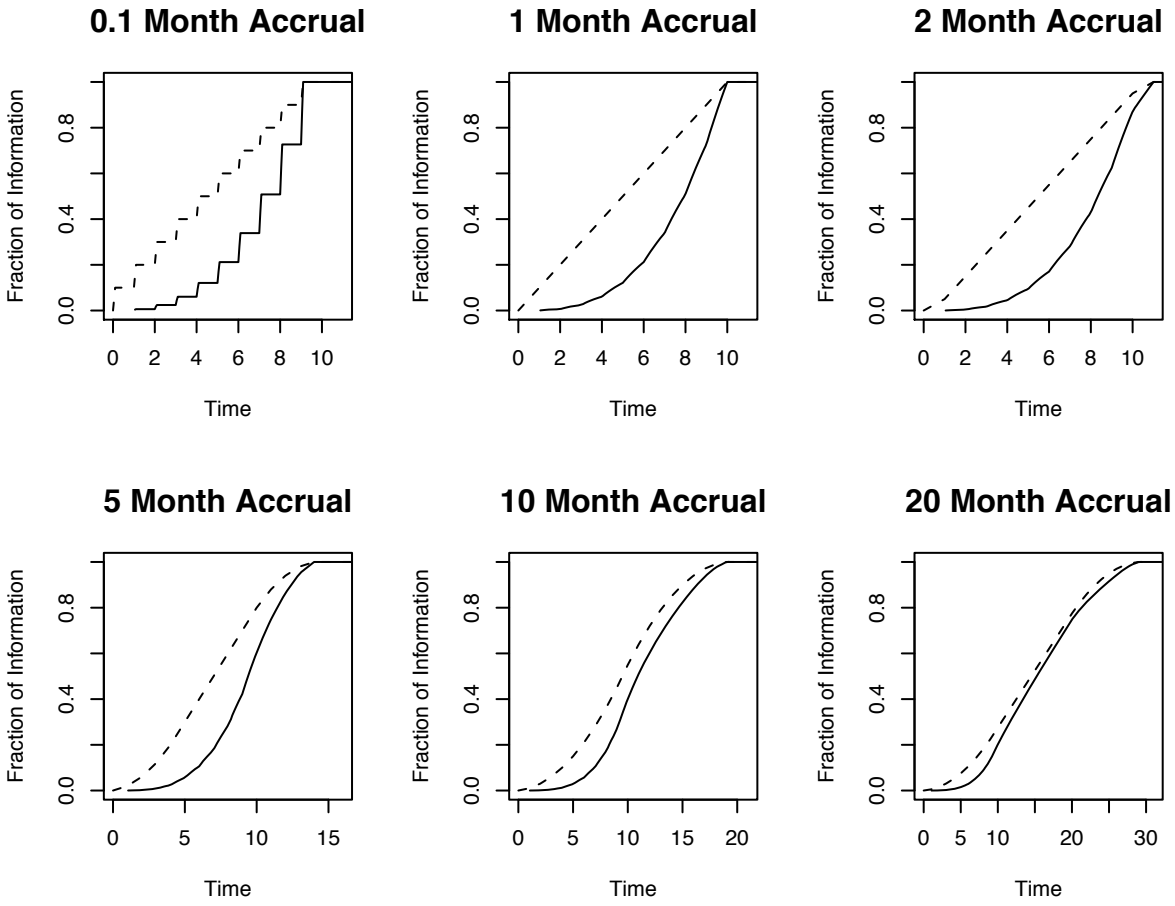

Figure 4.1: Plots showing the true information growth (solid line) relative to the information growth that would be estimated from the fraction of the total number of measurements (dashed line). In all cases, estimating the IG by the number of measurements overestimates the true information.

information growth for this setting, with varying accrual patterns. This figure demonstrates that in all cases a naive approach to estimating the information growth as simply a fraction of the sample sizes overestimates the true information, as noted previously. In cases with short accrual relative to the length of follow-up, this overestimation is dramatic. For example, with 2 month accrual and four analyses equally spaced in calendar time, the first analysis (after 2.75 months) has only 1.5% of the information that would be present at the final analysis. In contrast, the naive assumption that the information grows proportional to the number of measurements would estimate that this first analysis takes place with 22.5% of the final information. Similarly, at the second analysis, the true information is only 14% of the final, yet the naive estimate would be 50%, and at the third analysis the true information is 48% of the final but the naive estimate would be 77.5%. Table 4.1 shows the number of measurements at each study time for each interim analysis in this setting. The potential consequences of overestimating the information growth over the course of the study are discussed below.

Table 4.1: Distribution of observed study times at each interim analysis with 2 month accrual. The proportion of the final amount at each study time is given.

Analysis	0	1	2	3	4	5	6	7	8	9	True IG	Naive IG
1	1.00	0.87	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.015	0.225
2	1.00	1.00	1.00	1.00	0.75	0.25	0.00	0.00	0.00	0.00	0.14	0.50
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.62	0.13	0.00	0.48	0.775
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1

4.2 Consequences in Sequential Designs

The issue of correctly specifying the information growth at the design stage has been previously explored (Proschan et al., 1992; Jennison and Turnbull, 2000). Applied to the longitudinal setting, the consequences of the misspecified information growth depend on how badly the information growth is misspecified and what mechanisms were pre-specified

to account for different true information growth. These authors have noted that slight variations in the true information growth from the planned analysis did not have a large impact on the boundaries or inference. In the longitudinal setting, their observation most closely corresponds to the case of long accrual relative to follow-up where the naive estimate based only on the number of measurements is likely to be a close approximation to the true information growth.

4.2.1 Boundary Design

We consider four approaches to group sequential design and analysis in this setting to illustrate potential difficulties associated with incorrectly specified information growth. For each approach, we consider (1) how the boundaries for the group sequential design were constructed, which will be determined by the assumed information growth at the time of each analysis, (2) how the design is to be used during the course of the trial, which will include the scale on which interim statistics will be evaluated and if the design boundaries are fixed or can be adjusted during the trial, and (3) how the alternative with assumed 97.5% power was calculated.

As an example, we consider the case of uniform accrual over 2 months, measurements at baseline and months 1-9, and four analyses equally spaced in calendar time. We choose the maximal sample size to be fixed in this setting and allow the power to vary between designs. All designs are constructed to have fixed 0.025 type I error rate.

The first approach is to use boundaries that are fixed on the sample mean scale and based on either the incorrect, naive information growth or the correct information growth. Constraining on this scale implies a belief that not only is the information growth specified correctly, but that the σ^2 is also specified exactly. Specifically, for this approach – fixed boundaries on the sample mean scale – we have that:

- Boundaries are constructed using the specified information growth, which might be either the naive information growth or the true information growth (see table 4.2).

- At each interim analysis, the sample mean from the interim analysis is compared to the boundaries designed on the sample mean scale and the decision is made to

continue or stop the trial. The boundaries constructed in the design phase remain fixed throughout the study.

- The alternative with 97.5% power for comparison was calculated from the design with the correctly specified information growth.

The second approach uses boundaries that are fixed on the z-statistic scale based on either the incorrect, naive information growth or the correct information growth. For this approach – fixed boundaries on the z-statistic scale – we have that:

- Boundaries are constructed using the specified information growth, either the naive information growth or the true information growth (see table 4.2).

- At each interim analysis, the z-statistic from the interim analysis is compared to the boundaries designed on the z-statistic scale and the decision is made to continue or stop the trial. The boundaries constructed in the design phase remain fixed throughout the study.

- The alternative with 97.5% power for comparison was calculated from the design with the correctly specified information growth.

For both of these approaches, we consider boundaries that allow for early stopping for the alternative only (one-sided) and for both the alternative and the null (two-sided). The fixed boundaries for an O'Brien-Fleming design that result from the incorrect and correct IG for our example are shown in table 4.2.

The third approach we consider is that of constrained boundaries (Burington and Emerson, 2003). This approach is intended to allow for recalibration of the stopping boundaries during a trial to maintain type I error rate at the nominal level. The constrained boundary approach allows for boundaries to be adjusted at all interim analyses based on the observed information growth. For our example, we consider the case in which the boundaries at the first three analysis times are fixed on the z-statistic scale, but at the final analysis the boundaries are adjusted to account for the true information growth that was observed over the course of the trial. Thus, for this constrained boundary approach we have:

Table 4.2: Boundaries using incorrect (naive) and correct information growth fixed on the sample mean and z-statistic scales for an O'Brien-Fleming design.

Sample Mean Scale						
	Naive IG			Correct IG		
Analysis #	IG	a	d	IG	a	d
1	0.225	-4.90	8.91	0.015	-127.5	131.4
2	0.500	0.00	4.01	0.140	-10.14	14.08
3	0.775	1.42	2.59	0.480	-0.17	4.11
4	1.000	2.01	2.01	1.000	1.97	1.97
Z-Statistic Scale						
1	0.225	-2.32	4.23	0.015	-15.61	16.09
2	0.500	0.00	2.84	0.140	-3.79	5.27
3	0.775	1.25	2.28	0.480	-0.11	2.85
4	1.000	2.01	2.01	1.000	1.97	1.97

- Boundaries are constructed using the specified information growth, either the naive information growth or the true information growth.

- At each interim analysis, the z-statistic from the interim analysis is compared to the boundaries designed on the z-statistic scale and the decision is made to continue or stop the trial. The boundaries constructed during the design phase are fixed for the first three analysis, but then the boundaries for the final analysis are recomputed to account for the true information growth and maintain type I error rate.

- The alternative with 97.5% power for comparison was calculated from the design with the correctly specified information growth.

The fourth and final approach we consider is the error spending approach of Lan and DeMets (1983) which spends a certain amount of error at each interim analysis and thus it does not exceed the nominal level. The amount of error to be spent at each analysis is determined based on the information growth (either the naive or correct). However, the actual boundaries for each interim analysis are computed based on the observed standard error of the statistic and thus are computed to "spend" the correct amount of type I or type II error regardless of the assumed information growth. We note that the efficacy boundary is constructed to spend the correct amount of type I error and the futility boundary is constructed to spend the correct amount of type II error for the alternative with presumed 97.5% power. For the error spending approach, we have:

- An error spending function is constructed using the incorrect information growth to be consistent with the O'Brien-Fleming or Pocock design for this information growth. The error spending functions for the type I and type II error are identical in this case.

- At each interim analysis, the error spending functions are used to calculate a z-statistic critical value based on the observed true sampling density at that interim analysis that will spend the specified amount of type I or type II error. The final analysis critical value is calculated based only on the amount of remaining type I error to be spent,

as with a constrained maximal sample size it is impossible to maintain both the type I and type II error rate.

- The alternative with assumed 97.5% power for use in calculating the type II error spending was taken from the design with the incorrect, naive information growth. The alternative for comparison is the alternative with 97.5% power under the correctly specified information growth.

Boundaries from the constrained boundary and error spending function approach for an O'Brien-Fleming design originally constructed with the naive information growth are shown in table 4.3.

Table 4.3: Boundaries using the constrained boundary and error spending approach with the naive information growth and an O'Brien-Fleming design.

Z-Statistic Scale						
	Constrained Boundaries			Error Spending		
Analysis #	IG	a	d	%Error Spent	a	d
1	0.225	-2.324	4.226	0.000	-3.735	4.226
2	0.500	0.000	2.835	0.092	-1.336	2.836
3	0.775	1.252	2.277	0.488	0.468	2.310
4	1.000	1.876	1.876	1.000	2.105	2.105

4.2.2 Boundary Evaluation

Table 4.4 shows the dramatic increase in type I error rate when using the naive information growth estimates in this setting of boundaries fixed on the sample mean scale. The boundaries are constructed to maintain a fixed 0.025 error rate, yet the single boundary type I error rate is 0.32 using the Pocock boundary and 0.21 using an O'Brien-Fleming boundary. For two boundary designs, the type I error rates are 0.29 and 0.20, respectively (table 4.5).

Table 4.4: Stopping probability for the alternative (SP_{alt}) at each of the four analyses under the null and the alternative with 97.5% power using Pocock and O'Brien-Fleming (OBF) stopping boundaries fixed under the naive information growth on the sample mean scale.

One Boundary - Under Null								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.000	0.135	0.000	0.000	0.000	0.252	0.000	0.007
2	0.000	0.044	0.000	0.000	0.000	0.055	0.000	0.007
3	0.000	0.019	0.000	0.002	0.000	0.011	0.000	0.006
4	0.793	0.009	0.975	0.023	0.680	0.002	0.975	0.005
Total	0.793	0.207	0.975	0.025	0.680	0.320	0.975	0.025
One Boundary - Under Alternative								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.000	0.267	0.000	0.000	0.000	0.466	0.000	0.028
2	0.000	0.309	0.000	0.000	0.000	0.292	0.000	0.192
3	0.000	0.298	0.000	0.449	0.000	0.178	0.000	0.512
4	0.017	0.109	0.025	0.526	0.010	0.054	0.025	0.243
Total	0.017	0.983	0.025	0.975	0.010	0.990	0.025	0.975

Table 4.5: Stopping probability (SP) for the null or alternative at each of the four analyses under the null and the alternative with 97.5% power using Pocock and O'Brien-Fleming (OBF) stopping boundaries fixed under the naive information growth on the sample mean scale.

Two Boundary - Under Null								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.274	0.138	0.000	0.000	0.488	0.274	0.033	0.007
2	0.279	0.039	0.000	0.000	0.160	0.027	0.251	0.007
3	0.209	0.013	0.455	0.002	0.044	0.003	0.547	0.006
4	0.045	0.003	0.520	0.023	0.004	0.000	0.144	0.004
Total	0.807	0.193	0.975	0.025	0.696	0.304	0.975	0.025
Two Boundary - Under Alternative								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.139	0.271	0.000	0.000	0.265	0.498	0.007	0.033
2	0.041	0.273	0.000	0.000	0.024	0.165	0.007	0.251
3	0.014	0.209	0.002	0.455	0.002	0.042	0.006	0.547
4	0.004	0.048	0.023	0.520	0.000	0.003	0.004	0.144
Total	0.199	0.801	0.025	0.975	0.291	0.709	0.025	0.975

In this case, the estimated information is greater than the true information, which causes the boundaries at interim analyses to be too narrow on the sample mean scale. These narrow boundaries lead to some null trials being declared effective when they would not have been if stopping boundaries constructed with the correct information had been used. This effect can be seen by noting the difference in early stopping probabilities rejecting the null (SP_{alt}) at each analysis when using boundaries fixed by the naive estimates of the information and under the true information when the null hypothesis is true. The inflation of the type I error rates is slightly less for designs with both efficacy and futility boundaries, as some trials are stopped prematurely early for futility, thus preventing these trials from contributing to the type I error rate.

When using a single efficacy stopping boundary, designs constructed using the incorrect information growth have a very slight increase in power for alternatives compared to designs constructed using the correct information growth. Using the naive information growth, interim boundaries are closer to the null and this change makes it more likely for trials to be declared effective. When using both efficacy and futility boundaries, in addition to the inflation of the type I error rate, there is also a loss of power due to the overestimated information. Table 4.5 also shows that for the alternative with 97.5% power under the true information growth, the power is only 71% using Pocock boundaries and 80% using O'Brien-Fleming. This result is again due to the overestimated information causing the interim boundaries to be too narrow; in this case the boundary for futility causes some trials that would eventually reject the null to be stopped early for futility.

A more typical approach is to fix boundaries on the z-statistic scale, rather than on the sample mean scale. Intuitively, this approach scales the test statistic by the approximately true information available, so that the problem of overestimating the information at early analysis times is lessened. The use of these statistics is known as the significance level approach (Jennison and Turnbull, 2000). Indeed, the nominal type I error rate is nearly preserved by fixing the boundaries on the z-statistic scale, even with the dramatic overestimation of the true information growth (tables 4.6 and 4.7). With the one boundary design, there is a slight elevation in the type I error rate and a slight decrease in the power due to the increased correlation of measurements at the first interim analyses relative to what

Table 4.6: Stopping probability for the alternative (SP_{alt}) at each of the four analyses under the null and the alternative with 97.5% power using Pocock and O'Brien-Fleming (OBF) stopping boundaries fixed under the naive information growth on the z-statistic scale.

One Boundary - Under Null								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.007
2	0.000	0.002	0.000	0.000	0.000	0.008	0.000	0.007
3	0.000	0.010	0.000	0.002	0.000	0.008	0.000	0.006
4	0.971	0.017	0.975	0.023	0.969	0.006	0.975	0.005
Total	0.971	0.029	0.975	0.025	0.969	0.031	0.975	0.025
One Boundary - Under Alternative								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.028
2	0.000	0.081	0.000	0.000	0.000	0.172	0.000	0.192
3	0.000	0.586	0.000	0.449	0.000	0.461	0.000	0.512
4	0.027	0.307	0.025	0.526	0.051	0.286	0.025	0.243
Total	0.027	0.973	0.025	0.975	0.051	0.949	0.025	0.975

Table 4.7: Stopping probability (SP) for the null or alternative at each of the four analyses under the null and the alternative with 97.5% power using Pocock and O'Brien-Fleming (OBF) stopping boundaries fixed under the naive information growth on the z-statistic scale.

Two Boundary - Under Null								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.010	0.000	0.000	0.000	0.452	0.010	0.033	0.007
2	0.492	0.002	0.000	0.000	0.418	0.008	0.251	0.007
3	0.407	0.010	0.455	0.002	0.097	0.004	0.547	0.006
4	0.069	0.009	0.520	0.023	0.009	0.001	0.144	0.004
Total	0.978	0.022	0.975	0.025	0.977	0.023	0.975	0.025
Two Boundary - Under Alternative								
	OBF				Pocock			
Analysis	Using Naive IG		Using True IG		Using Naive IG		Using True IG	
Number	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
1	0.002	0.000	0.000	0.000	0.274	0.033	0.007	0.033
2	0.069	0.087	0.000	0.000	0.178	0.156	0.007	0.251
3	0.048	0.574	0.002	0.455	0.041	0.257	0.006	0.547
4	0.008	0.211	0.023	0.520	0.003	0.058	0.004	0.144
Total	0.128	0.872	0.025	0.975	0.495	0.505	0.025	0.975

is assumed under the naive information growth (table 4.6). With two boundary designs, this approach is very slightly conservative in the type I error rate and can have a dramatic loss of power for the alternative that would have 97.5% power using the correct information growth (table 4.7). Here, the increased correlation of the early interim analyses leads to trials being stopped prematurely early for futility, thus contributing to the (slightly) reduced type I error rate and the loss of power. This result is most noticeable with the Pocock designs; the critical values for O'Brien-Fleming tests at early analyses are so high that few trials are stopped at the earliest analyses, thus reducing this problem somewhat.

There are two more flexible methods to maintain the correct type I error rate during a group sequential study compared to the significance level approach. The constrained boundary approach (Burington and Emerson, 2003) is one such method. We use the constrained boundary method to use the fixed (z-scale) boundaries at the interim analyses and then recalculate boundaries at the end of the study, given the true information growth. This approach will maintain the nominal type I error rate as the amount of type I error that has already occurred can be computed when the final boundaries are calculated. This method will generally not preserve power completely depending on the spacing of the analyses in information growth time.

Constraining the first three analyses on the z-statistic scale for the case above, leads to attained power in the one boundary case of 96.8% using the O'Brien-Fleming boundaries. In the one boundary case, the constrained boundary after fixing the first three does not maintain the type I error for the Pocock design because the cumulative stopping probability for the first three analyses under the null is greater than the nominal value. As seen in table 4.6, the type I error rate through three analyses is already 0.025. Thus, the constrained boundary approach cannot maintain the nominal type I error rate in this setting. In the two boundary case, the O'Brien-Fleming design has power of 87.5% and the Pocock design has power of 50.8%.

The error spending approach of Lan and DeMets (1983) can also be used to ensure correct type I error rates. This method spends a certain amount of type I error at each interim analysis and thus does not exceed the nominal level. The attained power, however, can fluctuate as seen previously, based on the spacing of the true information growth relative

to the assumed information growth. For the one boundary case above, using the constrained error spending scale gives the correct type I error rate. For the alternative with 97.5% power under the true information growth, the attained power is 96.8% for the O'Brien-Fleming design and 92.3% for the Pocock designs. In the two boundary case, the type I error rate is again maintained at the nominal level. For the two boundary designs, the attained power is 96.1% using the O'Brien-Fleming design and 97.7% using a Pocock design. In the case of the two boundary Pocock design, the power for the alternative with 97.5% power under the true information growth is actually greater using the error spending function approach because the error spending function is spending the type II error for a different alternative (based on the naive information growth). For this presumed alternative (under the naive IG) the Pocock design has 96.7% power.

Although both of these approaches do maintain the type I error rate, they will not generally conserve the expected stopping probabilities for the design as it was originally conceived. We also note that comparing the attained power for designs in which the type I error rate is not maintained is not a fair comparison, however all results are presented for illustrative purposes only. The goal of the above section was to demonstrate the potential consequences of misspecified information growth in group sequential longitudinal trials.

4.3 Homoscedastic, Nonlinear Data

We next examine the case in which the true effect is not exactly linear, but the parameter of interest is still the linear change in time from randomization. We still assume that the data are homoscedastic around the true mean. This scenario might occur if the true effect of treatment were linear in time from randomization, but does not start exactly when treatment is initiated – a circumstance that could occur if the treatment effect were delayed. It could also occur if the true treatment effect were quadratic. Here, we assume that the linear contrast over time is of scientific interest, even if the data are not exactly linear, so we are interested in the behavior of the information growth in such circumstances. We further note that in the setting of clinical trials, the analysis plan must be fully specified in advance, so there is no way to adjust the model to accommodate the observed nonlinearities. However, as mentioned previously, we are not addressing the important scientific issue of

how the meaning of an average linear trend would change across analyses in a sequential design.

Our primary interest in this case is the behavior of the information growth in this setting to illustrate a case in which the information growth may differ based on an alternative due to model misspecification. In this setting with homoscedastic but nonlinear data, the model-based standard error of the parameter of interest is a function of both the variance of the measurements and the amount of model misspecification. Here again, we let X represent the design matrix for the fitted model (the same as equation 3.2) and our parameter of interest is β_1 from the linear model $E(Y|X) = \beta_0 + \beta_1 x$. However, in this situation, $E(Y|X)$ need not be the exact linear model – it may be quadratic ($E(Y|X) = \gamma_0 + \gamma_1 x^2$), for example. Let the true value for $E(Y|X)$ be equal to μ, regardless of what the true form of the data is. The true variance for the OLS estimate of the parameter of interest is still the same as in equation 3.4. However, estimating σ^2 is more difficult in this setting. S^2 is often used to estimate σ^2.

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In the setting of a linear contrast of a nonlinear model, S^2 is not unbiased:

$$E[S^2] = \sigma^2_{y|x} + \frac{||\mu - E(X\hat{\beta})||^2}{n-2}$$

Thus, the model-based estimate of the standard errors will be too large, depending on the degree to which the linear contrast differs from the true contrast. If the degree of the systematic error is increasing with study time, then each successive model misspecification term in S^2 could be larger than the last. If the model misspecification term becomes large, then it is possible for the model-based standard errors to become nonmonotonic.

If the true effect of the treatment only takes effect after a certain amount of time on the treatment, such as might occur if a drug must build up in the body before an effect is seen, the true model might include an indicator term;

$$E(Y|X = x) = \beta_0 + \beta_1 x(\mathbf{1}_{x \geq c})$$

Time c is the time at which the treatment takes effect. To illustrate the dependence of when the treatment takes effect on the model-based information growth, we examined

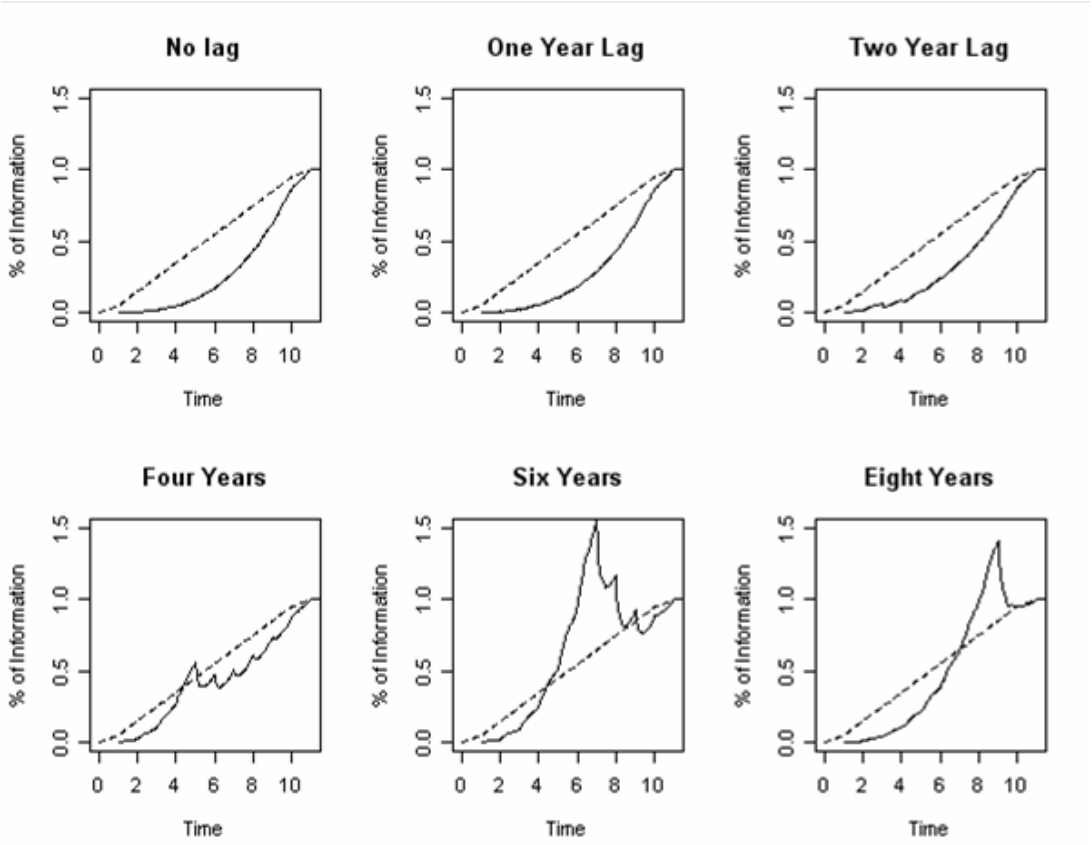

Figure 4.2: Plots showing estimated information growth from linear model (solid line) relative to the IG estimated from the total number of measurements (dashed line) under a nonlinear true effect.

different possible values for c. In all cases, the estimated linear effect at the end of the study was constant; this requirement implies that larger values of c (later treatment effects) have correspondingly larger values of the slope parameter after the treatment takes effect (β_1). We use a situation similar to before, with measurements at baseline (study time 0) and follow-up times 1-9.

Figure 4.2 shows some information growth curves estimated from the model-based standard errors for different values of the change point c. If the change point occurs after several follow up times, the curve can become nonmonotonic, indicating that the amount of "information" is decreasing with increasing measurements being made. More accurately, this phenomenon can be explained as the relative amount of information estimated at an earlier point in time was high compared to the amount of information that is eventually present at the end of the study, due to the apparent lack of systematic error in the estimated parameter.

This model misspecification case can potentially cause two types of nonmonotonicity. The first is the case in which the nonmonotonicity is such that the amount of information at an interim analysis time exceeds the amount of information that is present at the final analysis (due to the model misspecification). The second type of nonmonotonicity is when the nonmonotonicity causes the amount of information at an early interim analysis to be higher than the amount of information at a later interim analysis, but the information at both interim analyses is still less than the information at the end of the study. In the example, some nonmonotonicity is possible with a four year lag as the fraction of information when calendar time is at 4.5 years is higher than when calendar time is at 6 years, but both are still less than 1 (figure 4.2). More dramatic nonmonotonicity is observed with a lag of six (or eight years) as the amount of fractional information in the middle of these studies is greater than 1.

If the trial was planned for an appropriate sample size to detect the linear effect despite potential model misspecification (or potentially just accounting for increased estimated σ^2), then trials with dramatic nonmonotonicity are particularly problematic. Here, at an interim analysis where the estimated information fraction is greater than 1, the group sequential design might dictate that the trial be stopped for exceeding the amount of information

planned for at the beginning of the study. This scenario is potentially very problematic scientifically, as the very reason for allowing the final information to be lower (for a given σ^2) than the typical linear model was to allow for possible nonlinear models with scientifically meaningful linear trends. Some possible approaches in a setting in which a delayed treatment effect might be expected or anticipated would be (a) refusing to conduct (or plan) an analysis at which the estimated information fraction is above 1, (b) using fixed boundaries that will not necessarily match the true information growth achieved in the trial, and (c) using a bootstrap approach to separate the systematic error component from $\hat{\sigma}^2$ as explained later.

For approach (a), refusing to conduct an analysis at a time for which the information growth is above 1, may be unsatisfying scientifically; the original design must have had some motivation for wanting to conduct an interim analysis at this point in the study. Therefore, refusing to conduct an analysis due to statistical problems is less than ideal. In contrast, refusing to plan an analysis at a point in the study where the information growth may exceed 1 is more satisfying scientifically, especially if it is expected that the treatment may have a large effect late in the study (due to a lag or a quadratic effect). In such circumstances, it may be worthwhile to consider whether early stopping for futility is scientifically prudent. It may be that a stopping rule designed to stop for trending toward harm could be more scientifically relevant in this case (such that the futility boundary at interim analyses is not rejecting the alternative, but rather some less extreme value).

Option (b), using fixed boundaries at the interim analyses and then scaling appropriately at the last analysis (using either the error spending or constrained boundary approach), may be appropriate in some settings. One advantage of fixing the interim boundaries is that such boundaries can be evaluated in advance scientifically for the appropriateness of stopping for a particular effect size at the interim analyses. One difficulty using this approach is fully exploring the properties of the proposed boundaries under different potential true information growths. It is possible, though unlikely, that the combination of boundaries fixed for the interim analyses and the true information growth would together make it impossible to attain (or maintain) the nominal type I error rate.

Finally, approach (c) in this specific case of model misspecification with homoscedastic data is to remove the model misspecification component from the estimated standard error

and use the true information growth (that which is due to the constant σ^2 and the linear model) to construct and evaluate boundaries. In this setting, it is possible to remove the misspecification component through the method outlined by Kittelson et al. (2005) where σ^2 is estimated at each observed time from randomization separately or via the bootstrap. The method of Kittelson, et al. will work asymptotically if the timing of the accrual and interim analyses is such that some observations from all potential times from randomization have been observed (all x values observed). In lieu of all potential study times being observed at an interim analysis, a bootstrap approach can be used, which we outline below.

The general idea of the bootstrap approach is to simulate many replicates of the data under the "truth", calculating the slope parameter from the linear model each time and then using the variability of these replicated slope parameters to estimate the standard error of the slope without the additional component due to the lack of linearity. Conceptually this procedure is successful because in an extreme case where there was no error variance ($\sigma^2 = 0$) but the data were quadratic instead of linear, the model-based estimate of the standard error would be non-zero due entirely to the misspecified model. However, if the data were bootstrapped within the observed study times (explained in detail below), then the estimated slope would be the same each time; because there is no error variance, every bootstrapped dataset would yield exactly the same simulated data and the same slope parameter. In a less extreme case, with some error variance, the estimated slopes will be different due to the error variability but they will be more similar than the model based estimate of the standard error would predict, because every bootstrapped data set would be estimating the same linear trend from the nonlinear data.

There are several technical issues associated with using the bootstrap to simulate new data sets to estimate the true information at a particular point in time. In particular, the bootstrapped sample data sets need to maintain the same values of observed study times and the same accrual pattern.

To create a single bootstrapped estimate of the slope at a particular point in calendar time:

- Note how many measurements are taken at each study time. (i.e. 100 at study time

0, 50 at study time 1, 40 at study time 2, 0 at study time 3, etc.)

- At each study time, sample with replacement from the available measurements at that particular study time (i.e. sample 100 measurements from all measurements at study time 0, 50 from study time 1, etc.)

- Calculate the OLS slope for this bootstrapped data set

If study times are not exactly the same (our general setting), values for a particular study time can be bootstrapped within a neighborhood of the closest study time values. If the information is to be estimated at a time such that the last observed study time has only one measurement (or a very small number of measurements), then it may be necessary to consider a parametric bootstrap at this study time. In such a case, the population to be bootstrap sampled from could be estimated to have the mean of the observed data at that study time with a standard deviation estimated from the other study times.

One example of how the bootstrap can be used to recreate the true information growth in the presence of model misspecification is shown in figure 4.3. This example considers a scenario with uniform accrual over two years, and measurements at baseline and every two years thereafter (so $x = 0, 2, 4, 6$) in which the data are truly quadratic. Figure 4.3 shows the true information growth, the model based information growth (which is nonmonotonic) and 10 bootstrapped estimates of the information growth.

Although the model misspecification case is important to consider for group sequential designs, we will not consider it further here. We are also not addressing the issue of whether or not interim analyses of a linear contrast are relevant for settings in which there is strong reason to believe that the underlying contrast is dramatically nonlinear.

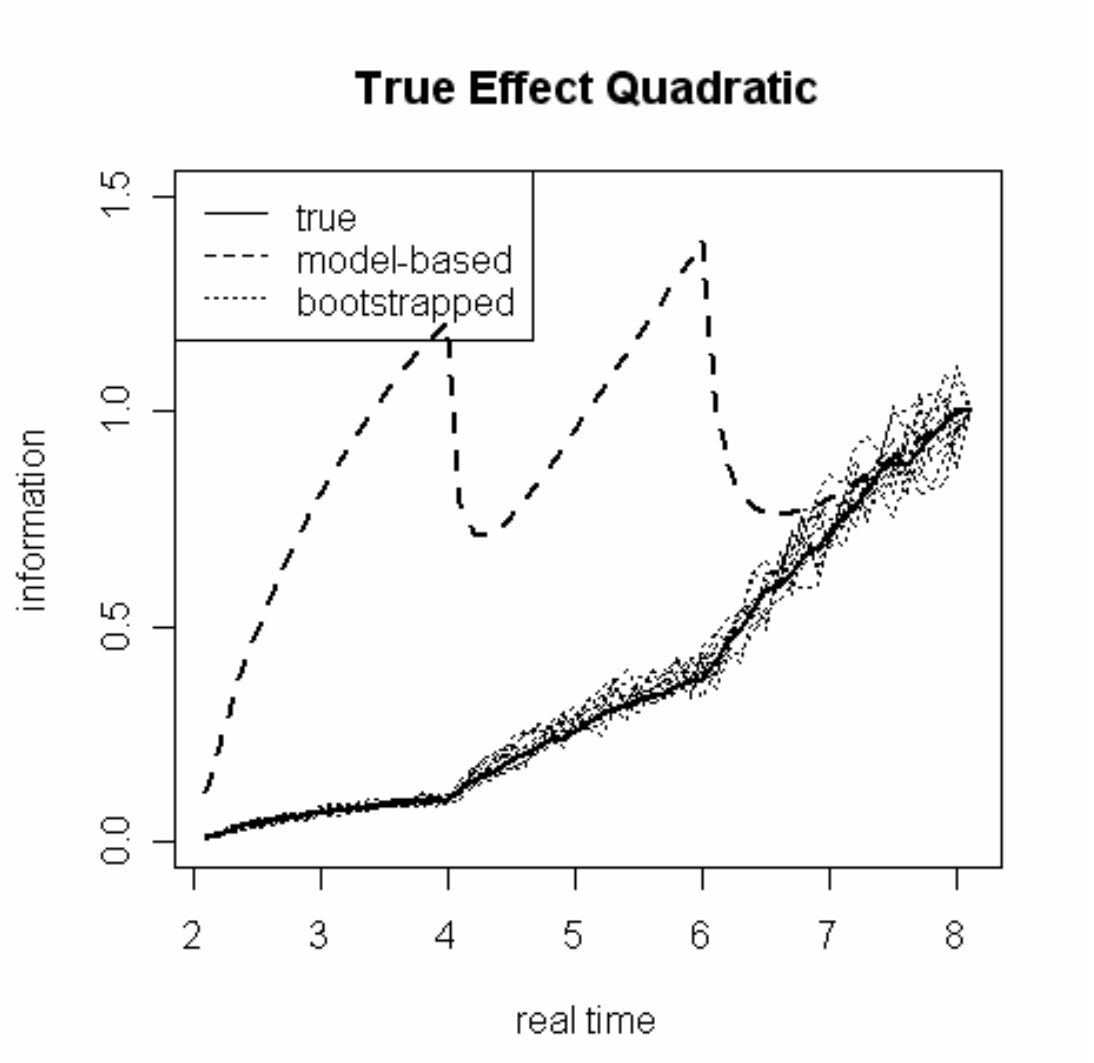

Figure 4.3: Information growth in the non-linear contrast setting is correctly estimated using a bootstrap approach when the model-based IG (dashed line) is nonmonotonic.

Chapter 5

INDEPENDENT DATA WITH PREDICTOR-VARIANCE HETEROSCEDASTICITY

We next examine the case in which the assumption of constant variance is incorrect due to a predictor-variance relationship. In this case, regardless of the alternative, the longitudinal measurements are heteroscedastic; the variance of the outcomes depends on the value of the predictor variable. In our case, this relationship would correspond to a model where the measurements were becoming more (or less) variable in study time. For example, such a relationship might occur if there were rigid entry criteria for the study (such as fasting glucose values within a narrow range), causing measurements close to study time 0 to be less variable than measurements after more study time has passed.

In this chapter, we examine the effect of this predictor-variance heteroscedasticity on the true information growth for a clinical trial, as well as consequences that stem from ordinary least squares regression in this setting.

5.1 Model

In this case our standard one-sample model:

$$\begin{aligned} E(\boldsymbol{Y}|\boldsymbol{x}) =& \boldsymbol{\mu} \\ =& \beta_0\mathbf{1} + \beta_1\boldsymbol{x} \\ Cov(\boldsymbol{Y_i}) =& \sigma^2 V(\boldsymbol{\mu}) \end{aligned}$$

has

$$\begin{aligned} V_{kk} &= (a + bx_k)^\gamma & (5.1) \\ V_{kk'} &= 0 & k \neq k' \end{aligned}$$

where $\gamma \geq 0$ and a and b are constants such that $(a + bx_k)^\gamma$ is non-negative over the range of x values observed.

This model allows the constants a and b to alter the amount of heteroscedasticity in study time. The case in which these constants are related to the alternative of interest is covered in the next chapter, in which we consider a mean-variance relationship. In the present chapter, we assume that a and b are not related to the parameter of interest.

5.2 Using Weighted Least Squares (Efficient, Known Weights)

In the case of heteroscedasticity, the ordinary least squares estimate, $\hat{\beta} = (X^T X)^{-1} X^T Y$, is not the best linear unbiased estimator. Rather, the best linear unbiased estimator (BLUE) is the weighted least squares estimate, $\hat{\beta}_w = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$, with W known and equal to $Var(Y)$. As such, if an *a priori* decision was made to use weighted least squares regression (with known weights) then the sequential analyses would still have an independent increment structure, as previously described (Jennison and Turnbull, 1997; Scharfstein et al., 1997).

Because it leads to an independent increment structure, using WLS with known weights is thus similar to the case of independent longitudinal data. The information again does not grow linearly with each added observation, but rather depends on the timing of the accrual pattern, the time from randomization and the analysis times. However, the true information growth can be estimated with knowledge of the expected study observation times (so that the $Var_j(x)$ can be estimated).

Two additional concerns in the WLS case are designing a study to account appropriately for the degree of heteroscedasticity and for the need to estimate the true weights in practice. The first potential problem is that the amount of heteroscedasticity over time in the study can affect the information growth. As might be expected intuitively, the amount of fractional information early in a study is higher for a study with a large amount of positive heteroscedasticity (measurements becoming increasingly variable) than for one with little to no heteroscedasticity. Assuming fast accrual relative to the timing of the follow up measurements, the measurements taken later in the study will generally be more variable and thus contribute less to the overall information growth. As an example of different information growths under different amounts of heteroscedasticity, consider a scenario with measurements taken at baseline and follow up times 1-5, and uniform accrual over two years.

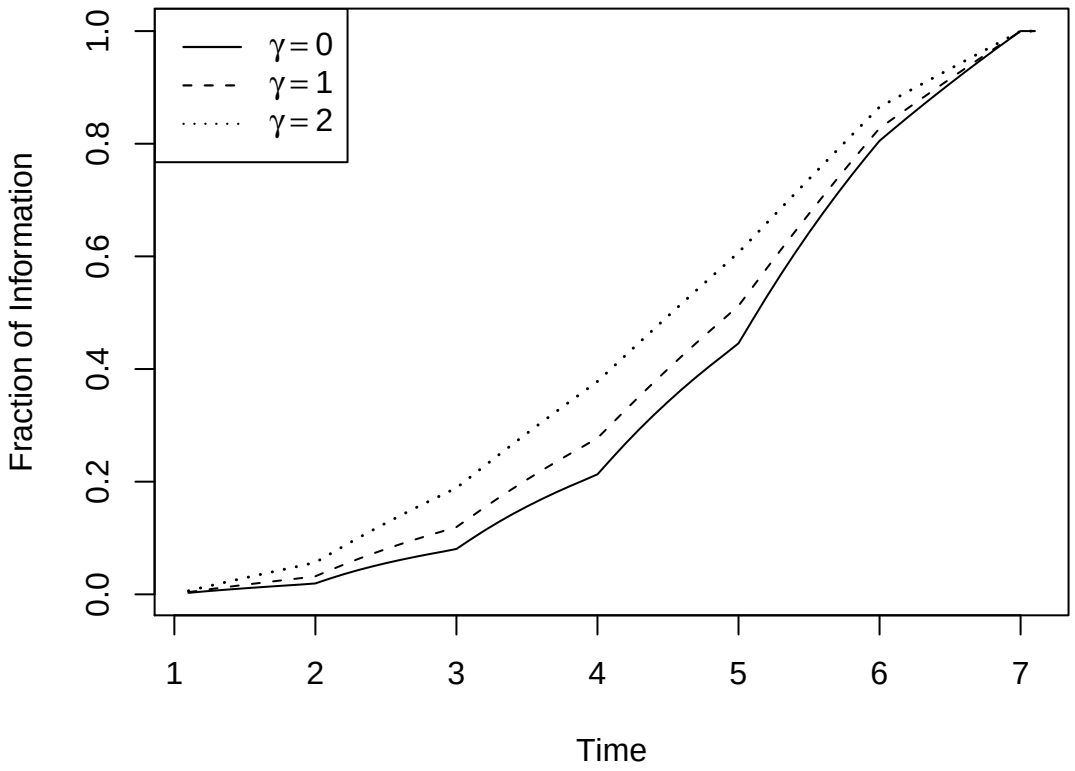

Figure 5.1: The true information growth using WLS with different amounts of heteroscedasticity. With no heteroscedasticity ($\gamma = 0$), the information grows more slowly than in cases with more heteroscedasticity.

Figure 5.1 shows the true information growth curves using WLS with different amounts of positive heteroscedasticity indicated by the parameter γ as in equation 5.1. As expected, greater heteroscedasticity leads to faster information growth. It should be noted here, that greater heteroscedasticity will lead to lower overall information for a constant sample size and similar values of the error variance σ^2.

If heteroscedasticity was unexpected, or if greater heteroscedasticity is seen than was planned for, then the estimated fractional information at interim analysis times will be too low. Such underestimation of the true information growth will generally lead to boundaries that are too conservative at interim analyses, compared to what was originally planned. This result is in many respects the opposite problem from the naive overestimation of the information growth with independent, longitudinal data.

5.3 Using Ordinary Least Squares (Inefficient (known) Weights)

We next turn attention to the case of using an inefficient estimator, in this case using ordinary least squares regression (OLS) with heteroscedasticity. Using OLS will have consequences both with the assumption of independent increments in the sampling density and in potential nonmonotonicities in the true information growth.

With simple linear regression, we will find it useful to note expressly the formula for the variance of the OLS estimate for the parameter of interest (β_1) with heteroscedastic data. Using a model for the predictor-variance relationship as in equation 5.1, we let σ_i^2 denote the variance of the ith observation for simplicity in the notation. Then,

$$Var(\hat{\beta}_j) = (X_j^T X_j)^{-1}(X_j^T V_j X_j)(X_j^T X)^{-1}$$

$$= \begin{bmatrix} n_j & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum \sigma_i^2 & \sum x_i \sigma_i^2 \\ \sum x_i \sigma_i^2 & \sum x_i^2 \sigma_i^2 \end{bmatrix} \begin{bmatrix} n_j & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1}$$

Solving just for $Var(\hat{\beta}_{1j})$ gives:

$$
\begin{aligned}
Var(\hat{\beta}_{1j}) &= \left(\frac{1}{n_j \sum_{i=1}^{n_j} (x_i - \bar{x}_j)^2} \right)^2 \left\{ \left(\sum_{i=1}^{n_j} x_i \right)^2 \sum_{i=1}^{n_j} \sigma_i^2 - 2n_j \left(\sum_{i=1}^{n_j} x_i \right) \left(\sum_{i=1}^{n_j} x_i \sigma_i^2 \right) + n_j^2 \left(\sum_{i=1}^{n_j} x_i^2 \sigma_i^2 \right) \right\} \\
&= \left(\frac{1}{n_j \sum_{i=1}^{n_j} (x_i - \bar{x}_j)^2} \right)^2 \left\{ n_j^2 \sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)^2 \right\} \\
&= \left(\frac{\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)^2}{\left(\sum_{i=1}^{n_j} (x_i - \bar{x}_j)^2 \right)^2} \right) \\
&= \left(\frac{1}{n_j Var_j(x)} \right) \left(\frac{\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)^2}{\sum_{i=1}^{n_j} (x_i - \bar{x}_j)^2} \right) \qquad (5.2)
\end{aligned}
$$

This formula illustrates that the variance of the slope when using OLS will be larger when the data at the extremes of the predictor space (i.e. with large values of $x_i - \bar{x}$) are more variable. Intuitively, this outcome is expected as data at the extremes will be points of high leverage when weighted as if the data were independent. The high leverage at points further in the predictor space will make the slope more variable if the possible values of the observations at such points are themselves more variable. In contrast, if data at the extremes of the predictor space were highly non-variable compared to data less extreme in the predictor space, the slope would be much less variable. The lack of variability at points of high leverage would in this scenario "anchor" the regression line over repeated samples (despite the variability of points in the middle of the predictor space) and cause the variability of the slope to be less than what would have been predicted assuming homoscedasticity.

5.3.1 Non-Independent Increments

As noted previously, the integration of the sampling density for a group sequential trial is simplified greatly by the assumption of independent increments, and all common software for clinical trials relies on this assumption. However, the assumption may not hold if an inefficient estimator is used.

For two analysis times, t_j and t_k with $j < k$, let β_j and β_k denote the OLS estimates at these analysis times. Then:

$$\begin{aligned} Var(\hat{\beta}_j) =& (X_j^T X_j)^{-1} X_j^T V_j X_j (X_j^T X_j)^{-1} \\ Var(\hat{\beta}_k) =& (X_k^T X_k)^{-1} X_k^T V_k X_k (X_k^T X_k)^{-1} \end{aligned} \tag{5.3}$$

Following the notation of Lee et al. (1996), we note that the estimate $\hat{\beta}_j$ must be based on a subset of the full data (Y_k), such that $\hat{\beta}_j = A^T Y_k$ for some $k \times k$ matrix A. Then,

$$Cov(\hat{\beta}_j, \hat{\beta}_k) = A^T V_k X_k (X_k^T X_k)^{-1}$$

In particular, A^T can be written as a block matrix, $[(X_j^T X_j)^{-1} X_j^T \quad 0_{2\times n_{j^*}}]$, where n_{j^*} denotes the number of additional measurements between analysis times ($n_k = n_j + n_{j^*}$). Under this arrangement, V_j will be a $n_j \times n_j$ submatrix of V_k, such that

$$\begin{aligned} Cov(\hat{\beta}_j, \hat{\beta}_k) =& \left[(X_j^T X_j)^{-1} X_j^T \quad 0_{2\times n_{j^*}}\right] V_k X_k (X_k^T X_k)^{-1} \\ =& \left[(X_j^T X_j)^{-1} X_j^T V_j \quad 0_{2\times n_{j^*}}\right] \begin{bmatrix} X_j \\ X_{j^*} \end{bmatrix} \left[X_k^T X_k\right]^{-1} \\ =& \left[(X_j^T X_j)^{-1} X_j^T V_j X_j + 0_{2\times 2}\right] \left[X_k^T X_k\right]^{-1} \\ =& \left[(X_j^T X_j)^{-1} X_j^T V_j X_j\right] \left[X_k^T X_k\right]^{-1} \end{aligned} \tag{5.4}$$

Using OLS will lead to independent increments if there is no heteroscedasticity (i.e. if $V_j = \sigma^2 I_j$) . In this case, $Cov(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 (X_k^T X_k)^{-1}$, which is equal to $Var(\hat{\beta}_k)$ when there is no heteroscedasticity.

Using OLS will also lead to independent increments if both designs are fully balanced and have the same vector of $\boldsymbol{x}$ measurements, but analysis time t_k has more observations. In this case, $cX_j^T X_j = X_k^T X_k$, where c can be interpreted as the multiplicative amount of

additional complete cases at time t_k than at time t_j (and there are no incomplete cases at either time due to the balance and complete sampling of the study times). Additionally, in this setting, note that $cX_j^T V_j X_j = X_k^T V_k X_k$, so independent increments are a direct result of applying the formulas in equations 5.4 and 5.3:

$$\begin{aligned} Cov(\hat{\beta}_j, \hat{\beta}_k) &= \left[(X_j^T X_j)^{-1} X_j^T V_j X_j\right] \left[X_k^T X_k\right]^{-1} \\ &= \left[(X_j^T X_j)^{-1} X_j^T V_j X_j\right] \left[cX_k^T X_k\right]^{-1} \\ &= \frac{1}{c} Var(\hat{\beta}_j) \end{aligned}$$

and

$$\begin{aligned} Var(\hat{\beta}_k) &= (X_k^T X_k)^{-1} X_k^T V_k X_k (X_k^T X_k)^{-1} \\ &= (cX_j^T X_j)^{-1} (cX_j^T V_j X_j)(cX_j^T X_j)^{-1} \\ &= \frac{1}{c} Var(\hat{\beta}_j) \end{aligned}$$

However, the use of OLS will not give independent increments in general. Specifically, even if there is balance at every interim analysis such that there are the same number of measurements at every study time that have been observed by a particular interim analysis, but the number of study times between analyses is different, there will not necessarily be an independent increment structure.

The amount of departure from independent increments will depend on the amount of change in the variability of the new measurements compared to the existing measurements, as well as the amount of change in the distribution of the x measurements. Specifically for the case of simple linear regression, we can use equation 5.4 and focus on the covariance of

the slope parameters of interest, ($Cov(\hat{\beta}_{1j}, \hat{\beta}_{1k})$, to get:

$$\begin{aligned} Cov(\hat{\beta}_{1j}, \hat{\beta}_{1k}) = &\left(\frac{1}{n_j^2 Var_j(x)}\right)\left(\frac{1}{n_k^2 Var_k(x)}\right) \times \\ &\{\sum_{i=1}^{n_j} x_i \sum_{i=1}^{n_k} x_i \sum_{i=1}^{n_j} \sigma_i^2 + n_j n_k \sum_{i=1}^{n_j} x_i^2 \sigma_i^2 \\ &- n_j \sum_{i=1}^{n_j} x_i \sum_{i=1}^{n_j} x_i \sigma_i^2 - n_k \sum_{i=1}^{n_k} x_i \sum_{i=1}^{n_j} x_i \sigma_i^2)\} \end{aligned}$$

$$= \left(\frac{1}{n_j Var_j(x)}\right)\left(\frac{1}{n_k Var_k(x)}\right)\left\{\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)(x_i - \bar{x}_k)\right\} \tag{5.5}$$

For independent increments to be true, this expression for the covariance must be equal to the variance at the later time point, t_k. Using equation 5.2, we note that:

$$Var(\hat{\beta}_{1k}) = \left(\frac{1}{n_k Var_k(x)}\right)\left(\frac{1}{n_k Var_k(x)}\right)\sum_{i=1}^{n_k} \sigma_i^2 (x_i - \bar{x}_k)^2$$

Then combining the expressions for $Cov(\hat{\beta}_{1j}, \hat{\beta}_{1k})$ and $Var(\hat{\beta}_{1k})$ gives:

$$\frac{Cov(\hat{\beta}_{1j}, \hat{\beta}_{1k})}{Var(\hat{\beta}_{1k})} = \frac{\left(\frac{1}{n_j Var_j(x)}\right)\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)(x_i - \bar{x}_k)}{\left(\frac{1}{n_k Var_k(x)}\right)\sum_{i=1}^{n_k} \sigma_i^2 (x_i - \bar{x}_k)^2} \tag{5.6}$$

$$= \frac{\left(\frac{1}{n_j Var_j(x)}\right)\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)^2 + \left(\frac{\bar{x}_j - \bar{x}_k}{n_j Var_j(x)}\right)\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)}{\left(\frac{1}{n_k Var_k(x)}\right)\sum_{i=1}^{n_k} \sigma_i^2 (x_i - \bar{x}_k)^2}$$

So the degree of non-independent increments will depend on the amount of change in the ratio of the weighted variance to the unweighted one at t_j and t_k, as well as the difference in the average of the predictor value x. This will be explored further below.

Assessing Departures from Independent Increments

In order to investigate the effect of departures from independent increments on the integration of the sampling density, we created two metrics: the sum of relative departures and

the linear trend in departures.

For the sum of relative departures metric, we assess relative departures from the covariance matrix that would be true under independent increments. Under independent increments, the covariance matrix of the observed statistic at four analysis times has the form:

$$\begin{bmatrix} a & b & c & d \\ b & b & c & d \\ c & c & c & d \\ d & d & d & d \end{bmatrix}$$

where a, b, c, and d are the variances of the statistic at each of the four analysis times. We define the sum of the relative departure from independent increments as:

$$\sum_{j=2}^{J} \sum_{i=1}^{j-1} \left| 1 - \frac{Cov(\beta_{1i},\, \beta_{1j})}{Var(\beta_{1j})} \right| \tag{5.7}$$

The use of relative departures, rather than absolute, ensures that two studies with different numbers of clusters (e.g. individuals) but identical relative accrual patterns, timing of measurements, and analysis times, would yield the same value for departures from independent increments.

This metric for examining the degree of departure from independent increments, however, does not reveal the direction of any departures. For example, the correlation of previous statistics with the final one could be either higher or lower than expected under independent increments, but could give the same value on the relative departures scale. For this reason, we created another metric that would use directional information.

For this second metric, the linear trend in relative departures, we examine the linear trend in the relative values of the covariance with the final observation. So the value of this linear trend in relative departures from independent increments metric is given by:

$$\left(\frac{1}{J \sum_{j=1}^{j} (j - \bar{j})^2} \right) \left(J \sum_{j=1}^{J} j \frac{Cov(\hat{\beta}_{1j}, \hat{\beta}_{1J})}{Var(\hat{\beta}_{1J})} - (\sum_{j=1}^{J} j) \left(\sum_{j=1}^{J} \frac{Cov(\hat{\beta}_{1j}, \hat{\beta}_{1J})}{Var(\hat{\beta}_{1J})} \right) \right) \tag{5.8}$$

which is the slope from a linear regression of the scaled covariances on the analysis number (j). The above equation uses $\bar{j}$ to denote the average of these analysis numbers, similar to $\bar{x}$ in the standard linear regression setting.

For this metric we standardize the values to the variance of the final statistic, to ensure again that the number of clusters does not yield different results for otherwise identical studies. Positive values of this metric reflect that previous analyses are less correlated than would be expected under independent increments, while negative values reflect that the final analysis is more correlated with the previous ones.

Consequences of departures from independent increments

In this setting of predictor-variance heteroscedasticity, the amount of the departure from independent increments is constant over all possible alternatives. However, because the assumption of independent increments is used when integrating the sampling density, departures from it may impact calculation of the type I error rate and power for a specific alternative.

To illustrate this possibility, we chose a scenario in which measurements are made at baseline and study times 1-5, and individuals are accrued uniformly over two years. Four total analyses are spaced equally in calendar time. We varied the amount of heteroscedasticity by varying the amount of baseline variability (a), the amount of additional variability with each one-unit increase in the predictor (b) and the power of the heteroscedasticity (γ), as per equation 5.1. For simplicity, we assumed $\sigma^2 = 1$ in all cases.

The impact of non-independent increments was assessed by comparing the expected type I error rates and power using the standard sequential methods with the results from simulating trials with non-independent increments and using the boundaries developed under the assumption of independent increments. We evaluated power at the alternative calculated to have 97.5% power using the independent increment assumption. To summarize, for each possible predictor-variance scenario, we construct a design such that:

- Boundaries are constructed using the true information growth for the scenario but assuming independent increments. (So the diagonal of the covariance matrix of the statistics is specified correctly, but the off-diagonal elements are not.)

- At each interim analysis, the z-statistic for the interim analysis is compared to the boundaries constructed during the design phase.

- The alternative with 97.5% power for comparison was calculated from the design that assumes independent increments.

The design is evaluated at both the null and the alternative by simulating values of the interim statistics from the true covariance matrix for the interim statistic (that may not have independent increments). For each power calculation, we simulated one million trials. If independent increments were true, then all of the designs would have a type I error rate of 0.025 and 97.5% power for the calculated alternative. The simulation of one million trials gives a standard error on the empirical estimate of 0.000156 if the true power were 0.025 or 0.975. If the true type I error rate were 0.025, 95% of the empirical estimates would be in the range of 0.02468 to 0.02531.

In general, assuming independent increments in the presence of predictor-variance heteroscedasticity did not lead to severe problems with either the type I error rate or the power (table 5.1). Extreme departures, where later outcomes were dramatically more variable, did cause the overall type I error rate of the study to decrease below the nominal 0.025 level.

As an illustration of a specific case of extreme heteroscedasticity, consider $a = 10$, $b = 5$, and $\gamma = 2.25$. With 100 individuals, the covariance matrix is:

$$Cov \begin{bmatrix} \hat{\beta}_{11} \\ \hat{\beta}_{12} \\ \hat{\beta}_{13} \\ \hat{\beta}_{14} \end{bmatrix} = \begin{bmatrix} 13.697 & 0.571 & -0.069 & -0.126 \\ 0.571 & 2.723 & 0.446 & 0.122 \\ -0.069 & 0.446 & 1.293 & 0.549 \\ -0.126 & 0.122 & 0.549 & 0.850 \end{bmatrix}$$

The sum of relative departures metric has the value 4.86; the linear trend in departures metric has the value 0.395. A γ of 2.25 is a level of heterscedasticity that is likely greater than what may occur in practice, yet assuming independent increments gives boundaries that are only slightly conservative under the null: type I error rates of 0.021 and 0.023 with the O'Brien-Fleming and Pocock designs, respectively (table 5.1).

The case of measurements becoming more variable in the predictor space leads measurements from previous analyses to be less correlated with the current one than would be expected under independent increments. Intuitively, this observation is reasonable. Additional measurements obtained at points far from the current value of $\bar{x}$ have high leverage

Table 5.1: Empirical type I error rate and power for the alternative calculated to have 97.5% power under an independent increment structure, under various predictor-variance relationships. The relative and linear departures from independent increments are as in equations 5.7 and 5.8, respectively.

		Relative	Linear	OBF		Pocock	
		Ind. Inc.	Ind. Inc.	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
a=10, b=1	$\gamma = 1$	1.1058	0.1086	0.0248	0.9751	0.0252	0.9746
	$\gamma = 2$	2.0721	0.1959	0.0246	0.9754	0.0250	0.9748
	$\gamma = 2.25$	2.2890	0.2142	0.0245	0.9756	0.0249	0.9749
a=10, b=5	$\gamma = 1$	3.1910	0.3015	0.0243	0.9758	0.0255	0.9751
	$\gamma = 2$	4.6718	0.3905	0.0220	0.9782	0.0233	0.9773
	$\gamma = 2.25$	4.8579	0.3945	0.0211	0.9791	0.0227	0.9780
a=30, b=5	$\gamma = 1$	1.6415	0.1596	0.0248	0.9752	0.0251	0.9746
	$\gamma = 2$	2.9261	0.2682	0.0241	0.9760	0.0246	0.9753
	$\gamma = 2.25$	3.1876	0.2878	0.0238	0.9762	0.0243	0.9755
a=30, b=-5	$\gamma = 1$	3.6718	-0.3877	0.0245	0.9755	0.0231	0.9767
	$\gamma = 2$	6.2328	-0.6840	0.0242	0.9758	0.0211	0.9789
	$\gamma = 2.25$	6.7175	-0.7421	0.0242	0.9758	0.0208	0.9793

using OLS. If these measurements are more variable than previous measurements, these points will have a greater potential influence on the slope than would be expected if there were no heteroscedasticity. In turn, the estimate of the slope at the later analysis will be less similar to the estimate at an earlier analysis due to the increased variability of the points further in the predictor space.

The opposite is true for the (rare) case in which measurements are becoming less variable as study time increases. In this case, the decreased variability at points of high leverage means that such points have less influence on the estimated slope than would have occurred if there were no heteroscedasticity. In this case, although the points have high leverage, they end up on average having less influence than expected due to decreased variability.

In both cases, the lack of independent increments is a result of using a weighting that is optimal in the independent, homoscedatsic setting and not optimal in the setting of heteroscedasticity. However, examining the properties of using OLS in these circumstances is important, as the variance structure of the data may not be known in advance.

In this setting, non-independent increments do not generally lead to gross departures from the nominal type I error rate. For intuition as to why this may be, we consider a trial with two interim analyses at times t_1 and t_2. Let S_1 and S_1 denote the value of the partial sum statistic at analysis times t_1 and t_1, respectively. Let S_{2-1} denote the additional incremental data added between the two analysis times ($S_{2-1} = S_2 - S_1$), and let C_1 denote the continuation region at the first analysis. The variance and expectation of the partial sum statistic at the second analysis time are given by:

$$\begin{aligned} Var(S_2|S_1 \in C_1) =& Var(S_1 + S_{2-1}|S_1 \in C_1) \\ =& Var(S_1|S_1 \in C_1) + Var(S_{2-1}|S_1 \in C_1) + 2Cov(S_1, S_{2-1}|S_1 \in C_1) \end{aligned} \tag{5.9}$$

$$\begin{aligned} E(S_2|S_1 \in C_1) =& E(S_1 + S_{2-1}|S_1 \in C_1) \\ E(S_2|S_1 \in C_1) =& E(S_1|S_1 \in C_1) + E(S_{2-1}|S_1 \in C_1) \end{aligned} \tag{5.10}$$

If the data have an independent increment structure, $Var(S_{2-1}|S_1 \in C_1) = Var(S_{2-1})$ and $Cov(S_1, S_{2-1}|S_1 \in C_1) = 0$. Further, $E(S_{2-1}) = E(S_1)$. Thus, equations 5.9 and 5.10

become:

$$Var(S_2|S_1 \in C_1) = Var(S_1|S_1 \in C_1) + Var(S_{2-1}) + 0$$
$$E(S_2|S_1 \in C_1) = E(S_1|S_1 \in C_1) + E(S_1)$$

If the increments are not independent, these simplifications do not hold due to three assumptions being violated. (1) The expectation of the second increment, $E(S_{2-1}|S_1 \in C_1)$, may be different. (2) The variance of the second increment, $Var(S_{2-1})$, will not be the same as when the data are truly independent, when the variance of the second statistic ($Var(S_2)$) is considered constant. (3) The covariance of the increments, $Cov(S_1, S_{2-1})$, is nonzero. Here, we summarize these competing effects on the distribution of S_2 when independent increments do not hold.
If $Cov(S_1, S_{2-1}) > 0$:

- **Expectation of** S_{2-1} With positive correlation between the increments, the expectation $E(S_{2-1}|S_1 \in C_1)$ will be more similar to $E(S_1|S_1 \in C_1)$ than is true when the increments are independent. Using a futility boundary and an efficacy boundary for a positive alternative, under the null, this term will be greater than zero, thus the expectation of the statistic at the second analysis time, $E(S_2|S_1 \in C_1)$, will be greater than what would be true with independent increments.

- **Variance of** S_{2-1} With positive correlation between the increments, the variance of the second increment, $Var(S_{2-1})$, will be less than if the increments were independent. This fact is true because we are comparing independent vs. non-independent increments with constant values for $Var(S_1)$ and $Var(S_2)$. Because $Var(S_2) = Var(S_1) + Var(S_{2-1}) + 2Cov(S_1, S_{2-1})$, if $Cov(S_1, S_{2-1}) > 0$, then $Var(S_{2-1})$ under positive correlation must be less than is true when $Cov(S_1, S_{2-1}) = 0$.

- **Covariance of the increments** S_1 **and** S_{2-1} By definition, with positive correlation between the increments, the covariance between the two is greater than what would be true with independent increments.

These factors will impact the distribution of the test statistic at the second analysis in different ways that depend also on the choice of boundary. They may approximately balance out, as the variance of S_{2-1} is smaller than expected (and thus generally $Var(S_{2-1}|S_1 \in C_1)$ will be smaller than expected), but the covariance term is greater than expected (and thus generally $Cov(S_1, S_{2-1}|S_1 \in C_1)$ will be greater than expected).

If $Cov(S_1, S_{2-1} < 0$:

- **Expectation of** S_{2-1} With negative correlation between the increments, the expectation $E(S_{2-1}|S_1 \in C_1)$ will be less similar to $E(S_1|S_1 \in C_1)$ than is true when the increments are independent. Using a futility boundary and an efficacy boundary for a positive alternative, under the null, this term will be less than zero, thus the expectation of the statistic at the second analysis time, $E(S_2|S_1 \in C_1)$, will be less than what would be true with independent increments.

- **Expectation of** S_{2-1} With negative correlation between the increments, the variance of the second increment, $Var(S_{2-1})$, will be greater than if the increments were independent. This fact is true because we are comparing independent vs. non-independent increments with constant values for $Var(S_1)$ and $Var(S_2)$. Because $Var(S_2) = Var(S_1) + Var(S_{2-1}) + 2Cov(S_1, S_{2-1})$, if $Cov(S_1, S_{2-1}) < 0$, then $Var(S_{2-1})$ under negative correlation must be greater than is true when $Cov(S_1, S_{2-1}) = 0$.

- **Covariance of the increments** S_1 **and** S_{2-1} By definition, with negative correlation between the increments, the covariance between the two is less than what would be true with independent increments.

These factors will impact the distribution of the test statistic at the second analysis in different ways that depend also on the choice of boundary. They may approximately balance out, as the variance of S_{2-1} is greater than expected (and thus generally $Var(S_{2-1}|S_1 \in C_1)$ will be larger than expected), but the covariance term is less than expected (and thus generally $Cov(S_1, S_{2-1}|S_1 \in C_1)$ will be smaller than expected).In both cases, the interplay of the effect of the non-independent increments on the expectation and variance of the second test statistic is dependent on the choice of boundaries as well.

5.3.2 Possible Nonmonotonic Information Growth

Besides non-independent increments, another possible consequence of using OLS with heteroscedastic data is information growth that is nonmonotonic. The intuitive explanation for this possibility is that if the design is nearly balanced with less variable measurements, adding an additional measurement further out in the predictor space that is much more variable than the other measurements (due to the heteroscedasticity) may cause the variability of the estimated slope to increase. This highly variable measurement is a point of high leverage under OLS (due to its extreme x value) and thus the increased variability of this point can add a great deal of variability to the estimated slope. In extreme cases, this situation can lead to a more variable estimate of the slope (less information) than at the previous interim time (with less variable measurements). Generally, even with a great deal of heteroscedasticity, as more measurements are obtained at time points further from randomization and the design becomes more balanced again, the resulting information will again increase such that the nonmontonicity observed in this setting is temporary nonmonotonicity: The final information is still higher than the interim.

The effect of the amount of heteroscedasticity and the accrual/measurement pattern (predictor space) on possible nonmonotonic information growth can be quantified directly. For a later analysis time t_k, the ratio of the variances, $\frac{Var(\hat{\beta}_{1j})}{Var(\hat{\beta}_{1k})}$, will depend on the ratio of the number of observations, the ratio of variance of the predictor space, and the ratio of the weighted variance terms. Using equation 5.2, we get:

$$\begin{aligned}\frac{Var(\hat{\beta}_{1j})}{Var(\hat{\beta}_{1k})} &= \left(\frac{n_k}{n_j}\right)^2 \left(\frac{Var_k(x)}{Var_j(x)}\right)^2 \left(\frac{\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)^2}{\sum_{i=1}^{n_k} \sigma_i^2 (x_i - \bar{x}_k)^2}\right) \\ &= \left(\frac{n_k}{n_j}\right)^2 \left(\frac{Var_k(x)}{Var_j(x)}\right)^2 \left(\frac{\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_j)^2}{\sum_{i=1}^{n_j} \sigma_i^2 (x_i - \bar{x}_k)^2 + \sum_{i=(j+1)}^{n_k} \sigma_i^2 (x_i - \bar{x}_k)^2}\right) \end{aligned} \tag{5.11}$$

This ratio will generally be greater than 1, indicating that, as expected, the variance of the statistic of interest decreases over time. Specifically, the ratios $\frac{n_k}{n_j}$ and $\frac{Var_k(x)}{Var_j(x)}$ must always be greater than 1 in a group sequential design. However, the third term in equation 5.11 will only usually be less than 1. In order for nonmonotonic information growth to occur, this term must be significantly less than 1 to overwhelm the gains in information from increasing numbers of observations and increasing variance of the predictor space. This result will occur if the added points at analysis time t_k are such that the variances of points with high leverage are large (so that $\sigma_i^2(x_i - \bar{x})^2$ is large).

It should be noted here that there are important scientific reasons to include measurements from times further from randomization, even if doing so detracts from the precision with which we can estimate the slope. From a statistical perspective only, if nonmonotonic information were observed, we could obtain a more precise estimate of the slope, by ignoring the newly acquired data. However, this approach places complete faith in the underlying linear model being true; the additional measurements further from randomization provide evidence (or lack thereof) of the linear trend further away from randomization. This evidence of a linear trend should be included, despite the statistical concerns about increased variability of the estimate.

One illustration of the possible nonmonotonicity of the information growth in this setting is shown in figure 5.2. The accrual pattern and measurement schedule match those from the independent, homoscedastic section (2 month uniform accrual, measurements at baseline and months 1-9 thereafter). The heteroscedasticity was generated as in equation 5.1, with $\sigma^2 = 1$, $a = 1$, $b = 1$, and $\gamma = 2$. The true information growth was obtained by simulation. In this setting, particularly later in calendar time (meaning when the measurements are becoming increasingly more variable), the true "information" does in fact decrease as the first few more variable measurements are obtained. Eventually, as more measurements that are more variable are accrued, the information again increases, as would be expected with increasing the number of observations. The initial drop in information is in part attributable to the amount of influence the first points can have on the estimated slope. As the first (more variable) measurements are of high influence, the actual variability of the $\hat{\beta}_1$ parameter is increased when these first points are added. When balance is achieved, such as at the end

of a study with no dropout and no missing measurements, heteroscedasticity is less of a concern, because no point is overly influential.

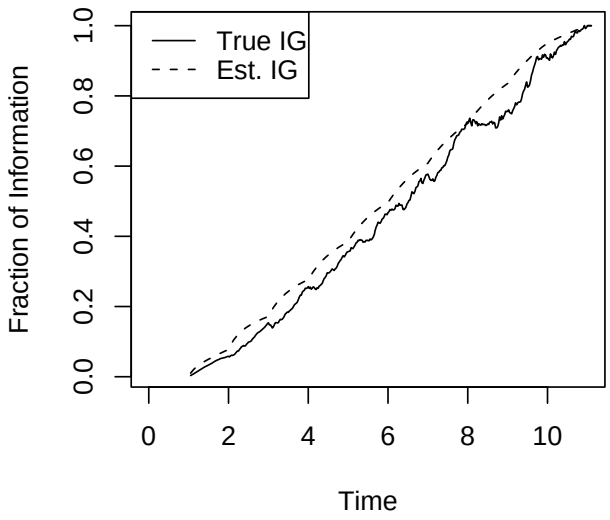

Figure 5.2: Plot illustrating information growth when the data are very heteroscedastic. The solid line represents the true, nonmonotonic information growth, simulated empirically. The dashed line represents the model-based estimates of the information growth.

An additional concern in the OLS setting occurs if the assumption of constant variance is used to estimate the standard error of the slope. If these "model-based" standard errors are used, the estimated information growth will be different from the true information growth. The model-based estimates of the standard error are constrained by the assumption of constant variance. Thus, when adding points that are more variable, the estimate of the constant variance is only slightly altered by adding a few measurements with more variability. The estimate of the (constant) variance increases as more measurements with increased variability are added, so the estimate of the variance is typically highest only after all additional measurements have been taken. This increase in variance is usually offset by gains in the number of measurements, leading to more typical behavior of the information

growth curve. However, under extreme heteroscedasticity (γ large in equation 5.1), assuming a constant variance could lead to dramatic cases in which the estimated information is lowest after all of the more variable measurements have been accrued.

5.3.3 Practical Concerns

Type I error rate

Although we have illustrated situations where even extreme heteroscedasticity does not lead to grave departures from the nominal type I error rate, if large amounts of heteroscedasticity were expected, specific boundary choices could be explored by simulation at the design stage of a trial. If large departures from the nominal type I error rate were found for approximate estimates of the heteroscedasticity, the design boundaries could be modified such that the rate could be maintained closer to the nominal level.

Power

The presence of heteroscedasticity can lead to power concerns if the design of the study is not adequately prepared for the potential increased variability. However, if a study is designed with a maximal final information (rather than a maximal sample size), and this information can be attained despite the heteroscedasticity, the power for the original alternative can be maintained (with the caveats about the differences in information growth under differing degrees of heteroscedasticity). However, practical considerations may limit the maximal number of observations that can be obtained. In this case, heteroscedasticity would contribute to a loss of power for many alternatives.

Nonmonotonic Information Growth

The possibility of nonmonotonic information growth when using OLS presents a problem for standard group sequential methods. If a study encounters nonmonotonic information growth in truth due to predictor-variance heteroscedasticity, there are three options available: (1) Do not conduct the planned interim analysis, or delay it until a time at which the information has increased again such that the later analysis is more statistically informative than prior

analyses. (2) Construct boundaries that correctly spend some amount of small, prespecified type I error at this analysis, despite the decreased information. This second approach will lead to nonmonotonic boundaries on the sample mean scale, in which estimates that would have generated stopping for efficacy at the earlier (but less variable) analysis time would not stop at this later analysis time. (3) To avoid nonmonotonic boundaries on the sample mean scale, use the sample mean scale boundaries from the previous analysis time; remaining analyses could be completed by accounting for the constrained boundaries. However, if there is extreme nonmonotonic information growth between two analyses, it is possible that this approach would spend all of the type I error for the trial (or even more than the nominal amount) at this point with nonmonotonic information growth, which would preclude conducting future analyses.

Chapter 6

INDEPENDENT DATA WITH MEAN-VARIANCE HETEROSCEDASTICITY

We next consider the case of a mean-variance relationship with independent, longitudinal data. Although we have previously considered violations of the homoscedasticity assumption centered on a predictor-variance relationship, this section considers the case in which this assumption is violated by a mean-variance relationship.

A mean-variance relationship is, by definition, one in which the variance of the data is a function of its mean value. In turn, the standard error of a statistic from these data will also depend on the mean value. As before, we are interested in the slope parameter β_1 from the linear model $E(Y|x) = \beta_0 + \beta_1 x$. A mean-variance relationship in this sequential longitudinal setting may have consequences for three possible reasons. First, as with all mean-variance relationships, the standard error of the estimated slope will depend on the true value of the slope. Second, the information growth over the course of the study will also depend on the true value of the slope. Finally, as with predictor-variance heteroscedasticity, using ordinary least squares is an inefficient estimator and may lead to a structure without independent increments.

6.1 Model

In this case our standard model,

$$\begin{aligned} E(\boldsymbol{Y}|\boldsymbol{x}) =& \boldsymbol{\mu} \\ =& \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x} \\ Cov(\boldsymbol{Y_i}) =& \sigma^2 V(\boldsymbol{\mu}) \end{aligned}$$

has

$$V_{kk} = (\beta_0 + \beta_1 x_k)^\gamma \qquad (6.1)$$
$$V_{kk'} = 0 \qquad k \neq k'$$

where $\gamma \geq 0$ and β_0 and β_1 are constrained such that $(\beta_0 + \beta_1 x_k)^\gamma$ is non-negative over the range of x values observed.

Note that this model is a specific case of equation 5.1 in which the constants a and b are equal to β_0 and β_1, respectively. As our scientific interest lies in β_1, the amount of heteroscedasticity present will depend on the value of the parameter of interest. By setting the parameter b equal to β_1, we note that under the null, $\beta_1 = 0$, there is no predictor-variance heteroscedasticity in this model.

In this model, the other parameters β_0 and γ will affect the amount of change in heteroscedasticity observed for different alternatives. We will therefore consider various effects of the alternative on the mean-variance relationship controlled by these parameters.

6.2 Power for a Specified Alternative

One potential problem with any mean-variance relationship is appropriate power calculations, because the mean and the final information are linked. In the model we are considering, a positive slope ($\beta_1 > 0$) would lead to lower final information, and therefore less power than what would be true if the information were the same as under the strong null hypothesis. Similarly, a negative slope would lead to greater information and therefore more power than what would be true if the information were the same as under the null. Properly accounting for this difference in information is the first step toward calculating correct power in a longitudinal sequential trial with a mean-variance relationship.

The longitudinal nature of the data adds another dimension to the mean-variance relationship in this setting. As calendar time increases during the study, more measurements will be made on each subject, which will have increasing x values. With a constant, positive slope, measurements made at these study times further from randomization are more variable, and thus the alternative (of $\beta_1 > 0$) leads to heteroscedasticity while the null ($\beta_0 = 0$) does not. Chapter 5 explored the fact that different amounts of heteroscedasticity

lead to different information growth curves, and here we have a similar situation. Different alternatives will lead to different information growth patterns, thus leading to a situation in which the information growth depends on the mean – a mean-information growth (mean-IG) relationship.

Table 6.1: Distribution of observed study times at each interim analysis under two accrual scenarios. The proportion of the final amount at each study time is given.

	0	1	2	3	4	5
Scenario 1 – 2 year accrual						
$t_1 = 1.75$	0.87	0.37	0.00	0.00	0.00	0.00
$t_2 = 3.5$	1.00	1.00	0.75	0.25	0.00	0.00
$t_3 = 5.25$	1.00	1.00	1.00	1.00	0.62	0.12
$t_4 = 7$	1.00	1.00	1.00	1.00	1.00	1.00
Scenario 2 – 5 year accrual						
$t_1 = 4$	0.80	0.60	0.40	0.20	0.00	0.00
$t_2 = 6$	1.00	1.00	0.80	0.60	0.40	0.20
$t_3 = 8$	1.00	1.00	1.00	1.00	0.80	0.60
$t_4 = 10$	1.00	1.00	1.00	1.00	1.00	1.00

If using ordinary least squares regression, rather than weighted least squares with efficient, known weights, the problem of a potential lack of independent increments is present as well. This is problematic for calculating the true power, because, as previously discussed, the standard methods for integrating the sampling density of a sequential design rely on the assumption of independent increments. However, the case of a mean-variance relationship is different than the case of a predictor-variance relationship, because under the strong null in the mean-variance setting there is no heteroscedasticity, and independent increments will be true when using ordinary least squares. However, when using ordinary least squares

under alternatives, there will be a change in the variance of the observations over study time and thus there is the potential for non-independent increments when using ordinary least squares regression.

To illustrate the effect of these factors in a group sequential design, we considered two accrual scenarios. In both scenarios, measurements are made at baseline and then study times 1-5. In the first scenario, individuals are accrued uniformly over two years so the entire study takes place over 7 years in calendar time. Four analyses are spaced equally in calendar time. In the second scenario, individuals are accrued uniformly over 5 years so that accrual is longer relative to the length of follow up. For this scenario, four analyses are planned at calendar times 4, 6, 8, and 10. The fraction of total measurements that have been observed for each study time at the interim analyses under these scenarios are shown in table 6.1. For the specific case of 100 individuals, $\gamma = 2$, and $\beta_0 = 10$, table 6.2 shows final information and the information growth under the null ($\beta_1 = 0$) and a specific alternative $\beta_1 = 1$. This comparison shows that under both scenarios, there are differences in the final amount of information ($I_{null} \neq I_{alt}$) and differences in the growth of the information over time. We consider the effect of the mean-final information relationship, the mean-information growth relationship and non-independent increments on the true power under these accrual and measurement time scenarios and under a variety of different strengths of the mean-variance relationship.

6.2.1 The Mean-Final Information Relationship

When examining the effect of the mean-final information relationship on the power of a group sequential design, we note that this issue is not unique to the group sequential setting. Specifically, the standardized z-statistic depends on both the observed value and the (estimated) standard error of the statistic which in turn depends on the true value of the parameter.

$$z = \frac{(\hat{\theta} - \theta_0)}{\sqrt{\widehat{Var}_\theta(\hat{\theta})}}$$

Table 6.2: The information growth under the null ($\beta_1 = 0$) and the alternative ($\beta_1 = 1$) using the accrual and analysis time schedule of scenario 1 and scenario 2. The relative amount of information under the null and the alternative ($\frac{I_{alt}}{I_{null}}$) at each analysis is also given.

	Scenario 1				Scenario 2			
Analysis Time	Calendar Time	IG_{null}	IG_{alt}	$\frac{I_{\text{alt}}}{I_{\text{null}}}$	Calendar Time	IG_{null}	IG_{alt}	$\frac{I_{\text{alt}}}{I_{\text{null}}}$
1	1.75	0.015	0.021	0.872	4	0.119	0.139	0.723
2	3.5	0.156	0.189	0.749	6	0.500	0.500	0.620
3	5.25	0.550	0.580	0.654	8	0.814	0.814	0.621
4	7	1.000	1.000	0.620	10	1.000	1.000	0.620

However, in most cases with a fixed sample test the sample size is large enough that the difference in standard error of the statistic under the strong null and under a hypothesized alternative is small and thus the difference in power due to a mean-variance relationship is small. However, notable differences in power may be an issue with strong mean-variance relationships and moderate sample sizes.

The Unadjusted Approach

To illustrate the effects of this potential problem in the group sequential setting, we first computed stopping boundaries for a fixed number of measurements and analysis times when assuming no mean-variance relationship. These boundaries are called "unadjusted" and would suffice if there were no mean-variance relationship. However, if the mean-variance relationship is as in equation 6.1, the variability of the test statistic increases under a positive alternative ($\beta_1 > 0$). Thus, the alternative that was calculated to have a specific power (we will use 97.5% for illustration) under the assumption of no mean-variance relationship will no longer have this power in truth. For this unadjusted approach:

- Boundaries are constructed using the true information growth under the null.

- At each interim analysis, the z-statistic for the interim analysis is compared to the boundaries constructed during the design phase.

- The alternative with 97.5% power for comparison was calculated from the design that assumes that the final information and information growth are the same as under the null and that the covariance matrix of the interim statistics has an independent increment structure.

Tables 6.3-6.6 show the alternative with assumed 97.5% power when assuming no mean-variance relationship ("Unadjusted"). These tables also show the empirical power (from 1,000,000 simulated trials) at each of these alternatives. The empirical power was calculated assuming that the z-scale boundaries derived for the null distribution would be used throughout. Assuming that the null is as anticipated and that the form of the variance is correctly specified, use of these boundaries will maintain the type I error rate of the study.

As might be expected, increasing the mean-variance relationship by increasing the γ parameter increases the effect of ignoring the mean-variance relationship. For example, using scenario 1, an O'Brien-Fleming design, 100 individuals, and a baseline value of $\beta_0 = 10$, at the alternative that has 97.5% power given no mean-variance relationship, we truly have 96.6% power when $\gamma = 1$. If instead $\gamma = 2$, at the alternative with 97.5% power given no mean-variance relationship, we truly have only 88.1% power. Increasing γ to even more extremes, $\gamma = 2.25$, at the alternative with 97.5% power given no mean-variance relationship, we only have 80.2% power (table 6.3). This decrease in power for a specific alternative due to the mean-variance relationship is more extreme with smaller sample sizes. As in the case of fixed sample tests, greater sample sizes reduce the value of the alternative for which the design has 97.5% power, which in turn reduces the size of the difference in the variance under the null and under this alternative.

The baseline value of the mean, β_0, influences the effect of the mean-variance relationship on the power as well. In this setting, the impact of the baseline value affects the calculated alternative, because the lower value of β_0 means that the outcome measurements are less variable and thus the alternative with 97.5% power is smaller. This in turn means that there is less of a difference in the variance under the null and under the alternative, similar to

Table 6.3: Alternative believed to have 97.5% under the assumption of no mean-variance relationship (unadjusted), adjusting for the mean-final information relationship only (mean-FI) and adjusting for the mean-information growth relationship (mean-IG). The empirical power of the alternative when using stopping boundaries derived under the null is also shown.

Scenario 1: $\beta_0 = 10$, $n = 100$							
		Unadjusted		Adjusted for Mean-FI		Adjusted for Mean-IG	
		Alt	Power	Alt	Power	Alt	Power
OBF	$\gamma = 1$	0.2990	0.9659	0.3104	0.9751	0.3102	0.9750
	$\gamma = 2$	0.9454	0.8805	1.2762	0.9764	1.2717	0.9758
	$\gamma = 2.25$	1.2607	0.8021	2.2127	0.9779	2.1963	0.9770
Pocock	$\gamma = 1$	0.3674	0.9676	0.3846	0.9760	0.3814	0.9746
	$\gamma = 2$	1.1617	0.8842	1.7214	0.9852	1.5899	0.9764
	$\gamma = 2.25$	1.5492	0.7943	3.5227	0.9928	2.8759	0.9807
Scenario 1: $\beta_0 = 10$, $n = 500$							
OBF	$\gamma = 1$	0.1337	0.9710	0.1359	0.9751	0.1359	0.9750
	$\gamma = 2$	0.4228	0.9426	0.4751	0.9755	0.4745	0.9752
	$\gamma = 2.25$	0.5638	0.9192	0.6809	0.9758	0.6794	0.9754
Pocock	$\gamma = 1$	0.1643	0.9718	0.1677	0.9753	0.1671	0.9747
	$\gamma = 2$	0.5196	0.9476	0.6019	0.9790	0.5852	0.9749
	$\gamma = 2.25$	0.6929	0.9247	0.8831	0.9817	0.8417	0.9754

Table 6.4: Alternative believed to have 97.5% after adjusting for various assumptions and the empirical power of these alternatives when using stopping boundaries derived under the null.

Scenario 1: $\beta_0 = 30$, $n = 100$							
		Unadjusted		Adjusted for Mean-FI		Adjusted for Mean-IG	
		Alt	Power	Alt	Power	Alt	Power
OBF	$\gamma = 1$	0.5178	0.9698	0.5291	0.9751	0.5290	0.9750
	$\gamma = 2$	2.8362	0.8805	3.8285	0.9764	3.8151	0.9758
	$\gamma = 2.25$	4.3389	0.7658	8.8834	0.9787	8.7933	0.9778
Pocock	$\gamma = 1$	0.6363	0.9707	0.6534	0.9754	0.6502	0.9746
	$\gamma = 2$	3.4852	0.8842	5.1642	0.9852	4.7696	0.9764
	$\gamma = 2.25$	5.3318	0.7516	16.7941	0.9957	12.0386	0.9836
Scenario 1: $\beta_0 = 30$, $n = 500$							
OBF	$\gamma = 1$	0.2316	0.9728	0.2338	0.9750	0.2338	0.9750
	$\gamma = 2$	1.2683	0.9426	1.4254	0.9755	1.4235	0.9752
	$\gamma = 2.25$	1.9403	0.9076	2.4246	0.9761	2.4182	0.9756
Pocock	$\gamma = 1$	0.2846	0.9730	0.2880	0.9751	0.2873	0.9747
	$\gamma = 2$	1.5587	0.9476	1.8056	0.9790	1.7556	0.9749
	$\gamma = 2.25$	2.3845	0.9128	3.1839	0.9829	3.0043	0.9757

Table 6.5: Alternative believed to have 97.5% after adjusting for various assumptions and the empirical power of these alternatives when using stopping boundaries derived under the null.

Scenario 2: $\beta_0 = 10$, $n = 100$							
		Unadjusted		Adjusted for Mean-FI		Adjusted for Mean-IG	
		Alt	Power	Alt	Power	Alt	Power
OBF	$\gamma = 1$	0.3033	0.9654	0.3150	0.9750	0.3150	0.9750
	$\gamma = 2$	0.9590	0.8767	1.3017	0.9753	1.3012	0.9753
	$\gamma = 2.25$	1.2789	0.7951	2.2752	0.9758	2.2706	0.9755
Pocock	$\gamma = 1$	0.3545	0.9658	0.3706	0.9752	0.3695	0.9746
	$\gamma = 2$	1.1211	0.8680	1.6307	0.9792	1.5840	0.9754
	$\gamma = 2.25$	1.4950	0.7721	3.2095	0.9837	2.9634	0.9774
Scenario 2: $\beta_0 = 10$, $n = 500$							
OBF	$\gamma = 1$	0.1356	0.9709	0.1380	0.9750	0.1380	0.9750
	$\gamma = 2$	0.4289	0.9416	0.4829	0.9750	0.4829	0.9750
	$\gamma = 2.25$	0.5719	0.9172	0.6930	0.9751	0.6929	0.9750
Pocock	$\gamma = 1$	0.1586	0.9710	0.1618	0.9749	0.1615	0.9746
	$\gamma = 2$	0.5015	0.9409	0.5776	0.9764	0.5718	0.9747
	$\gamma = 2.25$	0.6688	0.9138	0.8436	0.9777	0.8286	0.9749

Table 6.6: Alternative believed to have 97.5% after adjusting for various assumptions and the empirical power of these alternatives when using stopping boundaries derived under the null.

Scenario 2: $\beta_0 = 30$, $n = 100$							
		Unadjusted		Adjusted for Mean-FI		Adjusted for Mean-IG	
		Alt	Power	Alt	Power	Alt	Power
OBF	$\gamma = 1$	0.5253	0.9697	0.5369	0.9750	0.5369	0.9750
	$\gamma = 2$	2.8771	0.8767	3.9052	0.9753	3.9036	0.9753
	$\gamma = 2.25$	4.4014	0.7575	9.1925	0.9761	9.1605	0.9758
Pocock	$\gamma = 1$	0.6140	0.9697	0.6299	0.9750	0.6289	0.9747
	$\gamma = 2$	3.3632	0.8680	4.8920	0.9792	4.7520	0.9754
	$\gamma = 2.25$	5.1450	0.7285	14.5526	0.9860	12.7464	0.9788
Scenario 2: $\beta_0 = 30$, $n = 500$							
OBF	$\gamma = 1$	0.2349	0.9727	0.2372	0.9750	0.2372	0.9750
	$\gamma = 2$	1.2867	0.9416	1.4488	0.9750	1.4487	0.9750
	$\gamma = 2.25$	1.9684	0.9049	2.4695	0.9751	2.4689	0.9751
Pocock	$\gamma = 1$	0.2747	0.9726	0.2779	0.9748	0.2776	0.9747
	$\gamma = 2$	1.5046	0.9409	1.7329	0.9764	1.7155	0.9747
	$\gamma = 2.25$	2.3017	0.9003	3.0338	0.9783	2.9682	0.9751

the effect of the sample size above. However, for the same alternative, the smaller baseline value would lead to a greater relative change in the variance, which in turn should lead to greater departures from the assumed power. For example, for scenario 1 with $n = 100$ and an O'Brien-Fleming design, the alternative assumed to have 97.5% power when $\gamma = 1$ and $\beta_0 = 10$ is 0.303. When $\beta_0 = 30$, this alternative is 0.525. However, despite the larger alternative, the effect of ignoring the mean-variance relationship on the power is slightly smaller when $\beta_0 = 30$ (power = 97.0%) than when $\beta_0 = 10$ (power=96.6%), due to the fact that this alternative is small enough that the relative difference in variance under the null and alternative when $\beta_0 = 30$ is small. This trend is seen with $\gamma = 1$ under other designs and scenarios as well. However, when γ is increased to 2.25, the alternative when $\beta_0 = 30$ is so much larger than when $\beta_0 = 10$ (e.g. 4.34 vs. 1.28) that the difference in relative variances under the null and the alternative is now larger when $\beta_0 = 30$, leading to lower power compared to the case when $\beta_0 = 10$ (76.6% vs. 79.5%).

The difference between the Pocock and O'Brien-Fleming designs on the effect of the mean-final information on the power is similarly two competing processes. For the same fixed maximal sample size and analysis times, the Pocock design has lower power than the O'Brien-Fleming design and thus the alternative with 97.5% power is larger, causing a greater effect of the mean-variance relationship. However, because the Pocock boundaries are less conservative early, when the difference in the variance of the outcome measures under the alternative and the null is less extreme, this can lead to the Pocock design losing less power than would be expected for a fixed value of the alternative. Additionally, the (incorrect) assumption that the information growth is the same for the null and the alternative has differential effects on the Pocock and O'Brien-Fleming designs which would impact the loss of power from assuming no mean-variance relationship as well. The impact of the incorrectly estimated information growth is discussed in the next section.

The impact of the assumption of no mean-variance relationship is different in the two accrual scenarios. In order to observe this, we need to consider two designs with similar alternatives that would have 97.5% power if there were no mean-variance relationship. We consider the case of $\beta_0 = 30$, $n = 100$, and $\gamma = 2$. The alternative at which a Pocock design would have 97.5% power if there were no mean-variance relationship in this setting

is 3.49 under scenario 1 (table 6.4) and 3.36 under scenario 2 (table 6.6). If there were no effect of the accrual and timing of analyses, we would expect a greater loss of power under scenario 1 due to the increased variability of the alternative. However, the true power for these alternatives is is 88.4% under scenario 1 and 86.9% under scenario 2. Despite the smaller variance under the scenario 2 alternative, scenario 2 leads to a larger decrease in power; this difference is is due to differential effects of misspecifying the information growth for the alternative.

Method of Adjustment

Tables 6.3-6.6 also show the calculated alternative when correcting for the difference in final information but still assuming a constant information growth and independent increments ("Adjusted for Mean-FI").

To account for the mean-final information relationship, methods analogous to those for a fixed sample test were used. Specifically, the z-statistic critical value with 97.5% power is used to solve for the alternative θ_a while accounting for the mean-variance relationship. The straightforward method is to solve the equation

$$(z^*)^2 = \frac{(\theta_a - \theta_0)^2}{Var_{\theta_a}(\hat{\theta})/n} \tag{6.2}$$

for θ_a, where the relationship between the $Var_{\theta_a}(\hat{\theta})$ and θ_a is known (or presumed known). The value of z^* is obtained from the group sequential design with information growth as would be true under the null. If there were no interim analyses, z^* would be the critical value from the standard normal distribution; with $\alpha = 0.025$, it would be equal to 1.96. However, as discussed in chapter 2, for a sequential design, the sampling density is affected by the choice of stopping boundaries and thus this critical value is calculated by numerical integration of the sampling density.

In a group sequential trial, if we assume both that the information growth is constant for the different alternatives and that an independent increment structure is present, this adjustment can be done readily with standard software. For this method, we do not change the boundaries from the "unadjusted" design, we only revise our estimate of the alternative

for which the design has 97.5% power to account for the mean-variance relationship. We call this "adjusting for the mean-final information." For this approach, we have:

- Boundaries are constructed using the true information growth under the null.
- At each interim analysis, the z-statistic for the interim analysis is compared to the boundaries constructed during the design phase.
- The alternative with 97.5% power for comparison was calculated by solving equation 6.2 using the z-statistic with 97.5% power from the design under the null as z^* and the known mean-variance relationship to calculate $Var_{\theta_a}(\hat{\theta})$ at the end of the study.

As expected, adjusting for the increased variance due to the mean-variance relationship increases the magnitude of the alternative with 97.5% power. The amount of adjustment needed depends on the strength of the mean-variance relationship. If there were no additional problems (i.e. if the assumptions about a constant information growth and independent increments under the mean and alternative were true), this adjustment would correctly calculate the alternative at which the design has 97.5% power. However, despite adjusting for the effect of the mean-variance relationship on the total information, issues remain with calculating the correct alternative in this setting, which are discussed in the next section.

6.2.2 The Mean-Information Growth Relationship

In our setting of a longitudinal group sequential trial, a mean-variance relationship will also impact the information growth. As observed in chapter 5, differences in the amount of heteroscedasticity lead to different information growth curves (figure 5.1). In the case of a mean-variance relationship, we have the additional complication that the information growth will be different under the null and the alternative due to different amounts of heteroscedasticity in each case (e.g. table 6.2).

In general, with the setting of a positive alternative and a mean-variance relationship whereby a larger mean generates more variable data, adjusting only for the mean-final

information will overcorrect the alternative to one with greater than 97.5% power. The mean-variance relationship in these settings means that the true information fraction grows faster under the alternative than under the null. This is due to the increasing variance of measurements later in study time under the alternative, thereby making the fractional amount of information present at earlier analysis times greater. Once the effect of the mean-variance relationship on the final information is accounted for in computing an alternative, assuming the information growth for this alternative is identical to the information growth under the null is a simple case of underestimating the true information growth.

Chapter 4 illustrated the consequences of overestimating the information growth with a naive estimate of the information growth in a longitudinal setting. Overestimating the information growth leads to a loss of power in a two-boundary design with early stopping for futility. The opposite is true here: Underestimating the information growth leads to an increase in power due to interim boundaries that are too conservative for the true information growth under the alternative. These conservative boundaries increase the power by reducing the type II error spending at interim analyses without a corresponding increase at the end of the study. In this way, trials that would have been stopped early for futility under the correct information growth (as part of the type II error) are not stopped early and may go on to be declared successes. Thus, the actual empirical power for alternatives calculated adjusting for the mean-final information relationship have greater than the expected 97.5% power.

The timing of the analyses and the boundaries used (either O'Brien-Fleming or Pocock) will have an impact on how much the different information growth curves affect the analysis times. Differences between the information growths under the null and alternative hypotheses are most pronounced during the middle of the study (e.g. figure 5.1). Therefore, if all interim analyses were to take place at very early information times or very late information times, the differences in information growth under the null and the alternative could be less pronounced, and adjusting only for the effect of the mean-variance relationship on the final information would likely be adequate to calculate the alternative at which the design has 97.5% power.

Similarly, the conservatism of the boundaries at different analysis times will also impact

the effect of missestimating the information growth under the alternative. For a fixed information growth, O'Brien-Fleming boundaries are more conservative at early information times than are Pocock boundaries. This early conservatism means that even under the alternative, there is very little possibility of stopping at early analyses for futility. Thus, underestimating the true information at early analysis times does not have as dramatic an effect on the power for O'Brien-Fleming designs as it does for Pocock ones.

To see this conclusion from our simulations, it is important to note that the alternatives are different under O'Brien-Fleming and Pocock designs, and the larger alternatives will have a greater difference in information growth under the alternative compared to the null. For example, in scenario 1, with $\beta_0 = 10$ and $n = 100$ (table 6.3), the alternative for the O'Brien-Fleming design with $\gamma = 2.25$ is $\beta_1 = 2.21$, and the alternative for the Pocock design with $\gamma = 2$ is $\beta_1 = 1.72$. In all respects other than the choice of boundary, the alternative under O'Brien-Fleming would be expected to have greater departures from the nominal power level – greater γ values will lead to greater departures in the information growth and the the greater alternative will also increase the mean-variance relationship. Yet, the empirical power under O'Brien-Fleming is 97.8%, while it is 98.5% under Pocock for this scenario.

It is important to note that the number of individuals does not affect the differences in information growth curves between the null and the alternative because the information is a scaled multiple of the number of individuals (in our setting with fixed accrual and measurement times). The differences in power departures at the alternative with assumed 97.5% power seen for different sample sizes is due to the smaller alternative with the larger sample size reducing the amount of the mean-variance relationship.

Method of Adjustment

To account for the mean-IG relationship in calculating the alternative with 97.5% power, we again assume that the true form of the mean-variance relationship is exactly known (or presumed known). Then, instead of simply correcting the calculated z-statistic for the mean-final information relationship at the end of the study, a search is conducted for the

alternative that corrects both for the mean-final information relationship and the change in the information growth with this alternative. As before, boundaries are calculated and fixed under the null. Then we have:

- Boundaries are constructed using the true information growth under the null.
- At each interim analysis, the z-statistic for the interim analysis is compared to the boundaries constructed during the design phase.
- The alternative with 97.5% power for comparison was calculated by solving equation 6.2 as in the adjustment for the mean-final information, but this time uses the z-statistic with 97.5% power from the design that uses the exact boundaries calculated above but correctly specifies the information growth under the alternative (θ_a).

If the alternative yielded independent increments for the covariance matrix of the interim statistic, this method would yield the alternative with 97.5% power using fixed z-statistic boundaries under the null.

Tables 6.3-6.6 show the calculated alternative when correcting for the difference in final information and information growth but still assuming independent increments ("Adjusted for Mean-IG"). These alternatives are generally less than the alternatives that only adjusted for the change in final information under the alternative.

6.2.3 Independent Increment Concerns

Finally, we note that correcting for the mean-final information and the mean-information growth relationship still does not always yield alternatives with exactly 97.5% power. This difference is due to a violation of the independent increment assumption under the alternative. Using weighted least squares, correcting for the mean-final information and mean-information growth, does yield calculated alternatives with exactly 97.5% power.

As discussed in chapter 5, the degree of departure from independent increments will depend here on the strength of the mean-variance relationship. However, we note that for these scenarios, the effect on the nominal power is generally small, even for very strong mean-variance relationships.

Method of Adjustment

If the mean-variance relationship were so extreme that adjustment for non-independent increments were needed, then an adjustment could be accomplished by adding an iterative step to the search in the adjustment for the mean-information growth relationship. Here, instead of computing the true power of the alternative by integrating the sampling density with the true information growth and the exact constraint, we would need to compute the true power by simulation of data with this alternative and the constrained boundaries.

6.3 Design Properties

Another potential consequence of the mean-variance relationship on the information growth is on the stopping probabilities at each analysis time under the alternative for given designs. This in turn leads to different operating characteristics of a design in terms of average sample size than what would be expected if there were a constant variance.

Even if we correctly calculate the alternative at which there was 97.5% power, this alternative may not yield a pattern of stopping probabilities that is expected for the alternative with 97.5% power under a particular design. In error spending parlance, neither the "type II error spending" nor the "power spending" functions will behave exactly as would be expected for a given choice of boundary due to the mean-information growth relationship.

For example, consider the case of an accrual and measurement pattern like that in scenario 1. Further, let $\beta_0 = 10$, $n = 100$, and $\gamma = 2$. As seen in table 6.2 the information growth under the null in this case is $\boldsymbol{\pi} = (0.015, 0.156, 0.550, 1)$. Using an O'Brien-Fleming design, and adjusting for the mean-final information relationship, we calculate that the alternative at which have 97.5% power is $\beta_1 = 1.276$ (table 6.3). If this alternative had the same information growth as the information growth under the null, then we would truly have 97.5% power for this alternative. However, as discussed above, the information growth under this alternative is slightly faster than the information growth under the null and as such, using boundaries computed assuming the null information growth, yields slightly higher power for this alternative.

Using the information growth under the null, properties of the expected O'Brien-Fleming

design are shown in table 6.7. However, the change in information growth under the alternative leads to both a change in the power of this alternative (to 97.6%) and a change in the stopping probabilities at each analysis time (table 6.8). Here we see the effect of the underestimated information from the null design on the power and type II error rate for the alternative with presumed 97.5% power. Under both alternatives, the true probability of stopping for efficacy at the third interim analysis is greater than anticipated for an O'Brien-Fleming design with information growth that is true under the null. This result is because the efficacy boundary at the third analysis is designed to maintain the type I error spending at the third analysis for the information growth under the null. Under the alternative, there is more relative information at the third analysis time. Therefore under the alternative, the trial is more likely to stop for efficacy at the third analysis than would occur for an alternative with the same relative information at the third analysis as the null hypothesis (no mean-IG relationship). Similarly, the alternative is less likely to stop for futility (type II error) at the third interim analysis as it has more information than expected. Note that the percent of the type II error spent does not reach 100 because the power for this alternative is greater than than nominal level using the unadjusted boundaries.

Table 6.7: Probability of stopping (SP) for the null or alternative (with 97.5% power) at each of the four analyses under the null and the alternative with 97.5% power if there were no mean-IG relationship, with scenario 1, $\gamma = 2$, $\beta_0 = 10$, and $n = 100$.

Properties if no Mean-IG Relationship							
		Null			Alternative		
Analysis	IG	% Error	SP_{null}	SP_{alt}	% Error	SP_{null}	SP_{alt}
1	0.015	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.156	0.0000	0.0003	0.0000	0.0000	0.0000	0.0003
3	0.550	0.1531	0.6044	0.0038	0.1531	0.0038	0.6045
4	1.000	1.0000	0.3703	0.0212	1.0000	0.0212	0.3703
ASN		523.9			523.9		

Table 6.8: Stopping probabilities at each of the four analyses under alternatives with assumed 97.5% power (accounting for mean-FI) with scenario 1, $\gamma = 2$, $\beta_0 = 10$, and $n = 100$.

Using Unadjusted Boundaries								
	Null				Alternative			
Analysis	IG	% Error	SP_{null}	SP_{alt}	IG	% Error	SP_{null}	SP_{alt}
1	0.015	0.000	0.0000	0.0000	0.023	0.000	0.0000	0.0000
2	0.156	0.000	0.0003	0.0000	0.197	0.000	0.0000	0.0006
3	0.550	0.156	0.6044	0.0039	0.587	0.114	0.0029	0.6417
4	1.000	0.997	0.3704	0.0210	1.000	0.946	0.0208	0.3340
ASN	523.9				519.2			
Using Futility-Adjusted Boundary								
	Null				Alternative			
Analysis	IG	% Error	SP_{null}	SP_{alt}	IG	% Error	SP_{null}	SP_{alt}
1	0.015	0.000	0.0000	0.0000	0.023	0.000	0.0000	0.0000
2	0.156	0.000	0.0035	0.0000	0.197	0.000	0.0000	0.0006
3	0.550	0.156	0.6703	0.0039	0.587	0.199	0.0050	0.6417
4	1.000	0.998	0.3012	0.0211	1.000	0.987	0.0197	0.3330
ASN	514.7				519.0			

6.3.1 Futility-Adjusted Boundaries

Another option in the scenario of a mean-variance relationship is to change the futility boundary at the design stage to account for the different information growth under an alternative. As the futility boundary is intended to reject the alternative hypothesis, it makes sense to modify the futility boundary to account for the information growth under the alternative and try to recapture some of the intended behavior of the original boundary

design. To accomplish this goal, we use the constrained boundary approach again. For the alternative that is adjusted for the mean-final information as above (which corrects for the mean-variance relationship in the final analysis), we calculate the information growth under this alternative. We then compute the standard O'Brien-Fleming design for the information growth under this alternative. The futility boundary from this design is then used with the efficacy boundary from the design under the null and these boundaries are fixed for all interim analyses. At the last analysis, the constrained boundary approach is used with the information growth under the null to ensure that the correct type I error rate will be maintained.

For this case we have:

- The efficacy boundary is determined using the information growth under the null.
- The futility boundary is determined using the information growth under the alternative with 97.5% power.
- The boundaries are merged using a constrained boundary approach and the information growth under the null to maintain type I error rate.
- At each interim analysis, the z-statistic for the interim analysis is compared to the boundaries constructed during the design phase.
- The alternative with 97.5% power is the alternative from the design under then null with 97.5% power after adjusting for the mean-final information relationship.

Boundaries resulting from this approach for the O'Brien-Fleming design in the example above (scenario 1, $\beta_0 = 10$, $n = 100$, and $\gamma = 2$) are shown in table 6.9. The faster information growth under the greater alternative is reflected in stopping boundaries for futility that are less extreme.

The results from using this futility-adjusted boundary are shown in table 6.8. The design resulting from this approach does not perfectly recapture the desired type II error spending from an O'Brien-Fleming design, but it does come closer than the unadjusted boundaries.

Table 6.9: Sample boundaries using null information growth (unadjusted) and using a constrained boundary approach to adjust the futility boundary an O'Brien-Fleming design.

	Unadjusted		Futility-Adjusted	
Analysis #	a	d	a	d
1	-15.599	16.085	-12.415	16.085
2	-3.446	5.007	-2.701	5.007
3	0.266	2.667	0.450	2.667
4	1.977	1.977	1.970	1.970

Note that this approach also modifies the expected stopping probabilities under the null as it makes it more likely for the design to stop earlier for futility under the null. This increase in early stopping for futility results in a lower ASN.

6.4 Confidence Intervals

A final area of concern relating to the mean-variance relationship in this setting is post-trial inference. We focus here specifically on the construction of confidence intervals, although the issues are similar for all post-trial inferences that rely on the correct estimation of the sampling density.

As noted previously, confidence intervals following a group sequential test can be constructed by inverting hypothesis tests, such that the the 95% confidence set is given by:

$$CI\{(M,S)=(m,s)\} = \left\{\theta : \frac{\alpha}{2} \leq P((M,S) \leq (m,s)|\theta) \leq 1-\frac{\alpha}{2}\right\} \tag{6.3}$$

The difficulty in the mean-variance setting is the correct estimation of the sampling density as it changes under each potential value of θ. As explored above, the mean-variance relationship can affect the sampling density through (a) the mean-final information relationship, (b) the mean-information growth relationship and (c) the lack of independent increments if using ordinary least squares instead of weighted least squares.

6.4.1 Simulated Quantile Method

To combat all three of these issues, we use a simulated quantile method for constructing confidence intervals. This method will be robust to all three potential issues above, but does require knowledge of the mean-variance relationship. Conceptually, the simulated quantile approach generates an empirical version of the sampling density over a broad range of possible values of θ. Based on the sampling density under a particular value of θ, quantiles of the observed statistic $\hat{\theta}$ are calculated. For 95% confidence intervals under the sample mean and analysis time order, we calculate the 2.5% and 97.5% quantiles under each ordering.

Ideally, the relevant quantiles could be simulated for all possible values of θ, however clearly this is both theoretically and practically impossible. Instead, we chose θ values at regular intervals and used a loess smoother to extrapolate the quantiles to all values of θ (see figure 6.1 for an example using an O'Brien-Fleming design with scenario 1, $\beta_0 = 10$, $n = 100$, and $\gamma = 2$). An *ad hoc* adjustment was made to calculate more values where the transition from stopping at one analysis time to the next occurred. The location of this transition is critically important for the analysis time ordering and is a place of a large amount of change on the sample mean scale as well. By choosing more values of θ in the intervals in which such jumps occurred, the exact value of θ for which the quantile occurred at a new stopping time could be more precisely determined.

Once the quantile smoothers are well-estimated, confidence intervals can be determined based on a particular observed value $(m, \hat{\theta})$. At any particular observed value, the confidence set includes all values of θ for which the observed value would not be unusual (i.e. the observed value was greater than the 2.5 percentile but less than the 97.5 percentile of all observed values for that value of θ). Figure 6.1 provides an illustration of this method using the sample mean ordering. If the value observed were $\hat{\theta} = 0.5$ and the trial stopped for efficacy at the final analysis (m=4), the dashed line represents the observed value. The values of θ that are included in the confidence set are those in which the observed value of $\hat{\theta}$ is between the two quantiles. In this case, the bottom of the confidence interval is the value of θ for which the 97.5 percentile is equal to the observed value. This is seen as when the

horizontal line crosses the line showing the 97.5 percentile as a function of the true value of the slope. The upper limit of the confidence interval is the value of θ for which the 2.5 percentile is equal to the observed value, which is observed when our horizontal line for the observed value crosses the line showing the 2.5 percentile as a function of the true value. The situation is analogous for the analysis time ordering, except that in this case less than and greater than are defined both as a function of the observed slope value and when the trial would be stopped.

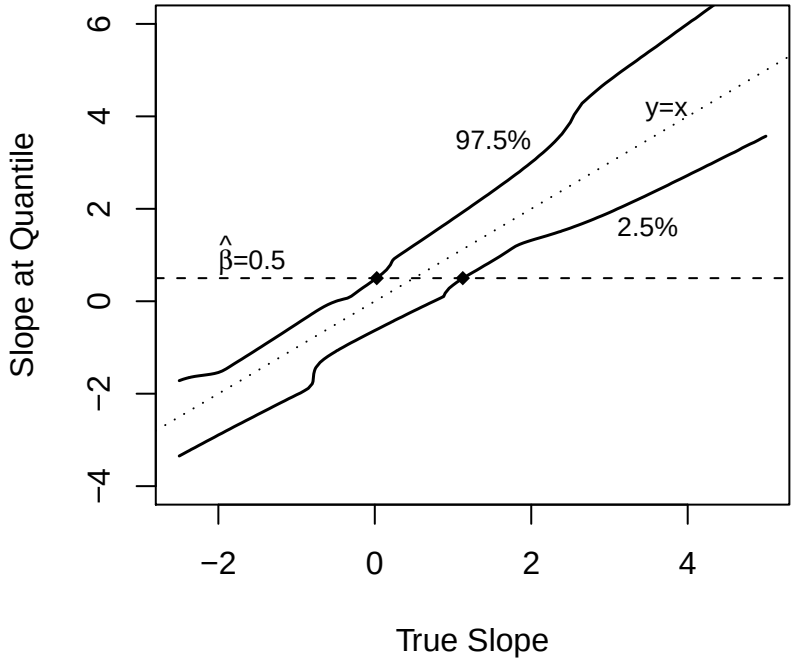

Figure 6.1: Plot illustrating the construction of confidence intervals under the sample mean ordering. The lines are the empirical 2.5% and 97.5% quantiles for various true values of the slope parameter.

The simulated quantile method for constructing confidence sets should provide appropriate coverage provided the quantiles can be well simulated, both by knowing the mean-variance relationship and by simulating the quantiles at enough true values of θ. We compared this method to three approaches:

Table 6.10: Empirical coverage probabilities (Cov.) and average length (Len.) of nominal 95% confidence intervals under the null and alternative with true 97.5% power. Scenario 1 with an O'Brien-Fleming design was used, with $\beta_0 = 10$, $n = 100$, and $\gamma = 2$.

	Under the Null $\beta_1 = 0$				**Under Alternative** $\beta_1 = 1.27$			
	Mean Order		Time Order		Mean Order		Time Order	
	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.
Unadjusted	0.958	1.17	0.958	1.19	0.842	1.23	0.842	1.22
Adjusting for Mean-FI	0.950	1.14	0.950	1.16	0.952	1.60	0.952	1.63
Adjusting for Mean-IG	0.952	1.15	0.952	1.17	0.950	1.57	0.950	1.58
Simulated Quantiles	0.959	1.18	0.958	1.25	0.940	1.49	0.946	1.69

- The naive approach which makes no adjustments for a mean-variance relationship and simply uses the sampling density as calculated under independent increments and no mean-variance relationship.

- An adjusted mean-final information approach which recalculated the sampling density using the final information that would be true for the observed slope at the end of the study.

- An adjusted mean-information growth approach which further adjusted the sampling density for the true information growth that would have been seen using the observed slope.

For illustration, we used scenario 1, with $\beta_0 = 10$, $n = 100$, and $\gamma = 2$. Under the null, when everything is correctly specified, the methods perform similarly (table 6.10).

However, under the alternative with 97.5% power, the unadjusted (naive) method performs poorly, with only 84.2% coverage. This result is directly attributable to continuing to assume the variance that was constant under the null, thereby failing to produce wider confidence intervals to account for the increased variability of the parameter under the alter-

native. Adjusting for the mean-final information relationship produces acceptable coverage in this circumstance.

Chapter 7

CORRELATED DATA

Finally, we consider the case of correlated data and its impact on standard methods for sequential clinical trials. We consider correlated data both when the data are homoscedastic and when the data are heteroscedastic.

As mentioned previously, we are considering the case in which the estimation of the slope parameter from the marginal model (using GEE) is the primary statistic of interest. The variance of the statistic from the GEE model is given by:

$$Var(\hat{\beta}) = (X^T W^{-1} X)^{-1} X^T W^{-1} V W^{-1} X (X^T W^{-1} X)^{-1} \tag{7.1}$$

where W is the "working" covariance and V is the covariance matrix of the observed values.

7.1 Model

In this case our standard model,

$$\begin{aligned} E(\boldsymbol{Y}|\boldsymbol{x}) =& \boldsymbol{\mu} \\ =& \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x} \\ Cov(\boldsymbol{Y_i}) =& \sigma^2 V(\boldsymbol{\mu}) \end{aligned}$$

has

$$\begin{aligned} V_{kk} &= (\beta_0 + \beta_1 x_k)^\gamma \\ V_{kk'} &= \rho_{kk'} * \sqrt{V_{kk} V_{k'k'}} \qquad\qquad k \neq k' \end{aligned} \tag{7.2}$$

where $\gamma \geq 0$ and β_0 and β_1 are constrained such that $(\beta_0 + \beta_1 x_k)^\gamma$ is non-negative over the range of x values observed.

Note that this model uses the same approach to possible heteroscedasticity due to a mean-variance relationship as in equation 6.1. It includes now the possibility for correlation between observation made on the same subject j.

We will consider two forms of the correlation structure, exchangeable, in which the correlation is the same on all measurements made on the same individual ($\rho_{kk'} = \rho$), and auto-regressive of order 1 (AR(1)), in which the correlation decreases for measurements made further apart in study time ($\rho_{kk'} = \rho^{|x_k - x_{k'}|}$).

7.2 Using Weighted Least Squares (Efficient, Known Weights)

As was the case with heteroscedastic data, using weighted least squares, with known, correct weight matrix ($W \propto Var(Y)$), will be the most efficient among all linear unbiased estimators in this setting by the Gauss-Markov theorem. When using the correct weights such that $W \propto Var(Y)$, the best linear unbiased estimator (BLUE) is $\hat{\beta}_w = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$. Because this estimator is efficient, the independent increment structure will hold and this situation is again similar to the case of independent longitudinal data.

However, even when using WLS with known weights, the information growth will be different for different amounts of correlation and different amounts of heteroscedasticity. We saw in chapter 5 that the amount of heteroscedasticity can impact the information growth, even when using the efficient weights. Heteroscedasticity will have a similar effect here, and the magnitude of the correlation will also impact the information growth.

As noted previously, the variance of the weighted least squares estimator is given by:

$$Var(\hat{\beta}) = (X^T W^{-1} X)^{-1} \tag{7.3}$$

Let w_{im} denote the element in the $i$ih row and the mth column of the matrix W^{-1}. In the case of simple linear regression we then have:

$$\begin{aligned} Var(\hat{\beta}) =& (X^T W^{-1} X)^{-1} \\ =& \begin{bmatrix} \sum_m \sum_i w_{im} & \sum_m \sum_i x_i w_{im} \\ \sum_m \sum_i x_i w_{im} & \sum_m \sum_i x_i w_{im} x_m \end{bmatrix}^{-1} \end{aligned}$$

Then, focusing only on the slope parameter, we have:

$$Var(\hat{\beta}_1) = \frac{\sum_m \sum_i w_{im}}{(\sum_m \sum_i w_{im})(\sum_m \sum_i x_i w_{im} x_m) - (\sum_m \sum_i x_i w_{im})^2}$$

$$= \frac{1}{\sum_m \sum_i w_{im}(x_i - \bar{x}_w)(x_m - \bar{x}_w)} \tag{7.4}$$

where $\bar{x}_w = \frac{\sum_m \sum_i x_i w_{im}}{\sum_m \sum_i w_{im}}$.

Thus, the relative amount of information at analysis time t_j relative to the final analysis t_J is given by:

$$\pi_j = \frac{\sum_m^{n_j} \sum_i^{n_j} w_{im}(t_j)(x_i - \bar{x}_w)(x_m - \bar{x}_w)}{\sum_m^{n_J} \sum_i^{n_J} w_{im}(t_J)(x_i - \bar{x}_w)(x_m - \bar{x}_w)} \tag{7.5}$$

From this equation, it is straightforward to observe that the information growth using the efficient, known weights will depend on the weights (the components of the matrix W^{-1}). These weights in turn depend on the amount of heteroscedasticity and the correlation parameters and structure – the values of $W^{-1} = (Var(Y))^{-1}$.

7.2.1 No Heteroscedasticity

If there is no heteroscedasticity (so $v_{kk} = \sigma^2$, for all k), then the information growth curves for the same measurement schedule and accrual pattern differ only with respect to the correlation. For illustration, we use scenario 1 and scenario 2 as described previously.

Figure 7.1 shows the true information growth curves when using WLS under various combinations of correlation structure and values. For the case of shorter accrual relative to follow up (scenario 1), the exchangeable information growth is very close to the information growth true when there is no correlation. This outcome is intuitively reasonable by noting that if the design is completely balanced at both analyses, then the relative amount of information is the same in the independent and the exchangeable case. However, if the design is not balanced at both analysis times, the fractional information for the exchangeable

case will be lower than for the independent, because future measurements will also be highly correlated with existing measurements, contributing to a larger percentage of information gained in the exchangeable case.

In contrast, the information growth with a true AR(1) structure is always above the information growth for independent data. This result is primarily due to the change in final information when the data have an AR(1) structure. In early analyses, data with an AR(1) structure will closely resemble an exchangeable structure (and is trivially true if there are only two measurements available on each individual). Thus the information at early analysis times is similar between data with an AR(1) structure and with an exchangeable one. However, by the final analysis, the information for data with an AR(1) structure will be less than the information for data with an exchangeable one, because of the decreased correlation of measurements further in study time. This difference in final information between the exchangeable and AR(1) cases means that the relative information at early analysis times is larger in the case of AR(1) data – at early analysis times a similar amount of information is present between exchangeable and AR(1) data, but this represents a larger fraction of the final information when the data have an AR(1) structure .

With both correlation structures, the information growth curves are most similar to the one under independence with smaller correlation values. In the case of exchangeable data, assuming the information growth under independent data will result in greater overestimation of the information growth with larger values of ρ. Likewise, with AR(1) data, assuming the information growth under independent data will result in greater underestimation of the truth with larger values of ρ (figure 7.1).

7.2.2 Heteroscedasticity

If the data are not only correlated but also heteroscedastic, then the information growth will again differ by the magnitude and structure of the correlation. We saw in chapter 5 that the true information growth using WLS with independent data differed by the amount of heteroscedasticity due to the parameter γ. We return to that issue here with heteroscedasticity due to a predictor-mean relationship as per equation 5.1.

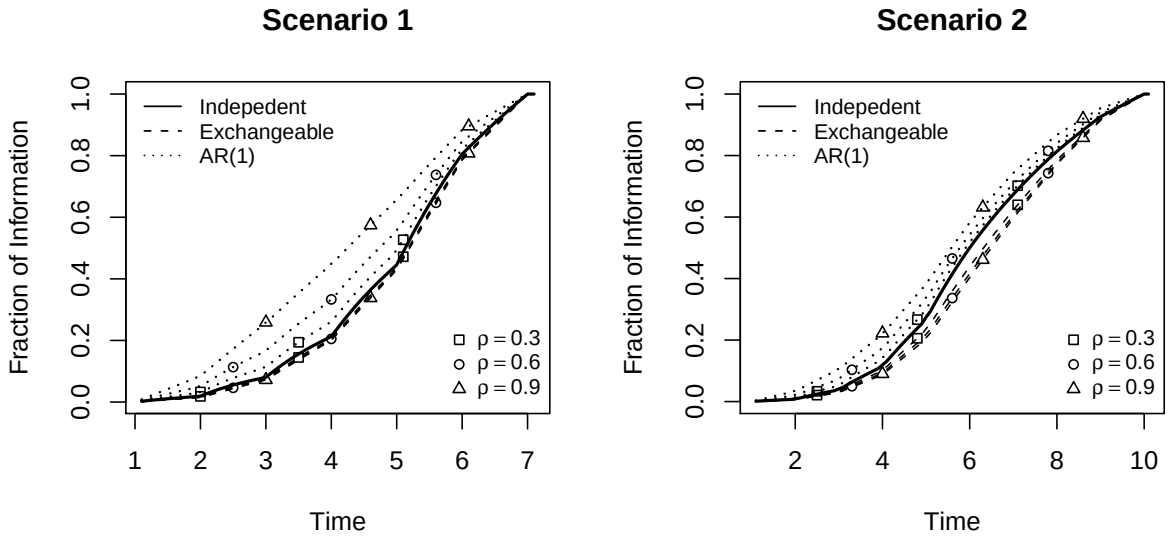

Figure 7.1: True information growth using weighted least squares with no heteroscedasticity but various amounts of correlation.

In the case of heteroscedasticity, both the exchangeable and the AR(1) correlation structures generally lead to information growth that is faster than that for independent data. Here the increased correlation at the earlier (less variable) measurement times contributes more to the overall information in both the exchangeable and AR(1) case. Thus, the information growth early in the study is higher for higher values of the correlation parameter for both structures.

Figures 7.2 and 7.3 show information growth curves for a dramatic predictor-variance relationship ($a = 1$, $b = 1$). The case in which $\gamma = 2$ (figure 7.3) hints at the limiting case for the information growth using WLS. With $\rho = 0.9$, the information growth curve increases rapidly and then levels off. For example, with $\rho = 0.9$, $\gamma = 2$ and exchangeable data in scenario 2 the fraction of information present when the calendar time is equal to 6 is 0.93, even though this is only 60% of the way through the study in calendar time (figure 7.3). This leveling off is due to significantly downweighting new measured values that are further in study time and greater in variability. In the extreme case with WLS, the

contribution to the slope of the new, more variable measurement would be nothing and the corresponding increase in the information would also be nothing, leading to an information growth curve that reached 1 before the end of the study. Such a situation is unlikely to occur in practice, as it would require dramatically more variable measurements at the end of the study compared to the beginning and high correlation among measurements.

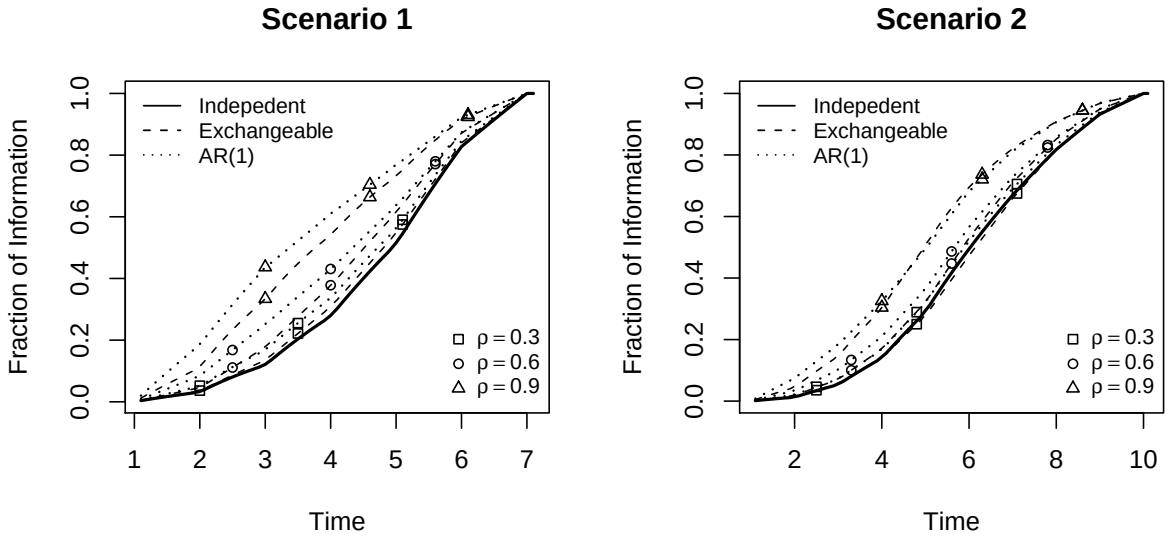

Figure 7.2: True information growth using weighted least squares with heteroscedasticity ($\gamma = 1$, $a = 1$, $b = 1$), and various amounts of correlation.

7.2.3 Consequences for Sequential Designs

As noted previously, using WLS with efficient, known weights will ensure that the independent increment assumption will hold and that the information growth will be monotonic. However, some care must be taken due to the dependence of the information growth on the correlation parameters and correlation structure, and due as well to possible heteroscedasticity from a predictor-variance relationship. If a study were planned assuming a different correlation parameter and predictor-variance relationship, then the true information growth would be different than what was planned, and the final inference should be adjusted for

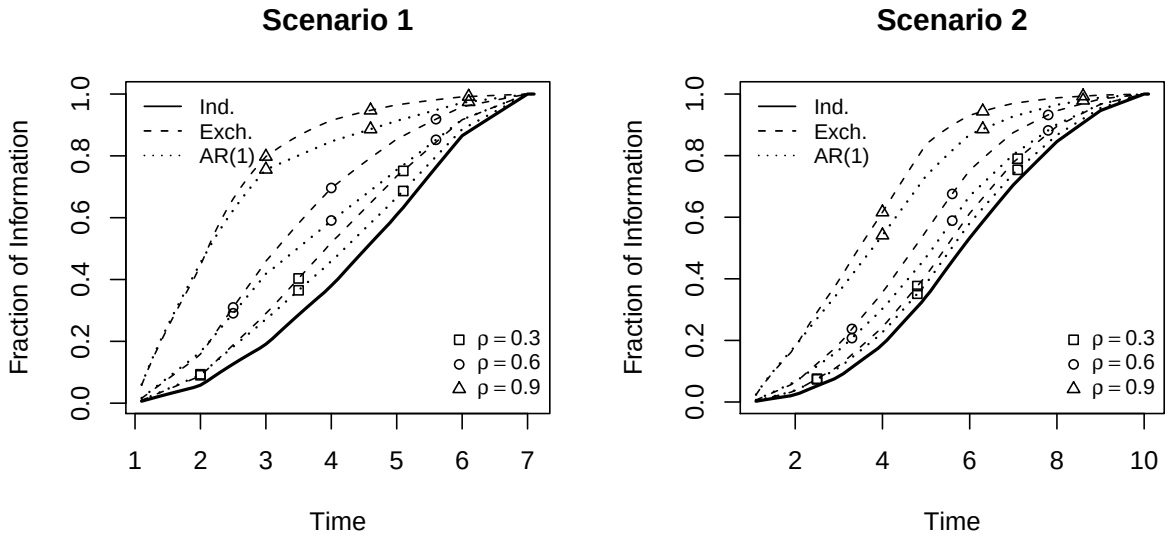

Figure 7.3: True information growth using weighted least squares with heteroscedasticity ($\gamma = 2$, $a = 1$, $b = 1$), and various amounts of correlation.

the best estimate of the information growth at the final analysis with the use of constrained boundaries, as discussed previously.

It is worth noting here that although the above results from weighted least squares assumed a known form of the covariance matrix, the results will hold asymptotically for GEE with a correctly specified working covariance matrix (both in the correlation and the possible heteroscedasticity), provided that the sample size is such that the parameters of the working covariance matrix can be estimated consistently. In this case, $W \rightarrow_p V$ and the above results will hold. We note that even in this case, the effect of the correlation on the information growth is still important, as the form of the correlation (AR(1), exchangeable, etc.) may be correctly specified in advance, but the parameter ρ need not be (but can be estimated from the data).

7.3 Using GEE with any Working Covariance

We now turn more generally to using GEE when the form of the working covariance matrix does not match the true covariance matrix exactly. For any choice of working covariance matrix, W, we can correctly specify the form of the true variance of the slope parameter. From equation 7.1, we have:

$$Var(\hat{\beta}) = (X^T W^{-1} X)^{-1} X^T W^{-1} V W^{-1} X (X^T W^{-1} X)^{-1}$$

Let w_{im} denote the elements of the inverse of the working covariance matrix W^{-1} and v_{im} denote the elements of the true covariance matrix V. Further, let w^*_{im} denote elements from the matrix $W^{-1}VW^{-1}$. Then, we have that the $Var(\hat{\beta})$ is equal to:

$$\begin{bmatrix} \sum\sum w_{im} & \sum\sum x_i w_{im} \\ \sum\sum x_i w_{im} & \sum\sum x_i w_{im} x_m \end{bmatrix}^{-1} \begin{bmatrix} \sum\sum w^*_{im} & \sum\sum x_i w^*_{im} \\ \sum\sum x_i w^*_{im} & \sum\sum x_i w^*_{im} x_m \end{bmatrix} \begin{bmatrix} \sum\sum w_{im} & \sum\sum x_i w_{im} \\ \sum\sum x_i w_{im} & \sum\sum x_i w_{im} x_m \end{bmatrix}^{-1}$$

Then, focusing only on the slope parameter, we have:

$$\begin{aligned} Var(\hat{\beta}_1) = &\left(\frac{1}{((\sum_m \sum_i w_{im})(\sum_m \sum_i x_i w_{im} x_m) - (\sum_m \sum_i x_i w_{im})^2)} \right)^2 \times \\ &\{ (\sum_m \sum_i x_m w_{im})^2 \sum_m \sum_i w^*_{im} + (\sum_m \sum_i w_{im})^2 \sum_m \sum_i x_m w^*_{im} x_i \\ &- 2(\sum_m \sum_i w_{im})(\sum_m \sum_i x_m w_{im})(\sum_m \sum_i x_m w^*_{im}) \} \end{aligned}$$

This equation can then be expanded into two parts; a first term that represents the minimal variance from the efficient estimator if the working covariance were correctly specified (if $Var(Y) = W$ so that $w^*_{im} = w_{im}$, for all i, m), and a second term from using the nonefficient weights. We let $Var(\hat{\beta}_{1\text{eff}^*})$ denote the variance of the estimated slope if the weights that were used were efficient – if $Var(Y) = W$ – and note that this variance is given by equation 7.4.

$$
\begin{aligned}
Var(\hat{\beta}_1) = &\left(\frac{1}{((\sum_m \sum_i w_{im})(\sum_m \sum_i x_i w_{im} x_m) - (\sum_m \sum_i x_i w_{im})^2)}\right)^2 \times \\
&\{(\sum_m \sum_i x_m w_{im})^2 \sum_m \sum_i (w_{im} + w^*_{im} - w_{im}) + (\sum_m \sum_i w_{im})^2 \sum_m \sum_i x_m (w_{im} + w^*_{im} - w_{im}) x_i \\
&- 2(\sum_m \sum_i w_{im})(\sum_m \sum_i x_m w_{im})(\sum_m \sum_i x_m (w_{im} + w^*_{im} - w_{im}))\} \\
= &Var(\hat{\beta}_{1\text{eff}^*}) + \left(\frac{1}{((\sum_m \sum_i w_{im})(\sum_m \sum_i x_i w_{im} x_m) - (\sum_m \sum_i x_i w_{im})^2)}\right)^2 \times \\
&\{(\sum_m \sum_i x_m w_{im})^2 \sum_m \sum_i (w^*_{im} - w_{im}) + (\sum_m \sum_i w_{im})^2 \sum_m \sum_i x_m (w^*_{im} - w_{im}) x_i \\
&- 2(\sum_m \sum_i w_{im})(\sum_m \sum_i x_m w_{im})(\sum_m \sum_i x_m (w^*_{im} - w_{im}))\}
\end{aligned}
$$

Note that the first term is not the variance of the efficient estimator for the true data being estimated (with $var(Y) = V$), but is the efficient estimate if the true data were such that $Var(Y) = W$. Nevertheless, by separating the variance of our estimate in this way, we can see that the term that may lead to non-independent increments and nonmonotonic information growth is the second term in the equation; if the second term were zero, then the first term would be that from the efficient estimator and therefore must have independent increments and grow monotonically.

7.4 Using GEE with Homoscedastic Data

This section considers the case of homoscedastic data, in which only the correlation structure can be misspecified. Focusing on this case, and assuming that the constant variance can be appropriately estimated, means that the $w^*_{im} - w_{im}$ terms above must be between -1 and 1, because the constant $\sigma^2 = v_{ii}$ can be removed from the terms of the matrices W^{-1} and $W^{-1}VW^{-1}$. We are most interested in the consequences of using GEE with working independence for reasons noted previously; using working independence will always produce a consistent estimate and will never have convergence issues. If we let $W = \sigma^2 I$, and simplify the notation by removing the σ^2 terms as noted earlier, we have $w_{im} = 1$ if $i = m$ and $w_{im} = 0$ otherwise. These changes lead to the following formula for the variance of the

estimate of $\hat{\beta}_1$ when using working independence:

$$\begin{aligned}
Var(\hat{\beta}_1) =& Var(\hat{\beta}_{1\text{eff}^*}) + \left(\frac{1}{(n^2 Var(x))}\right)^2 * \{(\sum_m x_m)^2 \sum_m \sum_i (v_{im} - w_{im}) \\
&+ (n)^2 \sum_m \sum_i x_m (v_{im} - w_{im}) x_i - 2n(\sum_m x_m)(\sum_m \sum_i x_m (v_{im} - w_{im}))\} \\
=& Var(\hat{\beta}_{1\text{eff}^*}) + \left(\frac{1}{(n^2 Var(x))}\right)^2 * \{(\sum_m x_m)^2 \sum_m \sum_{i \neq m} v_{im} \\
&+ (n)^2 \sum_m \sum_{i \neq m} x_m v_{im} x_i - 2n(\sum_m x_m)(\sum_m \sum_{i \neq m} x_m v_{im})\} \qquad (7.6)
\end{aligned}$$

Noting that the $Var(\hat{\beta}_{1\text{eff}^*})$ estimate in this case includes the terms missing from the sums (i.e. when $i = m$), we can rewrite this formula as:

$$\begin{aligned}
Var(\hat{\beta}_1) &= \left(\frac{1}{(n^2 Var(x))}\right)^2 n^2 \left(\bar{x}^2 \sum_m \sum_i v_{im} + \sum_m \sum_i x_i v_{im} x_m - 2\bar{x} \sum_m \sum_i v_{im} x_m\right) \\
&= \left(\frac{1}{(nVar(x))}\right)^2 \left(\sum_m \sum_i v_{im}(x_i - \bar{x})(x_m - \bar{x})\right) \qquad (7.7)
\end{aligned}$$

We note that if the data are truly exchangeable, then $V_{n\times n} = (1-\rho) I_{n\times n} + \rho \mathbf{1}_{n\times n}$ and equation 7.7 becomes:

$$Var(\hat{\beta}_1) = \left(\frac{1}{(nVar(x))}\right)^2 \left(\sum_i 1 * (x_i - \bar{x})(x_i - \bar{x}) + \rho \sum_m (x_m - \bar{x}) \sum_{i \neq m} (x_i - \bar{x}) * 1_{v_{im} \neq 0}\right)$$

If the design is balanced, then $\sum_i (x_i - \bar{x}) * 1_{v_{im} \neq 0} = 0$, and the above becomes:

$$\begin{aligned}
Var(\hat{\beta}_1) &= \left(\frac{1}{(nVar(x))}\right)^2 \left(\sum_i 1 * (x_i - \bar{x})(x_i - \bar{x}) - \rho \sum_i (x_i - \bar{x})(x_i - \bar{x})\right) \\
&= \frac{(1-\rho)}{nVar(x)}
\end{aligned}$$

7.4.1 Non-Independent Increments

As with any inefficient estimator, we are concerned about non-independent increments. In the case of GEE with homoscedastic data, and using an independent working covariance matrix, the degree of dependence will vary based on the timing of the analyses and on the amount and structure of the correlation.

As in section 5.3.1, we assess departures from independent increments on two metrics: the relative departures and the linear trend of the final analysis with the previous interim analyses. The impact of non-independent increments was assessed by comparing the nominal type I error rate and power using the standard sequential methods with the results from simulating trials with non-independent increments and using the boundaries developed under the assumption of independent increments. We evaluate the power at the alternative that would have 97.5% power if the independent increment assumption were true. To summarize, for each scenario we construct a design such that:

- Boundaries are constructed using the true information growth for the scenario but assuming independent increments. (So the diagonal of the covariance matrix of the interim statistics is specified correctly, but the off-diagonal elements are not.)

- At each interim analysis, the z-statistic for the interim analysis is compared to the boundaries constructed during the design phase.

- The alternative with 97.5% power for comparison was calculated from the design that assumes independent increments.

The design is evaluated at both the null and the alternative by simulating values of the interim statistics from the true covariance matrix for the interim statistic (that may not have independent increments). If independent increments were true, all of the designs would have a type I error rate of 0.025 and 97.5% power for the specified alternative.

Table 7.1 shows the departures from independent increments for both exchangeable and AR(1) true correlations when using GEE with working independence and homoscedastic data. Even somewhat large departures from independent increments on the relative scale do not lead to dramatic differences in the type I or type II error rates from the nominal levels. As before, under this scenario with an early interim analysis, the Pocock design is slightly more susceptible to departures from independent increments than the O'Brien-Fleming design due to higher stopping probabilities at the earliest analyses (greater error spent).

We also note from this table that the linear trend in departures metric with exchangeable data reflects no departures from independent increments because the final analysis has a balanced design and is thus efficient. However, the complete covariance matrix of the interim statistics does not have the complete independent increment structure and thus this metric does not adequately reflect departures from independent increments with exchangeable data.

Practical Considerations

For future studies, we investigate when non-independent increments may be of practical concern. To explore the effect of the timing of the interim analyses and the accrual pattern on possible non-independent increments, we consider a scenario in which every individual will have four measurements made (at equally spaced study times 0-3). We assume that an interim analysis will be conducted before a fourth measurement is obtained on any individual (at time t_j which is fixed), and we will calculate the covariance of the statistic from this analysis with the statistic from an interim analysis after we have the fourth measurement on a fraction of individuals (at time t_k, which will vary). To assess departures from independent increments in this setting, we used a modified relative departures approach and use $log(\frac{Cov(\hat{\beta}_{1j},\hat{\beta}_{1k})}{Var(\hat{\beta}_{1k})}$, so positive values indicate greater than expected covariance and negative values indicate less than expected covariance. We consider three different accrual patterns: (a) fast, such that all individuals had a third measurement before anyone had a fourth, (b) medium, such that two-thirds of the individuals had a third measurement before the first individual had the fourth, and (c) slow, such that only one-third of the individuals had a third measurement before the first individual had a fourth. Table 7.2 shows the fraction of individuals with measurements at each study time at the fixed analysis t_j and at two possible interim analysis times.

Figures 7.4 and 7.5 show how the amount of covariance between the statistics at two interim analyses varies as a function of the correlation of the data, timing of the interim analyses, and the accrual pattern. In these plots, very slight departures are observed in all cases when the correlation between measurements is less than 0.6. When the correlation is higher than 0.6, fast accrual (all individuals have a third measurement before anyone

Table 7.1: Empirical type I error rate and power at the alternative which has 97.5% under an independent increment structure, when using an independence working covariance matrix with GEE and homoscedastic data. The relative and linear departures from independent increments are as in equations 5.7 and 5.8, respectively.

Scenario 1							
		Relative	Linear	OBF		Pocock	
True Correlation		Ind. Inc.	Ind. Inc.	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
Exchangeable	$\rho = 0.3$	0.5903	0.0000	0.0250	0.9750	0.0249	0.9748
	$\rho = 0.6$	1.5971	0.0000	0.0250	0.9751	0.0245	0.9753
	$\rho = 0.9$	3.7219	0.0000	0.0251	0.9749	0.0232	0.9766
AR(1)	$\rho = 0.3$	0.9541	0.1553	0.0249	0.9749	0.0250	0.9748
	$\rho = 0.6$	1.0567	0.1781	0.0250	0.9750	0.0249	0.9750
	$\rho = 0.9$	2.0365	0.1290	0.0249	0.9750	0.0240	0.9758
Scenario 2							
Exchangeable	$\rho = 0.3$	0.2090	0.0000	0.0250	0.9749	0.0248	0.9751
	$\rho = 0.6$	0.4979	0.0000	0.0249	0.9750	0.0240	0.9758
	$\rho = 0.9$	0.9986	0.0000	0.0250	0.9749	0.0229	0.9769
AR(1)	$\rho = 0.3$	0.1890	0.0214	0.0250	0.9750	0.0251	0.9749
	$\rho = 0.6$	0.3009	0.0248	0.0249	0.9750	0.0247	0.9752
	$\rho = 0.9$	0.6385	0.0180	0.0250	0.9749	0.0235	0.9763

Table 7.2: Distribution of observed study times at possible interim analysis times under fast, medium, and slow accrual. The proportion of the final amount at each study time is given.

Fraction with 4th	Study time=0	1	2	3
Fast Accrual (all have third before anyone has fourth)				
0 (t_j)	1	1	1	0
0.33	1	1	1	0.33
0.66	1	1	1	0.66
Medium Accrual (2/3 have third before anyone has fourth)				
0 (t_j)	1	1	0.66	0
0.33	1	1	1	0.33
0.66	1	1	1	0.66
Slow Accrual (1/3 have third before anyone has fourth)				
0 (t_j)	1	0.66	0.33	0
0.33	1	1	0.66	0.33
0.66	1	1	1	0.66

gets a fourth) and interim analyses scheduled in places where there is not balance can lead to a large amount of deviation from the independent increment structure, particularly if the data are truly exchangeable (figure 7.4). If the data are truly AR(1), the departures from independent increments are not as extreme, even with very high correlation between measurements on the same individual (figure 7.5).

7.4.2 Nonmonotonic Information Growth

As is the case with heteroscedastic, independent data, it is possible to have nonmonotonic information growth when using OLS (or estimates that match OLS, such as GEE with an independent working covariance matrix).

For analysis times t_j and t_k with $j < k$, we use equation 7.7 and get:

$$\frac{Var(\hat{\beta}_{1k})}{Var(\hat{\beta}_{1j})} = \frac{\left(\frac{1}{(n_k Var_k(x))}\right)^2 \left(\sum_m^{n_k} \sum_i^{n_k} v_{im}(x_i - \bar{x}_k)(x_m - \bar{x}_k)\right)}{\left(\frac{1}{(n_j Var_j(x))}\right)^2 \left(\sum_m^{n_j} \sum_i^{n_j} v_{im}(x_i - \bar{x}_j)(x_m - \bar{x}_j)\right)}$$

$$= \left(\frac{n_j Var_j(x)}{n_k Var_k(x)}\right)^2 \left(\frac{\sum_m^{n_k} \sum_i^{n_k} v_{im}(x_i - \bar{x}_k)(x_m - \bar{x}_k)}{\sum_m^{n_j} \sum_i^{n_j} v_{im}(x_i - \bar{x}_j)(x_m - \bar{x}_j)}\right) \tag{7.8}$$

From this equation, it is clear that the first term involving n_j, n_k, and the variances of the predictor space at each analysis time must be less than 1. Therefore the information will be increasing if the correlation terms were all 0. However, there is no such guarantee with the term on the right side. Adding measurements that disrupt the balance noted previously with exchangeable data, for instance, could lead the top term to be larger than the bottom and thus could result in nonmonotonic information growth if the correlation were sufficient to do so.

Intuition might suggest that the reason for this possible absolute inefficiency is due to the weighting on the measurements when determining the estimate of $\hat{\beta}$. For example, we noted previously that if an independence working covariance matrix is used, the estimate of $\hat{\beta}_1$ will match exactly the estimate that would have been obtained through OLS regression ignoring the correlation within an individual. When all individuals have the same number of

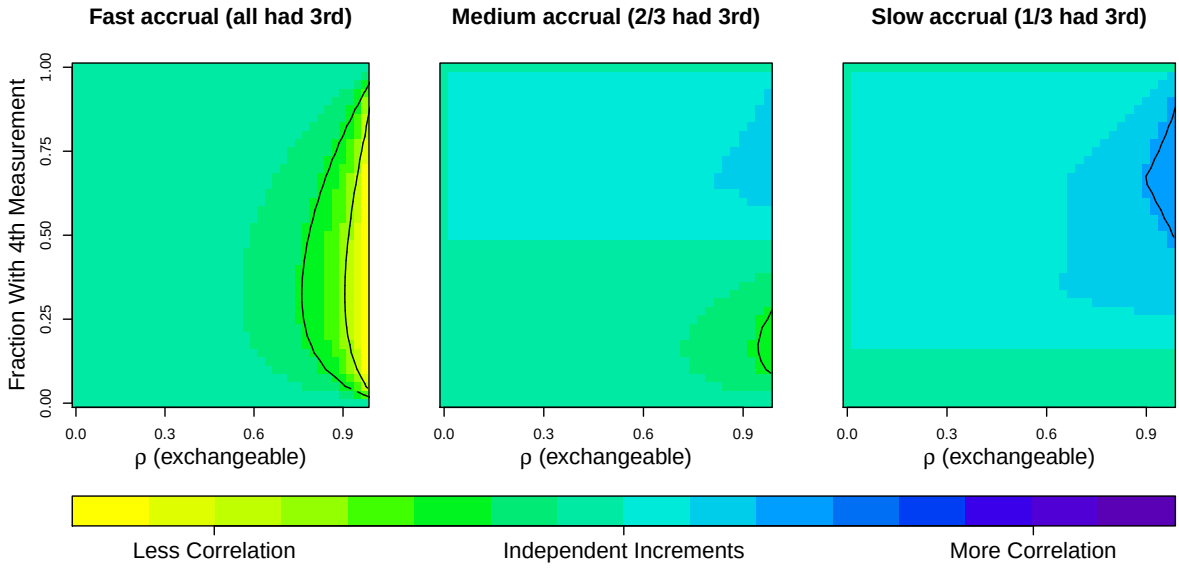

Figure 7.4: The relative amount of non-independent increments when the data are truly exchangeable. These plots consider the correlation between one interim analysis before any individuals have a 4th (final) measurement and one interim analysis at points later in calendar time when various fractions of 4th measurements have been obtained.

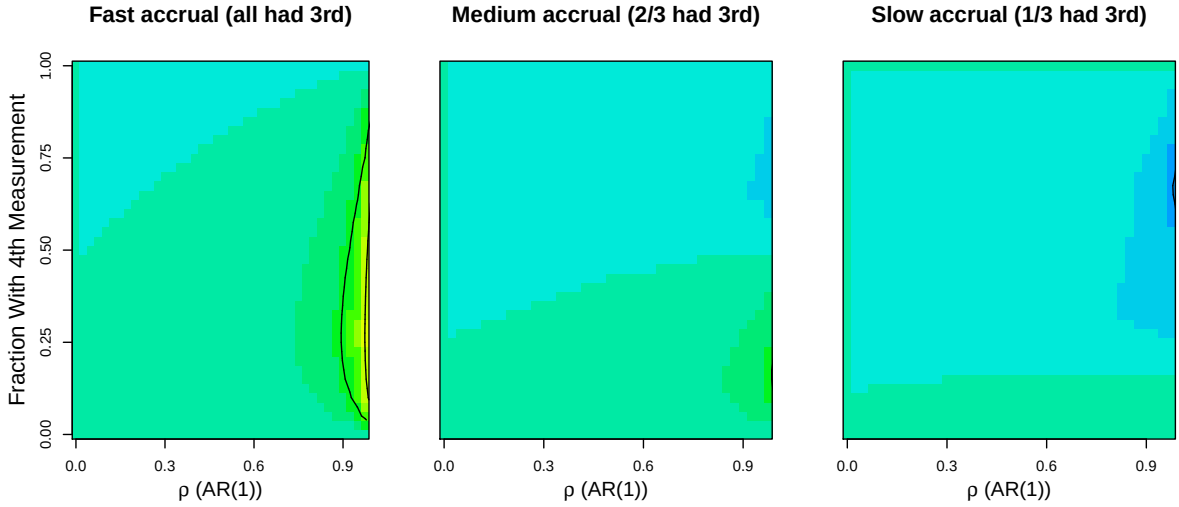

Figure 7.5: The relative amount of non-independent increments when the data are truly AR(1), using the same setting as in figure 7.4.

measurements at the same time points from randomization, all subjects are weighted equally and this balance does not generally result in a great loss of efficiency; when such balance is present and the data are exchangeable, it is equal to the efficient estimator. However, the independence working covariance weighting is not efficient when a few individuals have more measurements than the others. Compared to the case of all independent measurements, the line that is fit with just two observations on each subject when those observations are highly correlated is much less variable (there is a gain in information due to the positive correlation within an individual). If only a handful of these highly correlated subjects have measurements at a more extreme time point, then these subjects have greater influence on the slope (as if they were new, independent measurements), and the variability of the slope increases due to chance selection of different measurements over hypothetical repeated experiments. The potential influence of these new data points can actually increase the true variability of the slope, unless the correlation with other measurements is properly accounted for by downweighting the additional observations (using the working covariance matrix) relative to the weights that would be used in OLS regression.

We first explore an extreme scenario that could lead to nonmonotonic information growth. For this scenario, the true effect is linear, the data are homoscedastic, and the correlation within individuals is high. For our particular example, 10 measurements are made on each individual, one at baseline, and one at each of nine follow-up times. The accrual occurs over the first two months. For an example of high within individual correlation, we chose an AR(1) structure with $\rho = 0.95$. In an attempt to make the exchangeable correlation structure as equivalent as possible, ρ for the exchangeable case was chosen such that the average correlation between all pairs of measurements on an individual at the end of the study would be equal to that in the AR(1) structure ($\rho = 0.8338$). Finally, to ensure comparability, the number of individuals in the AR(1) case was increased such that the final amount of statistical information was equivalent between the AR(1) and exchangeable cases (2170 and 500 individuals, respectively). Four working covariance matrices were used in each simulation: independence, exchangeable, AR(1), and unstructured. These simulations were done using the `geepack` package in R (Yan and Fine, 2004).

We choose to include the "unstructured" working covariance matrix as well, because

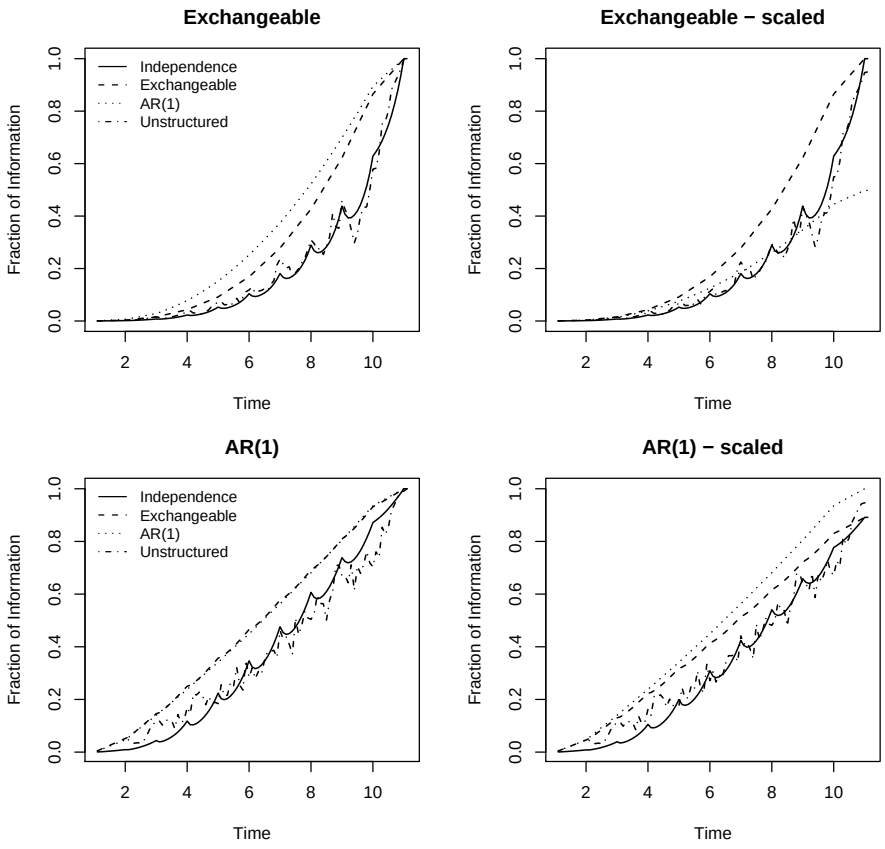

Figure 7.6: Plot illustrating information growth over time using GEE when the data are truly linear. The true correlation structure is either exchangeable or AR(1) and the plots show the information growth using each of four working covariance matrices. The scaled graphs show the true information growth relative to the amount of information when the working covariance matrix is exactly specified.

some authors (Gange and DeMets, 1996) have suggested the use of an "unstructured" working covariance matrix with GEE to provide nearly efficient estimation without prespecifying the form of the working covariance. Others have suggested that using a working covariance matrix with consistently estimated parameters (even if misspecified) will lead to nearly independent increments (Lee et al., 1996), suggesting that any choice of working covariance will lead to monotonic information growth.

Figure 7.6 shows the true information growth curves in our example. The plots demonstrate the nonmonotonic information growth when using an independence working covariance matrix in this setting. The scaled plots show the relative loss of efficiency compared to using the correctly specified form of the working covariance matrix. In this example, using an exchangeable working covariance matrix appears to be most desirable; it does lose some efficiency relative to using AR(1) when the truth is AR(1) (relative efficiency = 89%), but does not become nonmonotonic. In contrast, when the true data are exchangeable, using a working AR(1) structure leads to a dramatic drop in efficiency (relative efficiency = 50%).

Using an "unstructured" working covariance matrix does lead to nearly efficient estimation when the design is balanced (when all subjects have equal numbers of measurements), however, it performs poorly when the design is markedly unbalanced, yielding results similar to those using an independence working covariance matrix (figure 7.6). When the design is unbalanced, some estimated parameters in the working covariance structure are quite variable because there are only a few observations contributing to those estimates. For this reason, using the unstructured working covariance can lead to many of the same problems as using the independence working covariance, which suggests that in this circumstance the use of the exchangeable working covariance matrix would be preferred. In addition, in a small number of simulations (approximately 3%) using an unstructured working correlation matrix meant that the GEE estimates did not converge. These cases were excluded from our estimates, and hence the graphs do not reflect another practical difficulty of using an unstructured working covariance matrix in the setting of a group sequential clinical trial.

As might be expected, the degree of correlation within measurements on the same individual affects the true information growth when using an independence working covariance matrix (figure 7.7). When the true data are exchangeable with low correlation ($\rho = 0.3$)

and the same study design as before (2 month accrual, 10 measurements per individual), the information growth is nearly the same between the exchangeable and independence working covariance matrices (figure 7.7A). As the correlation increases, using working independence becomes less efficient at interim points in the trial and can lead to nonmonotonic information growth (figure 7.7: A-C).

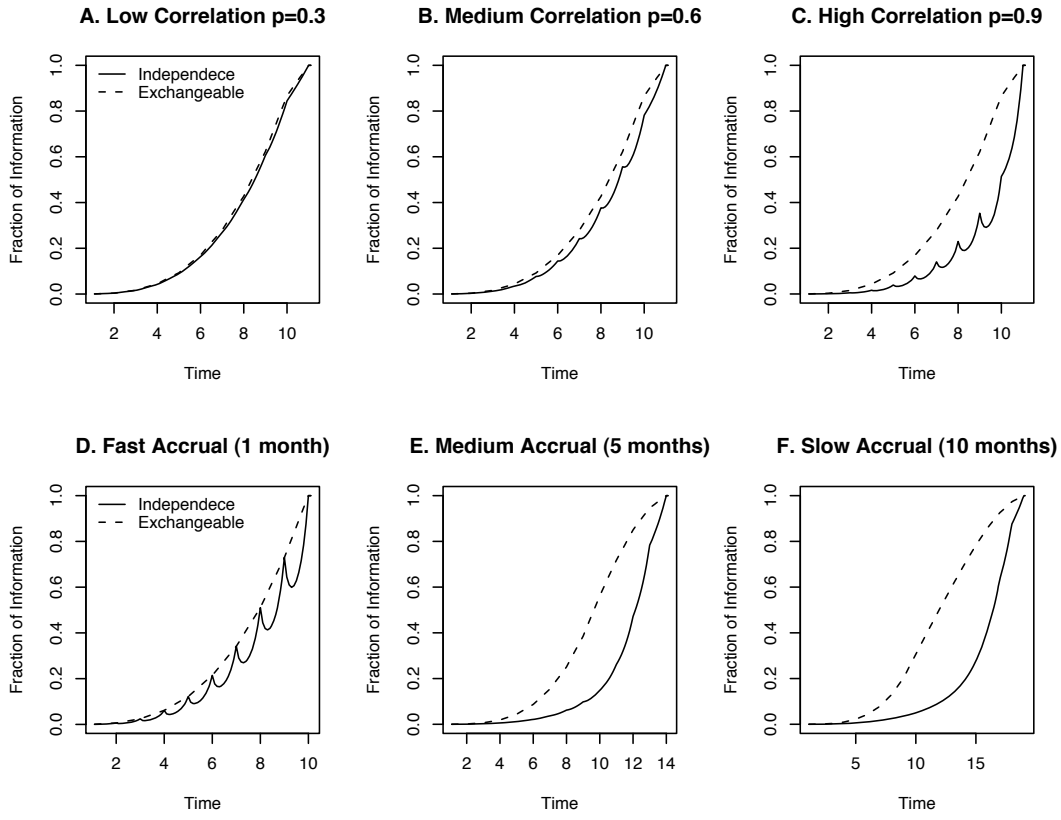

Figure 7.7: Plots illustrating the effect of the within individual correlation and the accrual pattern on the information growth over time using GEE. In all cases, the data are truly linear the covariance within individuals has an exchangeable structure, and 10 measurements are made on each individual (at baseline and months 1-9). For plots A-C, accrual was fixed at 2 months and for plots D-F the correlation was fixed at $\rho = 0.8338$.

When the design is completely balanced (as might occur at the end of a study with no dropouts), the estimates using independence and exchangeable working covariance matrices are the same. Such balance may also be achieved during a study if the accrual period is

shorter than the time between consecutive measurements on an individual. In our example where individuals are measured every month, this circumstance would occur if everyone were accrued within one month (e.g., every individual has a first measurement before anyone has a second as in figure 7.7D). However, when the design is far from balanced (as might occur during a long accrual period), working independence will be noticeably less efficient than exchangeable at interim points in the study when the true data are exchangeable (e.g. figure 7.7F).

A long accrual period when using an independence working covariance matrix leads to relative inefficiency, but does not tend to lead to noticeable nonmonotonic information growth. Nonmonotonicity is most pronounced when the accrual period is short relative to the follow up on each individual and if the correlation within an individual is high (figure 7.7D). Consider a case of high within subject correlation ($\rho = 0.8338$) and short accrual (so that all individuals have two measurements before anyone has a third). In this situation, the amount of statistical information decreases when the first individual gets at third measurement, and continues to decrease until slightly more than 10% of the study population has a third measurement. The amount of information present when everyone had two measurements but no one had a third is not surpassed until more than 50% of the new third measurements are obtained. This result becomes even more striking as the study continues. When everyone has nine measurements but no one yet has ten, the amount of statistical information decreases when the first person gets a tenth measurement and continues to decrease until approximately 30% have a tenth measurement. The amount of information is not greater than the amount when no one had a tenth until 70% have a tenth measurement.

Practical Considerations

As was the case with non-independent increments, we are concerned about when nonmonotonic information growth might occur in practice. We noted above the potential effects of the magnitude of the correlation within individuals and the timing of the analyses relative to the accrual pattern. To explore these effects more fully, we again consider a scenario in

which every individual will have four measurements made (at equally spaced study times 0-3), as we did when exploring possible non-independent increments. We assume that an interim analysis will be conducted before a fourth measurement is obtained on any individual (at time t_j which is fixed), and calculate the relative information of the statistic from this analysis with the statistic from an interim analysis after we have the fourth measurement on a fraction of individuals (at time t_k, which we will vary). We transform this ratio to a log scale for ease of illustration and plot ($log(\frac{Var(\hat{\beta}_{1j})}{Var(\hat{\beta}_{1k})})$, so negative values indicate nonmonotonic "information." As with our exploration of non-independent increments, we consider three different accrual patterns as seen in table 7.2.

Figures 7.8 and 7.9 show places in which an interim analysis occurring later in calendar time leads to a more variable estimate of the slope. Such nonmonotonic information occurs when the correlation is high (> 0.6), when accrual is fast, and when interim analyses are spaced close together.

7.5 GEE with Heteroscedasticity

We now allow for the possibility of heteroscedasticity among the observed data as well, either through a predictor-variance or a mean-variance relationship. In this case, the form of the variance does not simplify as it did with homoscedastic data. Using an independent working covariance matrix with heteroscedastic, correlated data, can lead to non independent increments and nonmonotonic information growth due to both the correlated data (as described above) and the heteroscedasticity (as described in chap 5). We are then particularly interested in the interaction between the two, namely how heteroscedasticity affects the potential non independent increments and nonmonotonic information growth due to using an independent working covariance matrix with correlation data.

7.5.1 Not accounting for heteroscedasticity

Table 7.3 shows the relative and linear departures from independent increments for data with a strong predictor-variance relationship and various degrees of correlation.

Interestingly, when not accounting for either the correlation or the heteroscedasticity, the magnitude of departure from independent increments is less than that for the equiv-

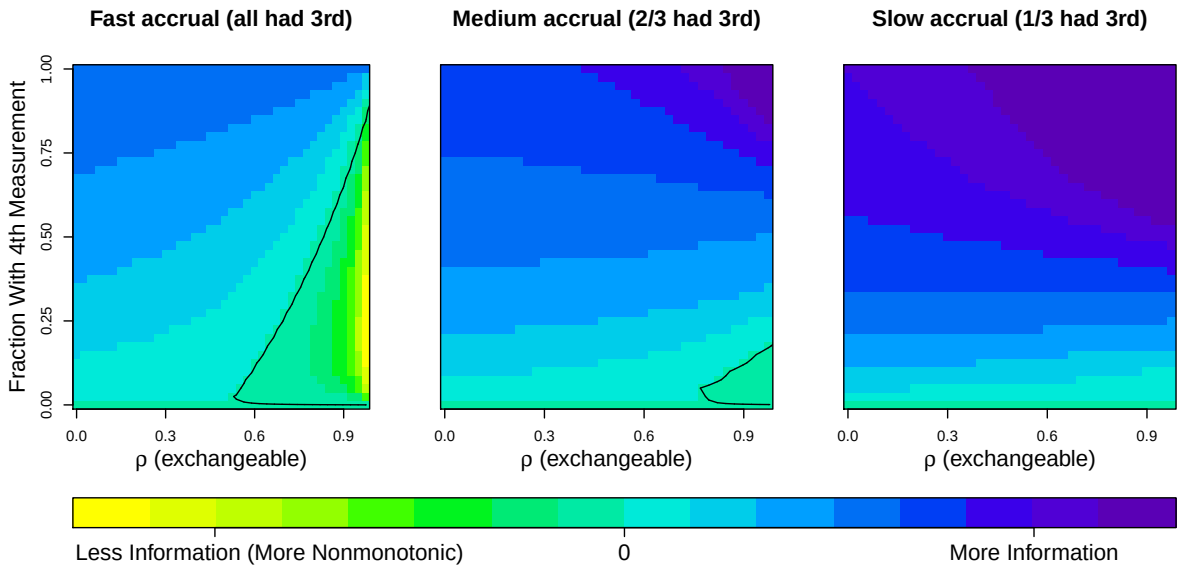

Figure 7.8: Places of possible nonmonotonic "information" when the data are truly exchangeable. These plots consider the relative amount of information between one interim analysis before any individuals have a 4th (final) measurement and one interim analysis at points later in calendar time when various fractions of 4th measurements have been obtained.

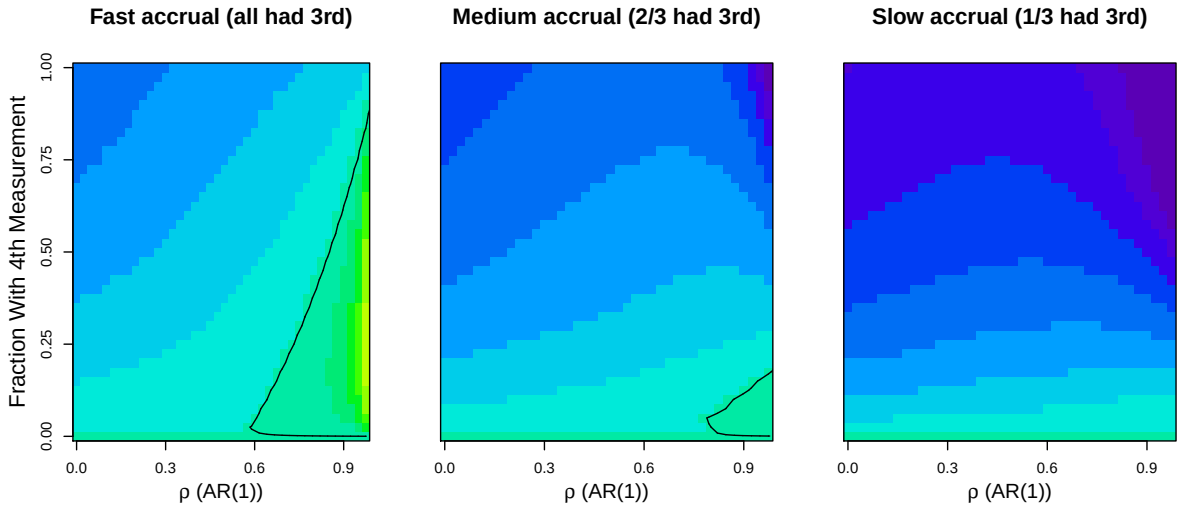

Figure 7.9: Places of possible nonmonotonic "information" when the data are truly AR(1), using the same setting as in figure 7.8.

alent amount of heteroscedasticity with independent data. For example, with a predictor variance relationship as in equation 5.1 and $a = 10$, $b = 5$, and $\gamma = 1$, we saw in table 5.1 that the relative departure from independent increments was 3.19 under scenario 1. For the equivalent parameters, but exchangeable data, table 7.3 shows that all three possible values have less relative departures from independent increments, and smaller differences from the nominal type I error rate as well. The same is true for data with an AR(1) correlation structure as well. The combination of inefficiencies appears to reduce the departures from independent increments and thus lessen one potential problem of using an inefficient estimator in this setting. An intuitive explanation for this effect is that the correlation between measurements within individuals will often make OLS estimates at interim analyses more correlated than they would be if independent increments were true, although this need not always be true. Heteroscedasticity such that later measurements are more variable will make the correlation between statistics at interim analyses less than it would be if there were independent increments. Thus, there are two competing factors affecting the amount of departure from an independent increment structure, which can lead to smaller departures than would be true under either correlated or heteroscedastic data alone.

7.5.2 Accounting for heteroscedasticity but not correlation

This paradox of the increasing variability of the measurements decreasing the problems due to non-independent increments continues even when properly weighting for the heteroscedsaticity. For this scenario, we allow for optimal weighting for the heteroscedasticity, but using working independence for the correlation structure. Such a model is similar to a model that assumes a particular mean-variance relationship (although not identical).

Table 7.4 shows the departures from independent increments under the same scenarios as in table 7.3. However, now that the heteroscedasticity is weighted appropriately, we again see the pattern of increasing departures from independent increments with increasing correlation. However, the amount of departure from independent increments is still measured to be less than in the totally independent data case.

Table 7.3: Empirical type I error and power for the alternative calculated to have 97.5% under an independent increment structure, when using an independence working covariance matrix with GEE and heteroscedastic data. The relative and linear departures from independent increments are as in equations 5.7 and 5.8, respectively.

Scenario 1: a=10, b=5, $\gamma = 1$							
		Relative	Linear	OBF		Pocock	
True Correlation		Ind. Inc.	Ind. Inc.	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
Exchangeable	$\rho = 0.3$	2.1541	0.2652	0.0247	0.9755	0.0250	0.9749
	$\rho = 0.6$	1.5150	0.1976	0.0250	0.9751	0.0249	0.9750
	$\rho = 0.9$	2.1388	0.0293	0.0250	0.9750	0.0237	0.9763
AR(1)	$\rho = 0.3$	3.1474	0.3546	0.0246	0.9755	0.0246	0.9752
	$\rho = 0.6$	2.3405	0.3187	0.0248	0.9752	0.0249	0.9751
	$\rho = 0.9$	1.5835	0.1851	0.0250	0.9750	0.0243	0.9755
Scenario 1: a=10, b=5, $\gamma = 2$							
Exchangeable	$\rho = 0.3$	3.1481	0.3012	0.0235	0.9765	0.0242	0.9757
	$\rho = 0.6$	1.4740	0.1916	0.0246	0.9754	0.0251	0.9749
	$\rho = 0.9$	1.0350	0.0540	0.0250	0.9751	0.0244	0.9756
AR(1)	$\rho = 0.3$	4.1844	0.3997	0.0227	0.9772	0.0230	0.9769
	$\rho = 0.6$	3.0843	0.3421	0.0237	0.9763	0.0241	0.9758
	$\rho = 0.9$	0.9795	0.1713	0.0248	0.9752	0.0249	0.9751

Table 7.4: Empirical type I error and power for the alternative calculated to have 97.5% under an independent increment structure, when using an independence working covariance matrix with GEE and heteroscedastic data, but accounting for the heteroscedastic data. The relative and linear departures from independent increments are as in equations 5.7 and 5.8, respectively.

Scenario 1: a=10, b=5, $\gamma = 1$							
		Relative	Linear	OBF		Pocock	
True Correlation		Ind. Inc.	Ind. Inc.	SP_{null}	SP_{alt}	SP_{null}	SP_{alt}
Exchangeable	$\rho = 0.3$	0.5726	-0.0093	0.0250	0.9749	0.0249	0.9749
	$\rho = 0.6$	1.4537	-0.0258	0.0250	0.9750	0.0244	0.9753
	$\rho = 0.9$	2.9966	-0.0624	0.0249	0.9750	0.0233	0.9765
AR(1)	$\rho = 0.3$	0.9550	0.1519	0.0250	0.9750	0.0251	0.9748
	$\rho = 0.6$	1.0602	0.1828	0.0249	0.9749	0.0250	0.9750
	$\rho = 0.9$	1.6250	0.1090	0.0250	0.9749	0.0241	0.9757
Scenario 1: a=10, b=5, $\gamma = 2$							
Exchangeable	$\rho = 0.3$	0.3809	0.0000	0.0250	0.9749	0.0250	0.9749
	$\rho = 0.6$	0.8251	0.0000	0.0250	0.9749	0.0247	0.9753
	$\rho = 0.9$	1.3577	0.0000	0.0250	0.9750	0.0240	0.9761
AR(1)	$\rho = 0.3$	0.8618	0.1312	0.0250	0.9749	0.0252	0.9748
	$\rho = 0.6$	0.9507	0.1560	0.0249	0.9749	0.0251	0.9750
	$\rho = 0.9$	1.0453	0.0809	0.0249	0.9750	0.0245	0.9756

7.5.3 Practical Consideration

As in the case of no heteroscedasticity, we are concerned about when non-independent increments and possible nonmonotonic information growth may occur in practice when using OLS. As before, we consider a scenario in which every individual will have four measurements made (at equally spaced study times 0-3), and we assume that an interim analysis will be conducted before a fourth measurement is obtained on any individual (at time t_j which is fixed). We are interested in the possibility of non-independent increments and nonmontonic information growth between this analysis and another interim analysis (at time t_k) which we vary based on the fractional number of individuals with a fourth measurement. We consider three different accrual patterns as seen in table 7.2.

We assess independent increments by plotting $log(\frac{Cov(\hat{\beta}_{1j},\hat{\beta}_{1k})}{Var(\hat{\beta}_{1k})})$, so that positive numbers indicated greater correlation between the interim statistics than would be true if independent increments were present. For possible nonmonotonic information growth, we plot $log(\frac{Var(\hat{\beta}_{1j})}{Var(\hat{\beta}_{1k})})$, so negative values indicate nonmonotonic "information."'

To assess these concerns with heteroscedastic, correlated data, we deliberately chose a great deal of heteroscedasticity to see what may be possible in extreme cases. We investigate the cases in which $a = 30$, $b = 5$, and $\gamma = 1$ or $\gamma = 2$ as in the equation $Var(Y|x) = \sigma^2(a + bx)^\gamma$.

Figures 7.10-7.13 show places of possible large deviations from independent increments. In general, heteroscedasticity with correlated data does not lead to larger deviations from independent increments than is true with homoscedastic data. With heteroscedasticity up to the case of $\gamma = 2$, non-independent increments is of minimal concern, unless the correlation within individuals is extremely high ($\rho > 0.6$) and the accrual is short relative to follow-up. If the data are truly AR(1), non-independent increments are even less of a concern than if the data are truly exchangeable.

Figures 7.14-7.17 show places of possible nonmonotonic information growth in this setting. With medium to slow accrual relative to follow-up, there is little problem with nonmonotonicity, as it occurs only with high correlations (ρ at least > 0.6) and occurs only when a small fraction of individuals have the fourth measurement, so that interim analyses

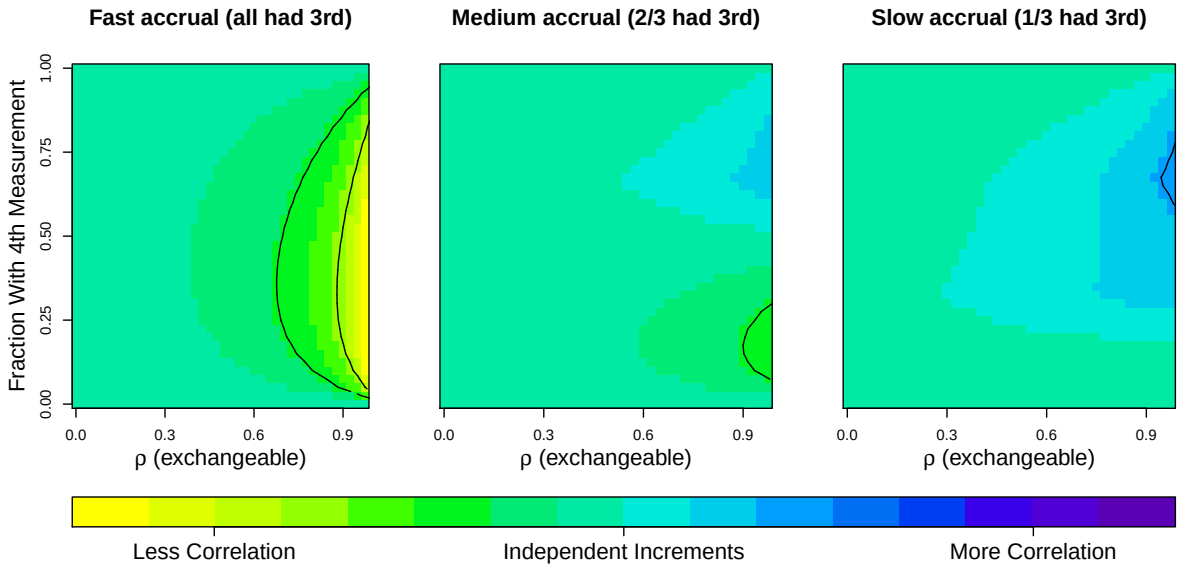

Figure 7.10: The relative amount of non-independent increments using OLS when the data are truly exchangeable and there is some heteroscedasticity ($\gamma = 1$). These plots consider the correlation between one interim analysis before any individuals have a 4th (final) measurement and one interim analysis at points later in calendar time when various fractions of 4th measurements have been obtained.

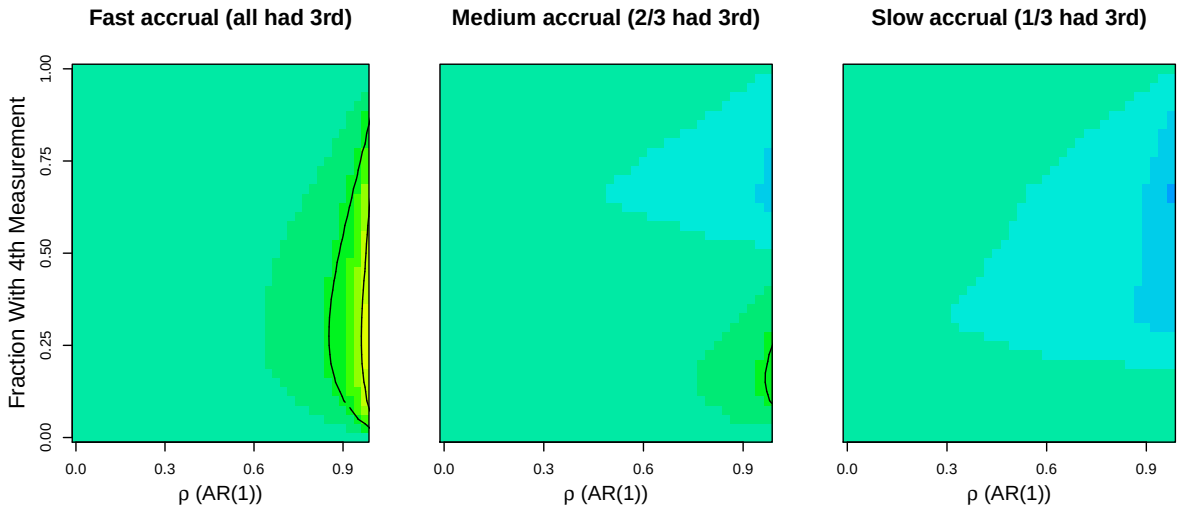

Figure 7.11: The relative amount of non-independent increments when the data are truly AR(1) and there is some heteroscedasticity ($\gamma = 1$), using the same setting as in figure 7.10.

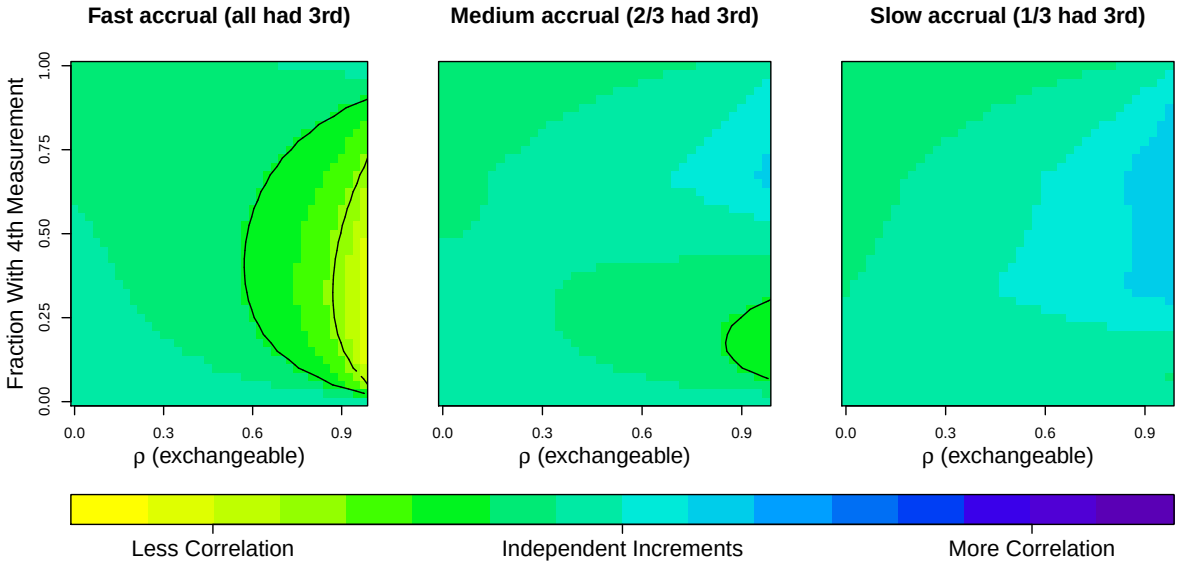

Figure 7.12: The relative amount of non-independent increments using OLS when the data are truly exchangeable and there is greater heteroscedasticity ($\gamma = 2$). These plots consider the correlation between one interim analysis before any individuals have a 4th (final) measurement and one interim analysis at points later in calendar time when various fractions of 4th measurements have been obtained.

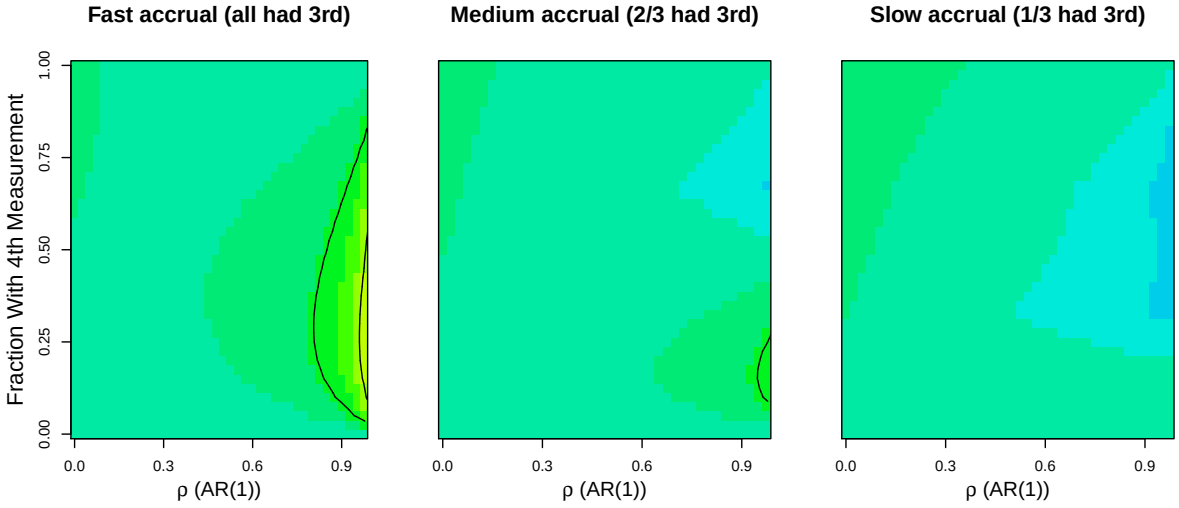

Figure 7.13: The relative amount of non-independent increments when the data are truly AR(1) and there is greater heteroscedasticity ($\gamma = 2$), using the same setting as in figure 7.12.

would have to be spaced very close together for nonmonotonicity to be of concern. However, if accrual is fast relative to follow-up (such that nearly all individuals are accrued before a second measurement is made on any individual), then nonmonotonic information is of more concern. With low to moderate correlations ($\rho < 0.6$), nonmonotonic information is possible with heteroscedasticity, but only when an interim analysis occurs shortly after one that was at a point of balance. As long as such a situation is avoided when planning a trial, nonmontonic information growth is unlikely with low to moderate correlations and heteroscedastic data. However, if the correlation is high ($\rho > 0.6$), care should be taken to schedule analyses such that the balance is similar between the interim analysis times. If one interim analysis is balanced, but the next is not, nonmonotonic information growth can occur.

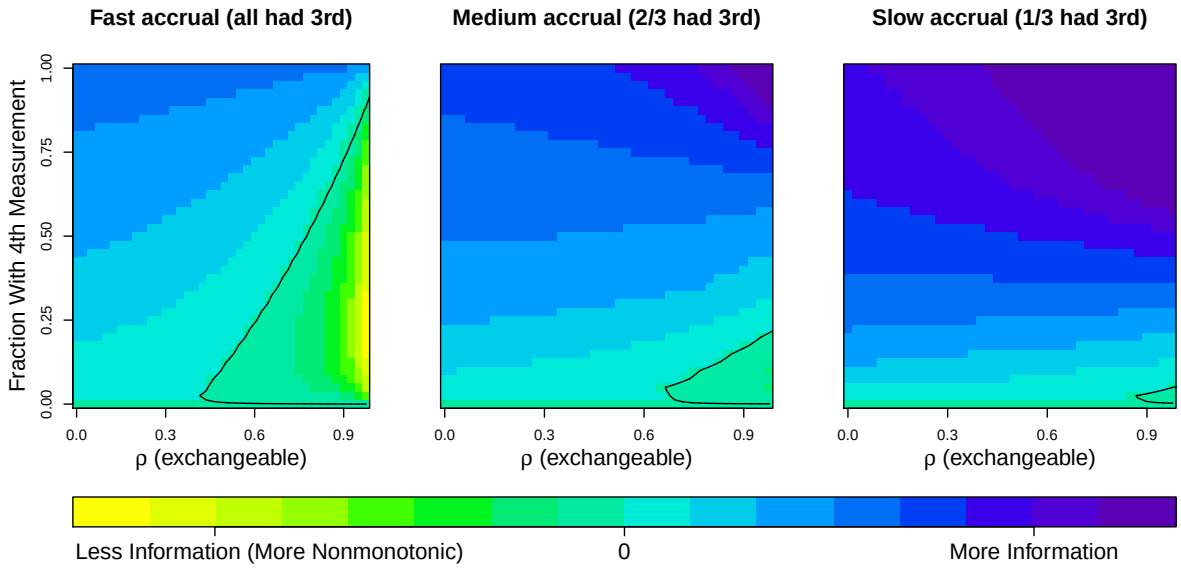

Figure 7.14: Places of possible nonmonotonic "information" using OLS when the data are truly exchangeable and there is some heteroscedasticity ($\gamma = 1$). These plots consider the relative amount of information between one interim analysis before any individuals have a 4th (final) measurement and one interim analysis at points later in calendar time when various fractions of 4th measurements have been obtained.

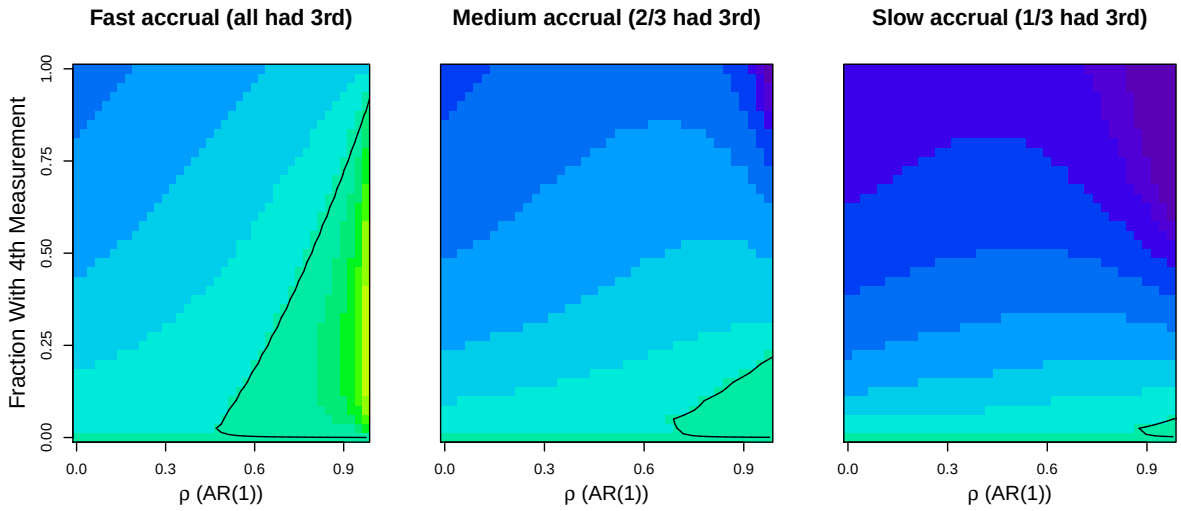

Figure 7.15: Places of possible nonmonotonic "information" when the data are truly AR(1) and there is some heteroscedasticity ($\gamma = 1$), using the same setting as in figure 7.14.

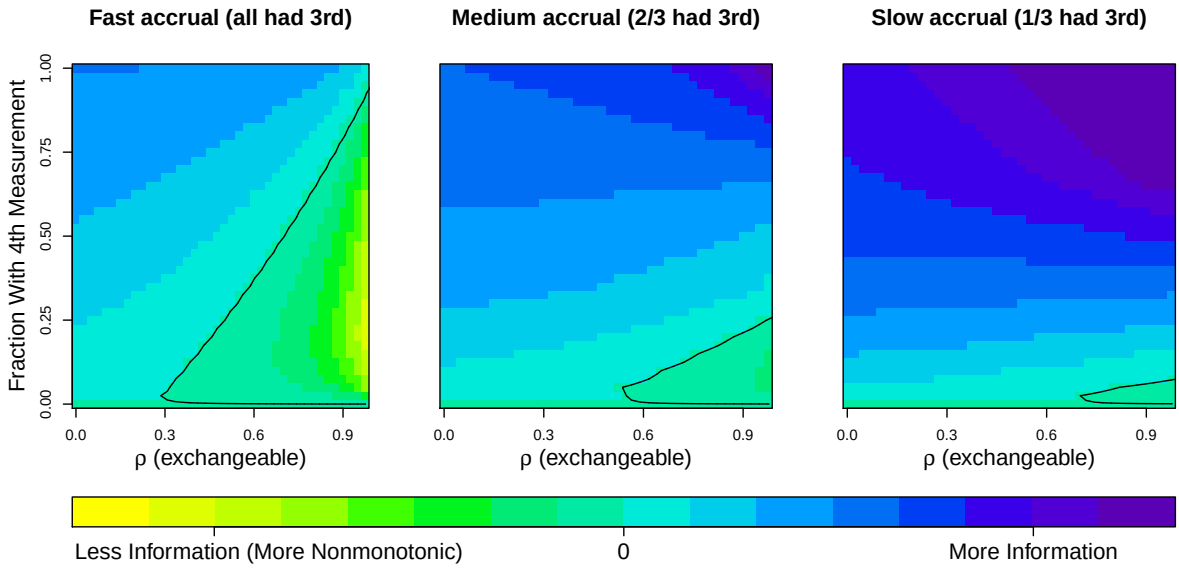

Figure 7.16: Places of possible nonmonotonic "information" using OLS when the data are truly exchangeable and there is some heteroscedasticity ($\gamma = 2$). These plots consider the relative amount of information between one interim analysis before any individuals have a 4th (final) measurement and one interim analysis at points later in calendar time when various fractions of 4th measurements have been obtained.

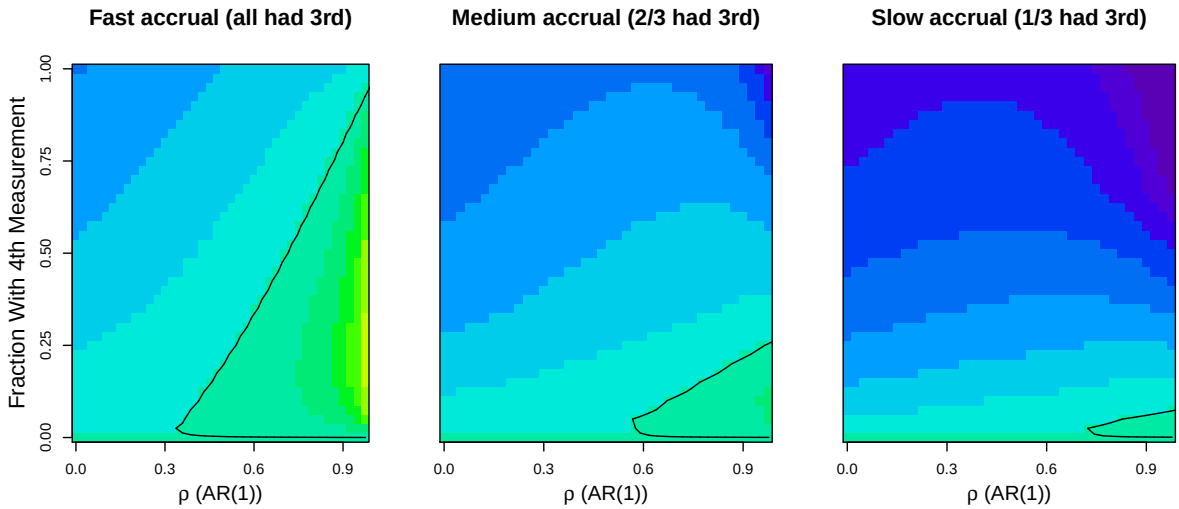

Figure 7.17: Places of possible nonmonotonic "information" when the data are truly AR(1) and there is some heteroscedasticity ($\gamma = 2$), using the same setting as in figure 7.16.

Chapter 8

EVALUATION OF RECOMMENDATIONS

To conclude, we present some recommendations for future studies and then evaluate how these recommendations would behave in an example study. These recommendations would be applicable to studies of the type investigated in this dissertation; namely studies of a linear rate of change over time with possible predictor-variance or mean-variance heteroscedasticity and possible correlations between measurements on the same individual. The methods are applicable to using GEE in this setting.

8.1 Recommendations

8.1.1 Design Stage

In most circumstances, studies can be planned to use working independence with GEE. Although this approach can result in some loss of efficiency (particularly at interim analyses where the design is not balanced) advantages of this approach are that the linear contrast will always be consistently estimated and that convergence is guaranteed. Both advantages are particularly important in clinical trials in which analyses must be fully prespecified and we are concerned about the implications of our statistical approach when the model is not completely correct and when non-convergence of a model would be particularly problematic.

At the design stage of a trial, the information growth should be estimated under the null with reasonable assumptions about any possible heteroscedasticity and correlations between measurements on the same individual. Assuming that the study observation times are fixed by scientific concerns, it will also be necessary to make a reasonable estimate of the accrual pattern. Unlike in trials with a single outcome measurement on each individual, the information growth in longitudinal trials depends on the study observation times and the accrual pattern.

For power calculations in the presence of an assumed mean-variance relationship, we

recommend updating the final information that would be true at the end of the trial based on the mean-variance relationship. Such an adjustment for a positive mean-variance relationship will adjust the power downward for a positive hypothesis or upward for a negative hypothesis.

Study designers can consider modifying the futility boundary based on the anticipated information growth for the alternative with estimated 97.5% power after adjustment for the mean-final information. In the presence of a greater alternative with a positive mean-variance relationship, this will generally bring the actual power closer to the nominal 97.5% level and will improve efficiency (lower ASN) under the null. However, this step is not necessary and may be skipped if the stopping probabilites of the design under the null are to be fully maintained.

Nonmonotonic information growth is very unlikely in most reasonable circumstances. If accrual is very short compared to the duration of follow up and either correlation is expected to be very high or extreme heteroscedasticity is expected, then nonmonotonic information growth maybe possible if interim analyses are spaced very close in calendar time. In these extreme situations, interim analyses should be planned to take place at points in the study where balance is anticipated (and avoid scheduling interim analyses where only a few individuals have measurements at the most extreme study times). Further, in such extreme situations, one should strongly consider using alternative working covariance matrices such as an exchangeable working covariance matrix. Based on our simulations, even extreme patterns of fast accrual relative to follow up (e.g. everyone has two measurements before anyone has a third) do not produce sustained nonomontonicity with correlations up to $\rho = 0.6$ or heteroscedasticity up to $\gamma = 2$.

8.1.2 Conducting the Trial

We recommend using boundaries on the z-statistic scale when conducting interim analyses for reasons illustrated in chapter 4. Boundaries at the final analysis should be constructed using a constrained boundary approach that adjusts for the true information growth observed over the entire study. This approach is easily implemented and will maintain close

to the nominal type I error rate except in extreme cases.

If the timing of interim analyses varies dramatically from what was expected at the design stage, the boundaries may be modified as per standard techniques using the expected information growth under the null (or the alternative for the futility boundary if a modified futility boundary is being used).

8.1.3 Post-Trial Inference

At the conclusion of the trial, we recommend calculating the sampling density based on the observed variance of the test statistic and either (a) the true information growth if the study continued until the final analysis or (b) the assumed information growth under the null if the study terminated early. The sampling density can then be used for confidence intervals and for adjusted estimates of the treatment effect.

8.2 Case Study

As an example, consider a one-arm trial to detect a decrease in the rate of cognitive decline. For the sake of this example, we will assume the measure of cognitive functioning is done with an instrument such that:

- Positive numbers indicating worse cognitive function
- The variability of the test increases with higher scores, such that they fit our model for a mean-variance relationship with $\gamma = 2$.
- The time between measurements is sufficient and the measurement instrument is such that there is no improvement in scores due to practice, which implies that repeated measures are also valid measures of cognitive function

As before, we use the model that the vector of measured outcomes on each individual, $\boldsymbol{Y_{ij}}$, is given by:

$$\boldsymbol{Y_{ij}} \sim (\boldsymbol{\mu_i}, \sigma_i^2 V(\boldsymbol{\mu_i}))$$

with $\boldsymbol{\mu_i}$ given by:

$$\boldsymbol{\mu_i} = \beta_{0i}\mathbf{1} + \beta_{1i}\boldsymbol{x}$$

For our example, we assume a known null in which $\beta_0 = 25$ and $\beta_1 = 0.50$. We assume that the clinically significant alternative for which 97.5% power is desired corresponds to $\beta_1 = 0.40$; a 20% reduction in the rate of decline.

To consider the extremes of possible mean-variance relationships and correlations that could occur in practice, we let $\sigma_i^2 = 0.1$ and elements of the covariance matrix be as follows, with $\gamma = 2$ and $\rho_{kk'} = 0.6$.

$$\begin{aligned} V_{kk} &= (\beta_0 + \beta_1 x_k)^\gamma \\ V_{kk'} &= \rho_{kk'} * \sqrt{V_{kk} V_{k'k'}} \qquad\qquad k \neq k' \end{aligned}$$

This example was loosely motivated by a trial done to test for a change in the rate of cognitive decline in Alzheimer Disease (Aisen et al., 2008). As in that trial, we assume that measurements will be taken at baseline, and every three months thereafter until the last measurement 18 months from randomization. For the sake of example, we assume that accrual takes place uniformly over six months and that interim analyses are scheduled at 12.5, 18.2, and 24 months in calendar time from the start of the study. This pattern of accrual relative to follow up is among the most extreme that could reasonably occur.

8.2.1 Design Stage

We plan the trial to use GEE with an independent working covariance matrix that does not account for any mean-variance relationship, knowing that correct standard errors can be obtained with the sandwich estimator as explained previously. Thus, we know that there will not be independent increments under the null, because at no time other than the very beginning and very end of the study is the design completely balanced.

In fact, with analysis times of 12.5, 18.2, and 24, and the assumed variance and corre-

lation structure, the covariance matrix for the slope statistic under the null is:

$$nVar_{null}(\hat{\beta}_1) = \begin{bmatrix} 1.192 & 0.380 & 0.115 \\ 0.380 & 0.350 & 0.138 \\ 0.115 & 0.138 & 0.156 \end{bmatrix}$$

This matrix corresponds to 0.46 on the relative departures from independent increments metric first described in chapter 5. It is 0.129 for the linear trend in departures. Of note, the first and second analysis are more correlated than what would be true under independent increments, but the first and second analyses are less correlated with the final analysis than what would be true under independent increments.

Under the alternative, there would also be departures from independent increments, and the covariance matrix for the slope statistic would be different due to the mean-variance relationship.

$$nVar_{alt}(\hat{\beta}_1) = \begin{bmatrix} 1.098 & 0.348 & 0.107 \\ 0.348 & 0.310 & 0.123 \\ 0.107 & 0.123 & 0.134 \end{bmatrix}$$

This covariance matrix corresponds to a relative departure from independent increments of 0.41 and a linear trend of 0.10.

The information growth is also different between the null and alternative. Under this null, the information growth would be: 0.13, 0.44, 1. Under the alternative, the information growth would be 0.12, 0.43, 1. In this case, the lower than expected increase in variance of measurements later in the study decreases the fractional amount of information at the earlier interim analyses relative to the final.

At the design phase, the relative conservatism of efficacy and futility boundaries should be decided upon for scientific and statistical reasons. For the sake of this example, we will plan for an O'Brien-Fleming efficacy boundary and a Pocock futility boundary. The z-statistic boundaries for this choice of efficacy and futility boundaries under the null information growth and under the alternative information growth are shown in table 8.1. In this example, we select boundaries that use the efficacy boundary under the null information growth, the futility boundary under the alternative, and then plan the final analysis using

the constrained boundary approach under the null. The mixed, final planned boundaries are also shown in table 8.1. For ease of illustration, these boundaries correspond to the difference between the observed and the null slope, so more negative numbers correspond to a larger treatment effect.

Table 8.1: Z-Statistic boundaries using the information growth under the null, the alternative, and a mixture of the null and alternative.

	Null IG		Alt. IG		Mixed	
Analysis #	a	d	a	d	a	d
1	-5.31	0.80	-5.50	0.85	-5.31	0.85
2	-2.88	-0.50	-2.92	-0.46	-2.88	-0.46
3	-1.92	-1.92	-1.92	-1.92	-1.93	-1.93

The appropriate sample size (number of individuals to be accrued) is then calculated from the z-statistic critical value with 97.5% power of the design, the alternative treatment effect with 97.5% power, and the other parameters which affect the standard error at the final analysis: the assumed mean-variance relationship, correlation, and σ^2. In this example, this works out to $n = 119$ individuals.

8.2.2 Evaluation

Under the null, the empirical type I error is 0.024. Under the alternative with desired 97.5% power, the realized power is 98.1%. To evaluate the design over a broader range of alternatives, we chose to look at alternatives that would have 25, 50, 80, 90, and 97.5% power if the final information and information growth were the same for these alternatives as under the null and assuming independent increments. The empirical power curve for these alternatives along with the assumed power curve is shown in figure 8.1.

It is worth noting that although the expected power does not match exactly what we would assume using standard methods, the differences are small and correspond to having greater than expected power for the alternative. We have seen that this result occurs under

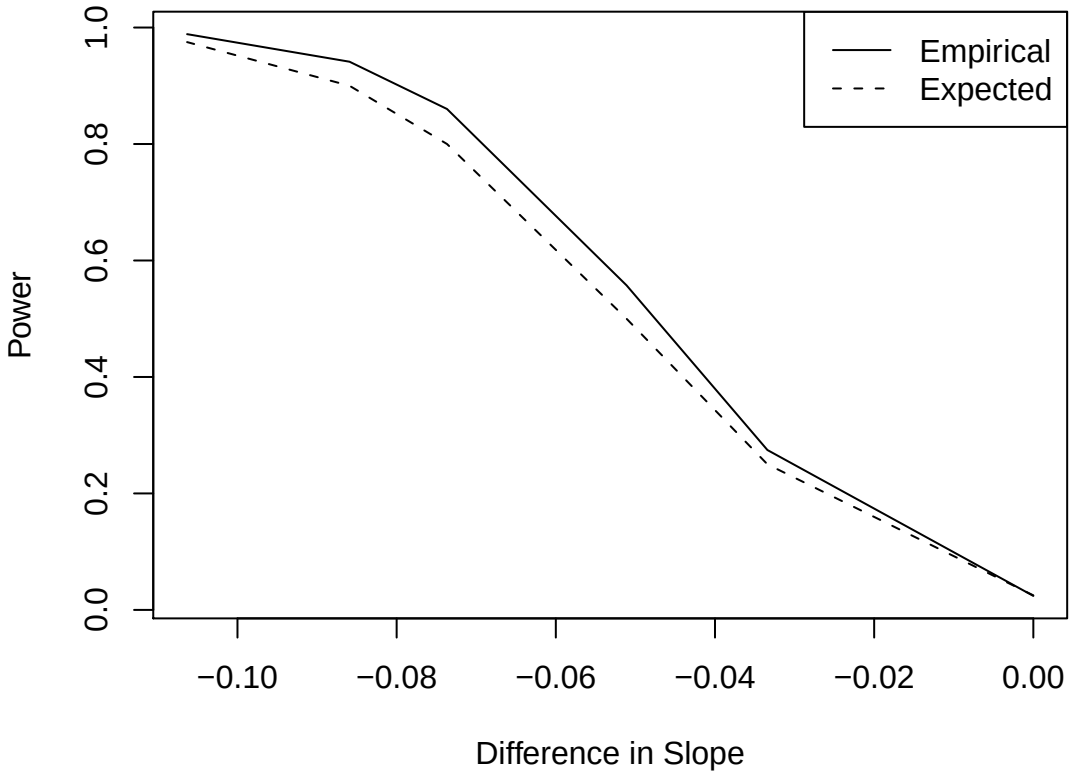

Figure 8.1: Power curves for the true (empirical) power at various alternatives and the expected power using all assumptions.

a positive alternative as well, once an adjustment is made for the change in final information due to the alternative.

The coverage of 95% confidence intervals constructed after adjusting the sampling density for true information growth (if observed) or simply the observed variance but assuming the null information growth (if the trial is stopped early) is shown in figure 8.2. Under a variety of alternatives, this method provides appropriate confidence intervals after simple adjustment.

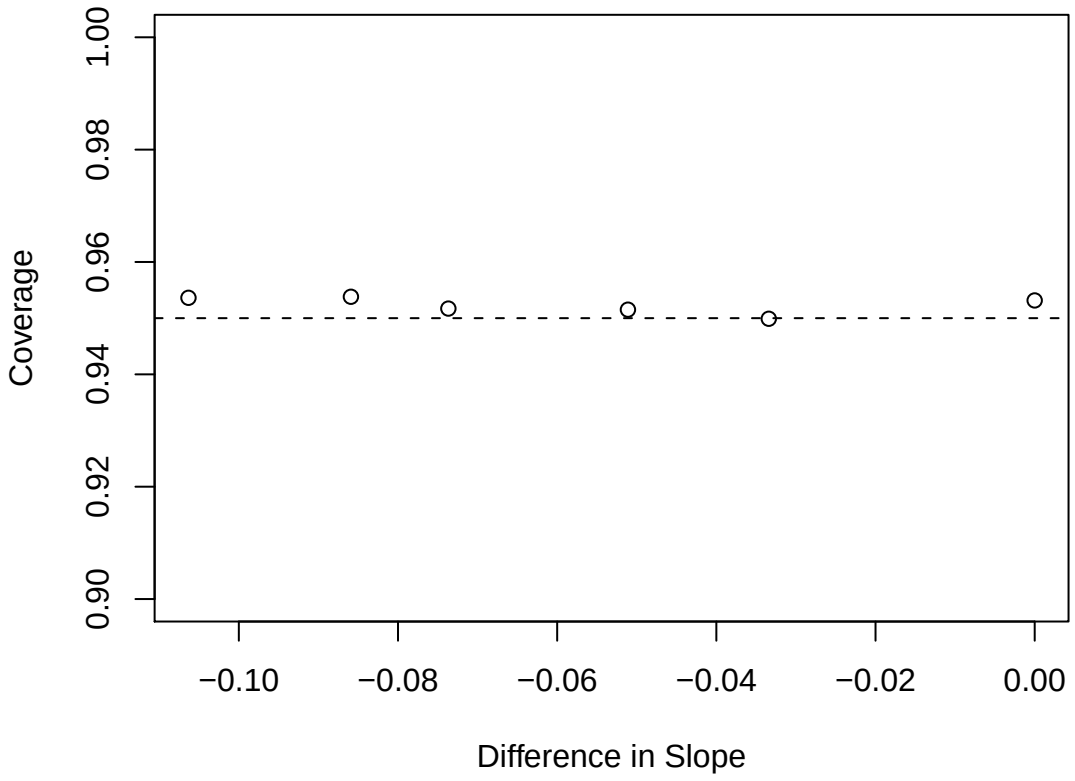

Figure 8.2: Empirical coverage probabilities for 95% confidence intervals at various alternatives.

8.3 Discussion

We have seen that mild to moderate departures from independent increments do not generally cause difficulty in estimating the sampling density. Under most reasonable boundaries, type I error rate is maintained very near the nominal level. Further, even under moderate departures, with standard boundaries the test becomes slightly conservative, which is certainly preferable to becoming anti-conservative in this setting. Similarly, with most standard boundaries, departures from independent increments do not have a great effect on the assumed power, and departures from independent increments tend to lead to greater than anticipated power when differences are observed.

In settings where there is a mean-variance relationship, it is important to account for the mean-final information relationship when calculating power. Once this adjustment is made, however, the differences in information growth due to different alternatives does not cause much difficulty under most reasonable conditions.

It is important for future study designers to make reasonable choices for boundaries to avoid extreme situations caused by very aggressive early boundaries. These situations can lead to poor statistical properties, as can extreme correlations or heteroscedasticity (due either to predictor-variance or mean-variance). If extreme correlations or heteroscedasticity is anticipated, planned interim analyses should not be spaced such that a time of near balance is followed immediately by a time when only a handful of individuals have a measurement at the final study time. Even in extreme situations, longer accrual relative to follow up will mitigate the situation somewhat.

Although we considered cases in which the study times were fixed and constant for all individuals, our results will likely generalize to the more common situation in which study times vary by subject – assuming that the observed study time is independent of the treatment group. The results should also generalize to nonlinear models with GEE, which might be used when looking at a change in rate of Poisson data.

BIBLIOGRAPHY

Aisen, P. S., Schneider, L. S., Sano, M., Diaz-Arrastia, R., van Dyck, C. H., Weiner, M. F., Bottiglieri, T., Shelia, J., Stokes, K. T., Thomas, R. G., and Thal, L. J. (2008). High-dose B vitamin supplementation and cognitive decline in Alzheimer disease. *Journal of the American Medical Association* **300,** 1774–1783.

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

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

Chang, M. N. and O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials* **7,** 18–26.

Crowder, M. (1995). On the use of working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82,** 407–410.

Emerson, S. S. (1996). Statistical packages for group sequential methods. *The American Statistician* **50,** 183–192.

Emerson, S. S. (2000). S+seqtrial technical overview. *Technical Report, Insightful Corporation, Seattle, Washington* .

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

Fitzmaurice, Garrett, M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51,** 309–317.

Gange, S. J. and DeMets, D. L. (1996). Sequential monitoring of clinical trials with correlated responses. *Biometrika* **83,** 157–167.

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

Hanley, J. A. (2005). Analysis of mortality data from cancer screening studies: Looking in the right window. *Epidemiology* **16,** 786–790.

Jennison, C. and Turnbull, B. W. (1997). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92,** 1330–1341.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods With Applications to Clinical Trials*. CRC Press.

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

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

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

Lee, S. J., Kim, K., and Tsiatis, A. A. (1996). Repeated significance testing in longitudinal clinical trials. *Biometrika* **83,** 779–789.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalizeed linear models. *Biometrika* **73,** 13–22.

Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50,** 270–278.

Mancl, L. A. and Leroux, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics* **52,** 500–511.

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

Pampallona, S., Tsiatis, A., and Kim, K. (1995). Spending functions for the type I and type II error probabilities of group sequential tests. *Technical Report, Harvard University, Dept. of Biostatistics* .

Pepe, M. and Anderson, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics Simulation and Computation* **23,** 939–951.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–200.

Proschan, M. A., Follman, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* **48,** 1131–1143.

Rudser, K. D. and Emerson, S. S. (2008). Implementing type I and type II error spending for two-sided group sequential designs. *Contemporary Clinical Trials* **29,** 351–358.

Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. (1997). Semiparametric efficiency and its implications on the design and analysis of group sequential studies. *Journal of the American Statistical Association* **92,** 1342–1350.

Tsiatis, A. A., Rosner, G. L., and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40,** 797–803.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43,** 193–199.

Wang, Y. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations. *Biometrika* **99,** 29–41.

Wei, L., Su, J. Q., and Lachin, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* **77,** 359–364.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48,** 817–838.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73,** 573–581.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials.* John Wiley & Sons.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions (corr: V39 p1137). *Biometrics* **39,** 227–236.

Wu, M. C. and Lan, K. G. (1992). Sequential monitoring for comparison of changes in a response variable in clinical studies. *Biometrics* **48,** 765–779.

Yan, J. and Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine* **23,** 859–880.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation using a partly exponential model. *Journal of the Royal Statistical Society, Series B, Methodological* **53,** 805–811.

VITA

Abigail Shoben was born in Urbana, Illinois. She earned a Bachelor of Science degree in Chemistry from Brown University in 2002. In 2003, she earned a Masters in Education in Science, Mathematics, and Technology Education from The Ohio State University. She taught high school chemistry in Rockville, Maryland prior to returning to graduate school In 2010, she earned a Doctor of Philosophy from the University of Washington in Biostatistics.