TARGETED THERAPIES: ADAPTIVE SEQUENTIAL DESIGNS FOR SUBGROUP SELECTION IN CLINICAL TRIALS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Baldur Páll Magnússon

January 2011

© 2011 Baldur Páll Magnússon

ALL RIGHTS RESERVED

TARGETED THERAPIES: ADAPTIVE SEQUENTIAL DESIGNS FOR SUBGROUP SELECTION IN CLINICAL TRIALS

Baldur Páll Magnússon, Ph.D.

Cornell University 2011

A critical part of clinical trials in drug development is the analysis of treatment efficacy in patient subgroups (subpopulations). Due to multiplicity and the small sample sizes involved, this analysis presents substantial statistical challenges and can lead to misleading conclusions. In this thesis, we develop methodology for statistically valid subgroup analysis in a variety of settings. First, we consider a number of trial designs of varying flexibility for the case of one subgroup of interest. Some procedures are novel, while others are adapted from the literature. Included is data-driven consideration of adaptive change of subject eligibility criteria—known as adaptive enrichment—whereby apparently nonresponsive patient populations are not recruited after data has been unblinded for an interim analysis. We conduct an extensive numerical study to investigate design operating characteristics, as well as sensitivity to subgroup prevalence and interim analysis timing. We observe that power gains can be substantial when a treatment is only effective in the subgroup of interest. Following this example, selected procedures are generalized to allow for analysis of an arbitrary number of subgroups.

Next, we propose a K-stage group sequential design that can be applied as a confirmatory seamless Phase II/III design. The design is specified through upper and lower spending functions, defined in terms of calendar times. After the first stage, poorly performing subgroups are eliminated and the remaining population is pooled for the duration of the trial. This procedure combines the elimination of non-sensitive subgroups with the definitive assessment of treatment efficacy associated with traditional group sequential designs. Numerical examples show that the procedure has high power to detect subgroup-specific effects, and the use of multiple interim analysis points can lead to substantial sample size savings. We address the challenges of adjusting for selection bias, and protecting the familywise error rate in the strong sense.

All designs are presented either in terms of standardized test statistics or the efficient score, making the analysis of normal, binary, or time-to-event data straightforward.

BIOGRAPHICAL SKETCH

Baldur was born in 1981 in Reykjavík, Iceland, but resided with his family in the southern part of Sweden until the age of ten. After moving back to Iceland, Baldur spent most of his teenage years planning to become a commercial airline pilot. However, after obtaining a private pilot license, he decided to check out this “university thing” everyone was talking about. On something of a whim, the subjects mathematics and computer science were chosen, and the career as a pilot was officially put on hold. After one year of studies at the University of Iceland, he took a music scholarship and transferred to Stetson University in sunny Florida. Over the next two and a half years, he divided his time between the Stetson orchestras and computer labs, before graduating in December 2005. After a brief foray into the job market in Iceland as a software developer, Baldur came to Cornell University in the fall of 2006 to commence his PhD studies. It was in Ithaca that Baldur discovered a passion for two new hobbies: ice hockey and enjoying various craft brews. Now, after four and a half years of studies, too many hours of hockey to count, and numerous trips to the Finger Lakes Beverage Center, Baldur is finally ready to move on. After graduation, Baldur will spend two months in Ithaca as a short term post-doc researcher before joining Novartis Pharma in Basel, Switzerland.

This thesis is dedicated to my wife Deanna, and to my parents Halla and Magnús.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and committee chair, Bruce Turnbull. His guidance and input—sometimes research-oriented and sometimes not—has been very helpful during my time at Cornell. I have learned a lot from Bruce, both during our research meetings and also as his teaching assistant. Bruce was also a valuable source of information and encouragement for me as I took on the role of course instructor during Cornell summer sessions.

I also wish to thank the entire ORIE faculty. In particular, I thank my other committee members, David Ruppert and Robert Jarrow, for their interest and helpful comments. I thank Philip Protter for his guidance during my early years as a graduate student. I also wish to thank Peter Jackson; it was a pleasure serving as a teaching assistant for him. Finally, I thank Bob Bland and Shane Henderson for their encouragement and positive feedback for my teaching duties as a course instructor.

Special thanks to the ORIE administrative staff. In particular, I thank Kathryn King for all her help, starting with her response to my first email to the department back in 2005.

Thanks to Collin, Kathy, Jie, and Tia for being great office mates. You made the stay here at Cornell very enjoyable! I want to thank fellow Icelanders Matthías Kormáksson and Ýmir Vigfússon for their enthusiasm and help as I made my way through the application process and also for providing great company throughout the years here in Ithaca. I also thank all the following for being good friends during the stay here: Joe, Dennis, Tim, Matt (McLean), Rolf, Sophia, Martin, Matt (Maxwell), Caroline, Gwen, Steve, Sam, and Tuohua. (This is but a sample; the list could go on for many pages.) Special thanks go to Stefan and Pascal for recruiting me to play hockey, and to Frans and Peter for taking on mentoring roles

during my first year here.

I give very special thanks to my family. My parents have been a constant source of invaluable support, and their positive outlook has always encouraged me to follow my interests and desires, wherever they may take me. Finally, I cannot express enough gratitude to my wife Deanna. Without her love and encouragement I would not be who I am today, and her positive attitude was always helpful during the more difficult times of my studies. Additionally, her patience while I worked on a PhD in Ithaca (of all places) is thoroughly appreciated!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

α	Type I error probability
β	Type II error probability
ℓ	Number of populations in a clinical trial
$\mathcal{I}$	Information level
Ω	Patient population
Φ	Standard normal cumulative distribution function
$\mathcal{P}$	Population index set
φ	Standard normal probability density function
f	Subgroup prevalence
K	Number of analyses/stages in a clinical trial
Y	Efficient score
Z	Standardized test statistic

AFP	Adjusted Fallback Procedure
ASD	Adaptive Seamless Design
CDF	Cumulative Distribution Function
CER	Conditional Error Rate
CP	Adaptive design with interim decisions based on Conditional Power
CTP	Closed Testing Procedure
DR-I/DR-II	Decision Rule I or II, see Chapter 5
EMEA	The European Agency for the Evaluation of Medicinal Products
FDA	The US Food and Drug Administration
FE	Adjusted Fallback procedure with Enrichment
FWER	Family-Wise Error Rate
GSDS	Group Sequential Design with Subgroup selection
HPP	Hybrid design with interim decisions based on Predictive Power
HS	Hochberg-Simes method
HUT	Hybrid design with Utility-based interim analysis
MCP	Multiple Comparison Procedure
OBFm	O'Brien and Fleming design adjusted for multiplicity
pCR	Pathologic Complete Response
RCT	Randomized Controlled Trial

Chapter 1

Introduction

1.1 Motivation

Randomized, placebo-controlled clinical trials are the gold standard in medical experimentation (Freidlin and Simon, 2005a; Jiang et al., 2007; Strassburger et al., 2004; Wang, 2007; Zhou et al., 2008); traditionally, the aim of such trials is to demonstrate the effectiveness of an experimental treatment in a broadly defined patient population. It is however increasingly apparent that patient responses to a particular treatment can vary considerably. For example, the antibody Herceptin is currently an approved treatment for breast cancer patients who show a positive HER2 expression (Romond et al., 2005). No definitive assessment of the effect of Herceptin on HER2 negative patients has been conducted, and such patients are therefore advised against its use. However, recent research has indicated that Herceptin may be effective in a larger population (Paik et al., 2007), which raises the question of whether HER2 positivity may need to be redefined. Had a larger clinical trial been conducted in the first place, allowing for both HER2 positive and negative patients, it is entirely possible that we would have a clearer indication of the true effectiveness of Herceptin today. This example is not unique, and it highlights the need for clinical trial designs that enable the assessment of treatment effectiveness in an overall population, as well as in interesting subgroups. Our research is focused on developing statistical methodology for precisely this purpose. We propose a number of testing procedures applicable in a variety of settings, and we find that our designs can outperform conventional trial designs when treatment effectiveness is limited to smaller populations, while still performing acceptably when the treatment is broadly effective.

This chapter is outlined as follows. We begin with a brief overview of the various stages of a therapeutic development program, followed by a more detailed motivation for subgroup analysis in medical experimentation, including several practical examples from current medical research. A literature review is given in Section 1.2. We define various important terms and concepts, detail some of the difficulties involved with subgroup analysis, and survey existing methods. Finally, a brief overview of the thesis is given in Section 1.3.

In order for an experimental therapeutic intervention to make the transition from laboratory testing to practical use, a development program assessing its efficacy and safety must be conducted. Such programs are traditionally split into four or five different phases, each designed to answer a separate research question (Chang, 2010, Ch. 3.4):

- **Phase 0** is a small study which involves no estimation of efficacy nor safety, as doses are normally too small to have any therapeutic effect. Single doses are given to few (often 10-15) subjects, with the purpose of assessing how the drug functions in, and is processed by the human body.
- **Phase I** studies involve the first testing of the drug in small groups (typically 20-100) to evaluate its safety and identify side effects. Phase I also involves dose escalation, which is a preliminary study aimed to find dose levels appropriate for therapeutic use. Usually, healthy volunteers are used for Phase I, though there are circumstances where terminally ill patients, who lack other treatment options, may be enrolled. Sometimes, so-called Proof-of-Concept (PoC) studies are conducted during Phase I. Such studies are conducted on small and well-defined target populations and are intended to allow for a

quick demonstration of therapeutic benefit. If a PoC study is successful, development may proceed aiming at a more broadly defined population.

- **Phase II** is conducted once initial safety concerns have been addressed in Phase I. Not all treatments make it from Phase I to Phase II. Larger groups are involved at this stage; typically 30-300 individuals are enrolled, and the drug is now tested on patients rather than healthy volunteers. The aim of a Phase II study is: safety assessment in larger groups of patients; learn about side effects and how to manage them; select the appropriate dose for therapeutic effect (referred to as Phase IIa); and learn whether the treatment is effective enough to warrant a larger trial (Phase IIb). Phase II is exploratory, and usually no definitive assessment of treatment efficacy is carried out.

- **Phase III** trials are carried out if Phase II investigations indicate that the treatment is efficacious and safe. These are large-scale, confirmatory trials, involving many hundreds or thousands of patients. Typically, trials are randomized, double-blinded and placebo-controlled to maintain the integrity of measured results. Randomized, placebo-controlled studies randomly assign patients to receive either the experimental treatment or a placebo. Studies are double-blinded if neither the subjects, nor the clinical researchers are aware of which treatment is being administered to any subject. Phase III studies are both lengthy and expensive, and are subject to regulations published by the FDA (US Food and Drug Administration, 1998). At the conclusion of a Phase III trial data are carefully analyzed in concordance with statistically valid inference procedures to determine whether the treatment is efficacious (and safe), and whether it should be made commercially available.

- **Phase IV** is known as a post-marketing surveillance trial as studies are con-

> ducted after the drug has been marketed. First, a retrospective analysis can be carried out on the data accumulated during Phase III, where information is gathered on possibly differing effects in various subgroups (subsets of the study population) and side-effects associated with long-term use. Second, Phase IV involves safety surveillance and technical support as the drug is sold to the general public. Harmful effects discovered during this phase may result in the drug being removed from the market.

Returning to the discussion in the opening paragraph, a frequent complication in clinical trials is the fact that patient populations are often heterogeneous with respect to therapeutic response. In the following paragraphs, we give several examples, which illustrate the clinical relevance of subgroup analyses, and highlight the need for statistically rigorous procedures with which subgroup-specific effects may be analyzed.

Kirsten rat sarcoma (KRAS) mutations are common oncogenic mutations that play a role in many cancers, and are associated with various malignancies. Retrospective analysis of data from several clinical trials has suggested that KRAS positive patients may be responsive to the antibodies *panitumumab* and *cetuximab*. Based on the preliminary evidence gathered in these trials, a targeted study (see Section 1.2.3) to further test the responsiveness of KRAS positive tumors to these antibodies is both scientifically and ethically valid (Mandrekar and Sargent, 2009).

Another example concerns HER2 (see opening paragraph), or "Human Epidermal growth factor Receptor 2," which is a protein giving higher aggressiveness in breast cancers. It has been estimated that roughly 20% of breast cancers exhibit an amplification of the HER2 gene (Nahta and Esteva, 2003). The antibody *trastuzumab* (marketed as Herceptin) has been associated with increased survival rates among HER2 positive breast cancer patients and is currently approved for treatment of such cancers. This was established with a targeted trial for which only HER2 positive patients were recruited, ostensibly resulting in substantial savings in cost and time compared to what might have been the case had an untargeted trial been carried out. There is, however, recent evidence that some HER2 negative patients may also benefit from this drug (Paik et al., 2007), and the question remains open whether Herceptin therapy may benefit a much larger group of breast cancer patients. We discuss this issue in more detail in Section 1.2.2, and in Section 1.2.3 when we introduce the concept of enrichment.

A BRCA mutation is a mutation in either of the genes BRCA1 or BRCA2 (Petrucelli et al., 1997). Patients carrying mutations in either of these genes are predisposed to breast cancer and ovarian cancer, as well as prostate cancer (BRCA1) and other cancers (BRCA2). The mutation is believed to be hereditary; each offspring of an individual with BRCA1 or BRCA2 mutation has a 50% chance of inheriting the mutation. From the previous paragraph, we know that HER2 overexpression is related to more aggressive breast cancers. There is however recent evidence to suggest that HER2-positive breast cancer and BRCA mutation-associated breast cancer are mutually exclusive diseases (Maynes et al., 2010).

Antidepressants have long been thought of as the best treatment for major depressive disorder. However, meta-analyses conducted by Kirsch et al. (2008) and Fournier et al. (2010) suggest that drug-placebo differences in antidepressant efficacy may increase as a function of baseline severity of depression. In other words, for patients with moderate depression, there may be no discernable difference in efficacy between taking an expensive antidepressant or a sugar pill. Only for the most severely depressed individuals is there a statistically significant benefit derived from treatment. Additionally, according to the analysis of Kirsch et al. (2008), the difference for patients suffering from most severe depression is relatively small. The meta-analysis conducted by Fournier et al. (2010) encompassed a larger collection of clinical trials than that of Kirsch et al. (2008), and the authors report substantial difference from placebo for severely depressed patients only. This topic has received attention in mainstream media, see for example Begley (2010). Today, commercially sold antidepressants are marketed for patients suffering from mild to the most severe depression, but the aforementioned meta-analyses suggest that any future clinical trials take into account the likely heterogeneity of patients.

BRAF is a serine/threonine kinase and is among the most frequently mutated proteins in human cancers (Greenman et al., 2007). The finding of Davies et al. (2002) that mutations in BRAF are particularly common in melanoma (mutations are found in approximately 60–70% of malignant melanomas), has offered hope that inhibition of BRAF kinase activity could benefit melanoma patients. One drug that is currently under study, PLX4032 of Plexxikon (Bollag et al., 2010), has passed through Phase I studies in which the treated group had a median increased survival time of 6 months over control (also see New York Times article by Harmon (2010)), and approximately 80% of patients showed partial to completed regression, though regression lasted only between two to eighteen months (Flaherty et al., 2010). Phase I and Phase II studies of the efficacy of PLX4032 are ongoing, and a Phase III trial has been started.

Any trial which does not account for the possibility of heterogeneity among patient subgroups is only effective for identifying treatments that work "on average" for the whole population of interest. Simon and Wang (2006) report estimates saying that only about 60% of prescriptions written produce the desired therapeutic benefits, while 7% of patients suffer from serious consequences due to negative side effects. Moreover, if an experimental drug only benefits a small portion of the target population, a confirmatory Phase III trial is unlikely to yield positive results as the estimated treatment effect will be diluted by nonresponsive patients. For example, the Phase III trial "Iressa Survival Evaluation in Lung Cancer" failed to show overall benefit of treatment with the tyrosine kinase inhibitor *gefinitib* versus placebo, but retrospective subgroup analyses have indicated that there are significant benefits for patients with somatic EGFR (Epidermal Growth Factor Receptor) mutations (Jänne and Johnson, 2006). Though further (retrospective) studies imply that this EGFR mutation can be used as a predictive marker for lung cancer patient response to kinase inhibitors, definitive assessments of drugs that specifically target the mutation have usually been conducted with conventional trials for broadly defined populations. This has received mainstream media attention; in a 2010 Newsweek article, Begley (2010) discusses some challenges of targeting cancer-driving mutations. Problems are exacerbated by the lack of patient genotyping in Phase III trials that are testing experimental drugs specifically designed to target a driving mutation; the result is usually, and perhaps unsurprisingly, a negative trial. It is therefore becoming increasingly apparent that the conventional one-size-fits all approach to randomized controlled trials (RCTs) is inappropriate, inefficient and nonsustainable.

The concept of subgroup analysis is not new; due to the cost and effort required for Phase III trials, subgroup analyses have commonly been carried out in order to extract as much information from the data as possible. At the start of clinical trials, investigators gather a great deal of baseline data on each patient, documenting the patient's current condition and medical history, and regulatory guidance already recommends that some analysis by demographics such as age, race and gender take place during the trial. Further, advances in human genomics-based studies have led to the recognition that phenotypically[1] homogeneous patients may be heterogeneous at the genomic level. In pharmacogenomics[2] (Simon and Wang, 2006) there has thus been an increased focus in searching for combinations of individual genes to form classifiers that predict a patients' therapeutic response.

The statistical validity of subgroup analysis in published clinical trial reports has often been the cause of contentious debate (Pocock et al., 2002; Yusuf et al., 1991). For instance, the presence of multiple endpoints (subgroups can be formed in many ways), and the possibility of "data-dredging," can lead to results being regarded with some suspicion. This is not wholly unwarranted; without use of proper statistical methodology, a post-hoc search of subgroups exhibiting (possibly spurious) statistically significant treatment effects is likely to be "successful." Triallists are therefore often encouraged to look at data for subgroups, but to not necessarily believe the implied conclusions (we briefly discuss regulatory guidance in Section 1.2.4). However, if in truth there are subgroups of patients for which a new treatment differs from the overall effect, either in terms of efficacy or toxicity, we have a scientific and ethical obligation to identify such subgroups. Hence there is a need to develop a statistically valid methodology to prospectively evaluate treatment efficacy in biologically plausible subgroups.

In this thesis, we consider various designs intended for subgroup analysis in clinical trials. We develop procedures that can handle multiple subgroups, nested or disjoint, as well as procedures that take place over several stages. Performance

[1] Phenotype refers here to any observable trait or characteristic of a patient.

[2] Pharmacogenomics refers to the science of determining how benefits and adverse effects of a drug vary among a population based on genomic features.

of these procedures is evaluated in a number of different scenarios, including comparison with designs that have been proposed in the literature.

1.2 Literature Review

Subgroup analyses are quite common in clinical research; in a survey conducted by Pocock et al. (2002), 70% (35 out of 50) of sampled articles included some form of subgroup analysis. When properly performed, such analysis can yield valuable insight into therapeutic effects; unfortunately, many commonly used inference procedures are inefficient and can produce spurious and misleading results. In this section, we define the types of subgroups that are of interest, highlight some of the difficulties encountered in subgroup analysis, and survey work that has been done over the last few decades.

1.2.1 Subgroups and Biomarkers

Subgroups can be categorized as *proper* or *improper*, see for example (Yusuf et al., 1991) and (Huque and Röhmel, 2010, p. 15). A "proper subgroup" is defined as a group of patients characterized by a common set of baseline characteristics that cannot be affected by treatment (e.g. age or gender), or prognostic disease characteristics defined before randomization (e.g. previous myocardial infarction). On the other hand, "improper subgroups" are defined as a group of patients characterized by a variable measured after randomization and potentially affected by treatment (e.g. subgroups defined by responders vs. non-responders in a cancer trial). Though analysis of improper subgroups can be tempting, such practices can be misleading and suffer from serious selection bias.

If a subgroup is specified before data are unblinded, we say that the subgroup is *prespecified* or *prospectively specified*. Analysis of subgroups that are not prespecified should usually be regarded with suspicion, and any reported significant effects need to be replicated in a separate and independent trial. In this thesis, we propose procedures that sometimes analyze only those subgroups that exhibit a positive early effect. Such subgroups may therefore be improper, and appropriate statistical methodology is required in order to ensure that final results are credible. A part of this methodology is the requirement that all subgroups of interest be prospectively specified in the trial protocol. As a result, procedures can be designed to adjust for the selection bias that results from analyzing improper subgroups. Henceforth, the term subgroup shall be used to refer to a prospectively specified subgroup.

The baseline characteristics that define a subgroup are referred to as biological markers or *biomarkers*, formally defined as follows (Atkinson et al., 2001):

Definition 1.1. *A biological marker (biomarker) is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.*

While “basic” biomarkers such as demographics or medical history have been studied for some time, the advent of genomic technologies, e.g. DNA sequencing or transcription profiling, has made it increasingly apparent that different genomic patient subsets can be heterogeneous in terms of treatment response (Maitournam and Simon, 2005; Simon and Wang, 2006; Wang, 2007). Studies that incorporate microarray data for the purpose of identifying genomic biomarkers in clinical trials are referred to as *pharmacogenomic* studies (Simon and Wang, 2006). The goal of

Table 1.1: Biomarkers as classifiers in clinical trials

	Response Rate		
Marker Status:	High	Low	
Treatment	20	20	No interaction
Placebo	5	5	
Treatment	15	20	Prognostic
Placebo	5	10	
Treatment	20	5	Predictive
Placebo	5	5	
Treatment	20	20	Prognostic-Predictive
Placebo	5	10	

such research is the eventual individualization of therapy.

Biomarkers are classified into *prognostic* markers and/or *predictive* markers (Mandrekar and Sargent, 2009; Sargent et al., 2005; Wang, 2007). A biomarker is prognostic if it separates a population based on disease prognosis (or long-term outcome) when untreated, or receiving standard (untargeted) treatment. On the other hand, if a biomarker separates the patient population based on the outcome of interest in response to a particular treatment or therapy, it is said to be predictive. Biomarkers that are both prognostic of disease state and predictive of drug effect are called prognostic-predictive biomarkers. In Table 1.1, we show various treatment effect outcomes that might be expected in a trial for which a genomic biomarker is used as a classifier to determine the treatment effect in a genomic subset. Prognostic markers hence only distinguish subgroups based on perceived long-term prognosis, but unlike predictive markers they cannot guide the choice of a particular treatment.

We refer to an *interaction* as the case when treatment effect differs by subgroups. When a treatment is beneficial (or harmful) in all subgroups, but the magnitude of effect varies among subgroups, we say that the interaction is *quantitative.* When a treatment is only beneficial in some subgroups, and completely ineffective or harmful in others, the interaction is said to be *qualitative.* For example, we might determine that a new agent is highly effective among male patients, but actually leads to shorter survival among female patients. Note that qualitative interactions can indicate either that positive treatment effect is confined to a particular subgroup (no effect elsewhere), or that treatment effects are in opposite directions in different subgroups (Simon, 2002; Wang et al., 2007b). Opposite treatment effects are generally considered to be highly unlikely (Pocock et al., 2002); when they are in fact observed they are "subsequently shown to be spurious" (Yusuf et al., 1991). Wang et al. (2007b) note that if qualitative interactions truly are present, targeted designs can substantially outperform conventional clinical trials. The magnitude of power gained in targeted trials versus untargeted trials diminishes when qualitative interactions are not allowed. In the construction of our procedures, we assume that opposite treatment effects *do not occur.*

As previously noted, subgroup analyses are frequently conducted in clinical trials, and there are two main reasons for these studies (Song and Chi, 2007; Wang et al., 2007a). First, and perhaps most common, is subgroup analysis carried out when a trial fails to show a statistically significant overall effect, ostensibly in an attempt to salvage the trial. Such analyses are frequently conducted on improper subgroups and/or without proper adjustment for multiplicity. For example, in separate surveys conducted by Pocock et al. (2002) and Wang et al. (2007a), a number of clinical trial reports were examined and subgroup results evaluated. In both surveys, several reports were unclear as to whether reported subgroup analyses were prespecified or *post hoc*, or if tests were adjusted for multiplicity and selection bias. Such results are clearly suspect, can be very misleading, and at best serve as exploratory studies for hypothesis generation. The second reason for subgroup analysis is to investigate the consistency in treatment effect across various subgroups of clinical importance. This may be done to further demonstrate the strength of evidence for treatment efficacy after overall significance has been established, or to assess those populations most (or least) likely to benefit from the treatment.

1.2.2 Subgroup Analysis Concerns

There are several issues and pitfalls associated with analysis and interpretation of subgroup-specific results: lack of statistical power; multiplicity; appropriate tools of inference; interpretations and generalizability; and misclassification.

Lack of Statistical Power

The issue of lack of statistical power is well known and has been discussed by several authors, see for example Maitournam and Simon (2005); Pocock et al. (2002); Sargent et al. (2005); Song and Chi (2007); Wang et al. (2007a); Wang (2007); Yusuf et al. (1991). As most clinical trials are only powered to detect an overall treatment effect, it should not be expected that subgroup effects are detected, even in relatively large subgroups. In particular, if a truly efficacious treatment is being tested and many subgroups are analyzed, we should expect false negatives in some subgroups by chance and lack of power. For example, the ISIS-2 trial (ISIS-2 (Second International Study of Infarct Survival) Collaborative Group, 1988) involved over 17,000 patients and clearly demonstrated the beneficial effect of aspirin for patients experiencing a heart attack. To give an example of the limitations of subgroup analysis, the authors note that "subdivision of the patients in ISIS-2 with respect to their astrological birth sign appears to indicate that for persons born under Gemini or Libra, there was a slightly adverse effect of aspirin on mortality, while for patients born under all other astrological signs there was a striking beneficial effect." The finding indicating that Gemini and Libra persons had an adverse effect was of course spurious, and serves to illustrate that subgroup analysis can quite easily lead to incorrect findings, even in trials that are positive overall. We note, however, that it is not impossible for subgroup analyses to be properly powered. Alosh and Huque (2009) point out that an increase in subgroup power, relative to that of the overall population, can be achieved through one of the three following: an increase in the subgroup prevalence, a higher treatment effect in the subgroup relative to the overall population, and higher measurement precision in the subgroup compared with the overall study population. While these factors are of course not controlled by clinical researchers, the fact is that contrary to what has previously been claimed, subgroup studies are not *necessarily* underpowered.

Multiplicity

Given the large amount of observable clinical characteristics and underlying genomic characteristics, there are many possible subgroup analyses that could be performed. When multiple subgroups are analyzed, the probability of a false positive can be substantially inflated. For example, consider a trial comparing an ineffective treatment to a control. If subjects are divided into 20 mutually exclusive subgroups, we should expect one spurious finding on average, using a nominal significance level of 5%. There are many ways to adjust for multiplicity, some of which we will discuss in Chapter 2. However, for coherent inferences on treatment effects in subgroups, multiplicity adjustment alone is not enough. Subgroups should be prespecified[3], based on biomarkers developed in earlier stages (or from previous trials), and have a strong biological and clinical rationale (Pocock et al., 2002; Yusuf et al., 1991). Medical reasons supporting a particular subgroup hypothesis should be explicitly stated at the outset and analysis of the resulting data should be carried out in concordance with the trial protocol. Subgroups that are defined *a posteriori* can serve as a basis for hypothesis generation for future research, but to avoid selection bias these hypotheses should only be tested with new and independently obtained data. Careless analysis of multiple subgroups can be highly misleading, in particular if subgroup definitions are data-driven (improper) and *post hoc* emphasis is placed on the "most interesting" subgroup finding. Prospective planning is essential to all effective trial designs, and this is holds true no less when subgroup analysis is envisaged.

Appropriate Tools of Inference

Recognizing the likely very low power of multiplicity-adjusted interaction tests, Yusuf et al. (1991) recommend that the prudent way to conduct subgroup analysis is to rely mostly on the overall results to indicate likely "true" effects in subgroups. Further, they advocate that medically interesting data-derived subgroup effects be reported clearly as *post hoc* analysis so that the resulting hypotheses can be studied in future trials. However, considering the high cost and significant time involved in a confirmatory trial, this approach is likely very inefficient. Alternatively, Pocock et al. (2002) advise that the only appropriate statistical method for making inferences from subgroup analyses is the test for interaction. Such tests are notoriously underpowered, a fact that has frequently been used to argue

[3] See Section 1.2.4 for a discussion on regulatory guidance for subgroup analysis.

against their use, but Pocock et al. (2002) suggest that their lack of power recognizes the limited information available for subgroup analysis. However, interaction tests only aim to demonstrate that treatment effect is heterogeneous across the subgroups of interest, and there are instances where presence of heterogeneity is not the appropriate question (Moyé and Deswal, 2001). Indeed, when validating a predictive biomarker, we are interested in whether or not the experimental treatment is efficacious in the subgroup of interest (defined by the biomarker). In such cases interaction tests may not be suitable.

Sometimes there is compelling evidence that a treatment affects only a specific population, and a confirmatory trial can be conducted randomizing only patients from said population. However, when there is no biological plausibility or no well established drug target, it is generally not advisable to exclude patients on the basis of a potential biomarker. In such cases, a composite objective should be prospectively defined, whereby researchers will test the hypotheses that the treatment works in all randomized patients, or in the subgroup defined by the biomarker. As there is more than one opportunity to declare a treatment effect, adjustment for multiplicity is necessary.

Interpretations and Generalizability

Care needs to be taken when interpreting trial results that involve subgroup analysis. While researchers need to evaluate whether a treatment effect is generalizable to a larger set of patients, they also need to guard against making sensational conclusions regarding subgroup results that are not robust to validation trials. Surveys of clinical trial reports (Pocock et al., 2002; Wang et al., 2007a) have led to the view that results of subgroup analyses are generally overinterpreted by authors, which can weaken the foundation on which such research is built. Investigators should prospectively specify biologically plausible subgroups at the onset of Phase III trials (development of biomarkers can take place over earlier phases), and all medically interesting hypotheses should be clearly defined. During specification of trial protocols, the consequences of prospectively defining subgroups of interest should be carefully weighted, in particular with respect to the potential generalizability of the trial results. If analysis of the resulting data leads to data-driven definitions of subgroups not specified in the trial protocol, the medical relevance of these subgroups should be evaluated and if deemed important enough, clearly stated in the trial results as potential for further research.

Mandrekar and Sargent (2009) suggest that validation of predictive biomarkers may well be achieved by utilizing data from a previously conducted randomized controlled trial (RCT), known as *retrospective* analysis. They argue that this is a valid strategy when "(1) a prospective RCT is ethically impossible based on results from previous trials, and/or (2) a prospective RCT is not logically feasible (large trial and long time to complete)." A known problem for retrospective analyses is a "convenience sampling bias" resulting from the fact that not all patients may have consented to give samples required for marker validation (e.g. cancer biopsies). Samples for a large majority of patients must therefore be available, and all subgroup hypotheses must be clearly stated up front. The authors go on to argue that findings from well designed retrospective analyses may be sufficient to establish predictive utility of a biomarker and to move it into clinical practice.

Retrospective exploration may also be conducted as a sort of meta-analysis, combining data from several RCTs. In such studies, the right biomarkers need not necessarily be known beforehand, and data can be divided to validate biomarkers defined *a posteriori* in one of the studies. Retrospective exploration also allows for evaluation of treatment effect in all patient subgroups (even if some trials were not designed to answer such questions), and the refinement of previously defined biomarkers. However, presence of convenience samples can make statistical inference problematic. When retrospective development of a genomic biomarker is of interest, difficulties may arise as sample collection and handling may have been suboptimal, and some sample data may be missing with no apparent method for imputation available (Wang, 2007). Simon (2005) points out that if a commercial therapy is already available, the drug sponsor may not have sufficient incentive to engage in the lengthy and expensive process of biomarker validation, even if there is strong suspicion that responses truly differ by patient subsets.

Misclassification

In all but the simplest of cases, most baseline data are measured with some chance of error. This holds true in particular for the evaluation of a molecularly targeted treatment when there is an assay predictive of which groups of patients will be more responsive than others. When this occurs, observed subgroups will include misclassified patients, which can adversely affect the efficiency of any proposed clinical trial. Simon and Maitournam (2004) and Maitournam and Simon (2005) study the effect of assay performance on the relative efficiency of targeted trials (recruit only patients predicted to be responsive) compared to untargeted trials (recruit all patients). The designs are compared in terms of number of patients required for randomization and the number of patients required for screening, as a function of the subgroup prevalence (prevalence of marker-positive patients) and the assay performance. The two measures used for assay accuracy are (1) *specificity*, the probability that a marker-negative person be correctly identified as such; and (2) *sensitivity*, the probability that a marker-positive person be correctly identified as such. The authors highlight that when treatment effect is limited to marker-positive patients, relative efficiency is primarily affected by assay specificity. When specificity is large enough (greater than 0.8), the targeted design is generally more efficient in terms of number of patients randomized. When the treatment effect for marker-negative patients is half of that for marker-positive patients, efficiency gains are minimal unless the assay is perfect. Comparing relative efficiency in terms of number of patients screened, assay sensitivity is of some importance as well. When there is no treatment effect in marker-negative patients there are only efficiency gains when prevalence is low, and specificity and sensitivity are both high. When there is some treatment effect in marker-negative patients (half of marker-positive), an untargeted design is more efficient even if the assay is perfect.

Case Study: HER2 marker status and Herceptin

As stated in Section 1.1, the antibody *trastuzumab* (Herceptin) is currently an approved treatment for HER2 positive breast cancers. A patient is defined as HER2 positive if either (1) immunohistochemistry (IHC 3+) indicates over-expression of the HER2 protein, or (2) HER2 gene amplification by fluorescence in situ hybridization yields a FISH-HER2:CEP17 ratio of ≥ 2.0 (Perez et al., 2006). Following analysis of preclinical evidence, the National Cancer Institute (NCI) decided to enroll HER2 positive patients in two trials intended to evaluate the efficacy of Herceptin, combined with standard chemotherapy, in the adjuvant setting[4]. The two trials, (1) the National Surgical Adjuvant Breast and Bowel Project trial (NSABP B-31), and (2) the North Central Cancer Treatment Group trial (NCCTG

[4] In oncology, adjuvant therapy refers to additional treatment usually given after surgery where all detectable disease has been removed, but where there remains a statistical risk of relapse due to occult disease.

N9831), were analyzed in a combined fashion. Roughly 3,700 patients were enrolled in total and after 394 events had been observed (by March 2005), early stopping boundaries were crossed with results indicating a clinically and statistically significant beneficial effect for HER2 positive patients (Romond et al., 2005). Though the goal of the trial was accomplished, some concerns linger regarding the use of HER2 as a marker for Herceptin.[5]

Originally, patients were eligible for the two trials if they tested positive in either IHC or FISH, and tests were carried out at various locations such as community hospitals, medical centers and national laboratories. In 2002, the trial protocol was modified to make central laboratory testing mandatory, and all previously tested specimen were re-tested at the Mayo Medical Laboratories, Rochester, MN. Analysis of these tests was conducted by Perez et al. (2006), wherein discordance between local and central testing was reported as high as 25% in some cases. While this discordance is alarming, it has permitted retrospective analysis of Herceptin benefits for HER2 negative patients. Paik et al. (2007) report results from subsequent analysis of available tumor tissues from the NSABP B-31 trial which indicate that patients which were negative for FISH and had less than 3+ staining intensity on IHC may derive benefit from Herceptin (relative risk was 0.36, CI = $(0.14, 0.92)$, $p = 0.032$). In a similar analysis of data from the NCCTG N9831 trial, Perez et al. (2007) observe a non-significant reduction in risk (relative risk = 0.51, CI = $(0.21, 1.2)$ and $p = 0.13$). Though small sample-size concerns are present in the analysis, as well as issues relating to convenience sampling bias, the results should not be ignored. Perez et al. (2007) suggest that the findings be used as hypothesis

[5] This has received national media attention; for example, Kolata (2010) discusses the problematic lack of reliability of HER2 tests today, where some patients are diagnosed as partly HER2 positive, and partly negative. Similar difficulties are highlighted in a recent discussion by Speed (2010). This, coupled with the dangerous side-effects of Herceptin, makes the choice of therapy less than straightforward for many patients.

generation for further studies, while Paik et al. (2007) conclude that the current definition of HER2 positivity may need to be modified, as some HER2 “negative” patients may yet benefit from Herceptin.

Following the results discussed above, Mandrekar and Sargent (2009) point out that, though the trials were positive and Herceptin is proven to be a highly effective drug for HER2 positive patients, exclusion of HER2 negative patients means that no definitive assessment of the predictive utility of HER2 can be conducted. The authors highlight the possibility that Herceptin potentially may benefit a larger population than was originally believed, and that further studies to this effect are warranted. This raises the question of whether an alternate design, one which allows inclusion of HER2 negative patients, might have been more useful (and ultimately more efficient (Mandrekar and Sargent, 2009)) than the targeted design that was employed.

1.2.3 Survey of Existing Methods

One of the likely causes of low success rates for untargeted RCTs is disease heterogeneity or varying response rates among patients. Prospectively planning the analysis of (predictive) biomarker validity is therefore key in the planning of future clinical trials. In this section we briefly discuss some designs that have been proposed as a solution to this problem.

Sargent et al. (2005) and Mandrekar and Sargent (2009) discuss the efficiency of various randomization schemes designed to allow for biomarker validation without needlessly excluding a number of patients. There are two main categories of predictive biomarker studies (for one putative marker). First, *Marker by treat-*

ment interaction designs split the population into two groups by marker status. Patients in each group are randomly assigned to one of the two treatments under consideration (one may be placebo), and testing is carried out either in the form of separate tests of superiority, or a formal statistical test for interaction between marker status and treatment assignment. Second, *Marker-based strategy designs* begin by testing each patient for marker status, after which each patient is randomly assigned to have his/her treatment determined by their marker status or to receive therapy independent of marker status. Patients in the marker-based arm receive placebo treatment if their marker status is negative, and the experimental treatment if their marker status is positive. Patients in the non-marker-based arm undergo a second randomization to determine whether they receive placebo or the experimental treatment. The purpose of this second randomization is to clarify whether findings regarding the predictive/prognostic utility of the biomarker is truly due to the marker, or just due to an improvement offered by the experimental treatment in the whole population.

Marker by treatment interaction designs using separate tests are essentially two separate (independent) clinical trials, which requires that they both be powered individually. The trial may fail to provide a clinically useful result if the biomarker is both prognostic and predictive (marker positive patients have worse prognosis with no treatment, but benefit more from the experimental treatment). Further, not all subjects are utilized in one test, making the design inefficient. Using a test of interaction would alleviate concerns of inefficiency, but a significant result would only provide evidence that the magnitude of treatment effect differs in the two patient subgroups. In addition, tests of interaction generally require a larger sample size compared with a study sized for an overall effect (Wang, 2007). In spite of the concerns mentioned above, Sargent et al. (2005) recommend that marker by treatment interaction designs be used when the purpose of a clinical trial is to assess the clinical utility of a single putative predictive biomarker.

When there is a panel of markers to be evaluated, marker by treatment interaction designs can be problematic and marker-based strategy designs may have significant merit (Mandrekar and Sargent, 2009; Sargent et al., 2005). For instance, if there are there are more than two treatments to which patients can be assigned, or if there are multiple outcomes of interest, the marker-based design may be applied to randomize patients based on marker status. Further, the marker-based design also allows investigation of the prognostic value of a biomarker by comparing results among patients who received placebo (or standard treatment) by marker level. The authors do remark that the marker-based design is inefficient due to an overlap arising from the additional randomization of patients in the non-marker-based arm. This results in a significant number of patients receiving the same treatment in both arms, and the overlap increases as the prevalence of the subgroups of interest increases.

As mentioned in Section 1.2.2, when there is convincing evidence that effectiveness of an experimental treatment is confined to a particular subgroup, efficiency gains may be achieved by only recruiting patients of that particular marker status (Maitournam and Simon, 2005; Simon and Maitournam, 2004). This is referred to as *enriching* the study population with potential responders (Temple, 1994, 2005) and the goal is to understand the safety and clinical benefit of the treatment in only a subgroup of patients. A requirement for enrichment is the availability of a robust diagnostic assay, whose validity must be confirmed before enrichment is carried out. The classical enrichment success story concerns the demonstration that the antibody *trastuzumab* (Herceptin) combined with conventional chemotherapy significantly improves disease-free survival outcomes among women with surgically removed HER2 positive breast cancer. However, as mentioned in the introduction and discussed at length in Section 1.2.2, retrospective analyses have suggested that Herceptin may be beneficial for a more broadly defined population. A new confirmative trial may therefore be needed to re-assess the utility of HER2 as a predictive marker, something that was not achieved in the original trials due to the exclusion of all HER2 negative patients. It is recommended by Freidlin et al. (2010) that in most settings, "biomarker-stratified designs be used to obtain a rigorous assessment of biomarker clinical utility."

Screening enrichment designs assume a substantial level of confidence in the accuracy of the biomarker as a classifier, as well as the validity of preliminary data (from Phases I and II). Early stage data is highly variable, and enriching the study population based purely on such results may lead to exclusion of populations that in truth benefit from the experimental treatment. Further, if the diagnostic assay is not sufficiently robust, there could be serious concerns about the validity of conclusions drawn from an enrichment trial. As a compromise, Simon and Wang (2006) suggest sizing a trial to test the overall population at a reduced significance level, and to include a contingency plan to test a (single) prospectively defined subgroup of patients predicted to be particularly responsive, in case the overall trial is negative. Such a design would provide sponsors with an incentive to develop classifiers without running the risk of labeling restrictions if their findings indicate overall efficacy. This could be described as splitting the significance level of the trial, also known as alpha-splitting, or prospective alpha-allocation. Though perhaps not developed with subgroup analysis in mind, several alpha splitting schemes have been proposed, such as a weighted Bonferroni adjustment (Dmitrienko et al., 2010, Ch. 2.6), the prospective alpha allocation scheme (Moyé, 1998, 2000), or the fallback procedure[6] of Wiens (2003) and Wiens and Dmitrienko (2005). These procedures all assume independent endpoints.

More applicable to subgroup analysis however, is the combination of prospective alpha allocation with the assumption that test statistics may be positively correlated. Alosh and Huque (2009) propose a flexible strategy for subgroup testing which, in addition to prospectively splitting the significance level and incorporating correlation, requires a certain consistency of findings between the subgroup and the overall population. In their design, the overall population is tested first at a reduced significance level, and if the overall hypothesis cannot be rejected, but findings are "good enough," testing of a subgroup hypothesis may be conducted at a significance level obtained as a function of subgroup prevalence, measurement precision and desired Type I error. The authors argue that based on the aforementioned consensus that opposite direction treatment effects are unlikely, a weak level of significance should be met in the overall population in order for subgroup analysis to be allowed. Another method based on the same requirement was developed by Song and Chi (2007), in which they use a generalized conditional rejection region to optimize study power.

It may sometimes be the case that an assay or well-defined biomarker is not available at the onset of a confirmatory trial. However, if in truth the patient population is heterogeneous, using a traditional RCT is inefficient and unlikely to yield a positive result. Freidlin and Simon (2005a) have proposed an adaptive design that combines prospective development of a genomics-based biomarker to select sensitive patients with a properly powered test for an overall effect. The procedure is split into two stages (the authors recommend that sample size be evenly split between each stage), and after the first stage a logistic regression

[6] This procedure is described in detail in Chapter 2.

model is used to identify the genes that have most significant treatment-interaction coefficients. Following the second stage, the final analysis consists of an overall test of treatment efficacy using patients accrued over both stages, and a test of patients in the subgroup developed over the first stage, using only patients accrued over the second stage. Both tests are carried out at a reduced significance level to appropriately protect the overall Type I error. To avoid bias, sample size for stage two may not be altered at the interim analysis. The authors conduct simulations which indicate that the development of a genomics-based biomarker to identify subsets of sensitive patients can be incorporated prospectively into a Phase III design without substantially compromising overall power. It should be noted that the biomarker is not used to restrict patient entry in stage two, and hence its development can be conducted at the final analysis (using data assigned to the first stage). The design is hence particularly appropriate for survival endpoints.

Jiang et al. (2007) propose a Phase III design suitable for settings when a prespecified biomarker is measured on a continuous scale. This design combines the test for an overall treatment effect with the establishment and validation of a cut point for the prespecified biomarker. The procedure also provides an estimate of the appropriate biomarker cutoff for sensitive patients. The authors perform simulation studies which indicate that their design retains adequate power to prove overall effect when the treatment is in truth broadly effective. When only a subgroup is responsive, substantial gains in efficiency over traditional randomized designs were reported, in particular when subgroup prevalence is low.

An alternative enrichment design, for use when a reliable assay to select sensitive patients is not available, was reviewed by Freidlin and Simon (2005b). In this Phase II design, known as a *randomized discontinuation design*, patients are initially randomized to experimental treatment or placebo. After a predetermined period of time, early responses are obtained, and responsive patients are allowed to continue while patients for whom the disease progresses are removed from the study. Stable patients are randomized again for another fixed period before being re-evaluated. The authors find that if all patients are sensitive, this design is considerably less efficient than upfront randomization. However, in a heterogeneous response setting, and with a relatively small sensitive population, efficiency gains can be considerable. Some issues regarding interpretation need to be addressed; if the study is positive, generalization can be difficult as the target population might be hard to identify. The authors conclude that randomized discontinuation can be a useful tool when reliable assays do not exist, though its application must be carefully structured to provide sufficient enrichment to sensitive patients. Further, the procedure should only be used when there is sufficient knowledge of disease history and biology to allow meaningful analysis of the results.

In recent years, so-called *adaptive designs* have received increased attention in clinical trials. Phase III trials are both costly and lengthy, and therefore substantial care must be taken in the planning process to ensure that the trial is designed in an efficient and robust way. However, many planning parameters are unknown after Phase II (early estimates of drug efficacy are highly variable), and hence it is appealing to allow for some mid-study changes that are prospectively planned in order to increase the likelihood of a successful trial. Following the seminal paper of Bauer and Köhne (1994), much research in the statistical literature has focused on providing valid and efficient procedures that permit various degrees of adjustments without jeopardizing the integrity of the trial. This can be applied in the context of subgroup analysis when previous data does not clearly suggest upfront use of an enrichment design. Namely, we can prospectively plan to enrich the study population with patients from a predefined subgroup, depending on data observed over early stages of the Phase III trial (Hung et al., 2006).

In principle, adaptive designs can be executed with great flexibility, and adaptations need not be planned in advance. However, as we have discussed above, significant difficulties can arise from cavalier subgroup analysis, and therefore it may be advisable to limit the flexibility to a selected number of options. Wang et al. (2007b) explore an adaptive enrichment approach which allows prospectively planned adaptation to limit a study to a predefined subgroup. The authors assume that a single subgroup has been identified in earlier phases (or earlier trials) in which a therapeutic intervention may be particularly effective. Data is evaluated during an interim analysis at which point the decision is to either (1) complete the trial recruiting patients from the overall population, or (2) exclude patients from the subgroup complement for stage two, and enrich the subgroup population. In this case enrichment implies that all planned sample size for stage two is allocated to the subgroup of interest for the purpose of increasing power. The population adaptation is made only if the therapy appears to be ineffective or unsafe in the complement of the predefined subgroup. Hence the adaptation rule is independent of interim results in the subgroup of interest. Moreover, as the subgroup is enriched due to safety or ethical concerns about the subgroup complement, concerns about selection bias should not invalidate the study. The authors compare their design to that suggested by Freidlin and Simon (2005a) and report substantial power gains for the subgroup. These gains are mostly apparent when only the subgroup is sensitive to the experimental treatment.

In a later paper, Wang et al. (2009) consider a slightly different adaptive enrichment design in the setting of two nested subgroups, S_1 and S_2 (say S_1 contains

S_2). In addition to adaptive enrichment, sample size modification is allowed based on conditional power evaluated at an interim analysis stage; if conditional power for the overall population is high enough, second stage sample size is increased to a prespecified maximum. If overall conditional power is too low at the interim analysis, the study is enriched to S_1. If conditional power for S_1 is also too low, the study is enriched to S_2, and early stopping for futility is allowed if no population looks promising. Simulation studies are conducted to evaluate the performance of this design relative to more traditional designs. Findings indicate that when the proposed biomarker is predictive, power gain (for the subgroups) can be substantial. If the marker is only prognostic, there is not much gain, if any. Interim analysis timing effect on power varies depending on the true underlying treatment effects; when there is a favorable nesting pattern in treatment effects (effect is highest in S_2, then S_1, then overall) an early interim analysis is beneficial. If there is no such nesting pattern, a later interim analysis is more informative. In the numerical study in Chapter 3, we include in our comparisons a similar design as the one proposed by Wang and colleagues. The authors do not advocate use of this design over any other, but rather emphasize that when "subgroup adaptation is a prespecified option, any multiple test procedure that has strong control over experimentwise type I error rate is an appropriate multiple test procedure."

For situations in which frequentist analysis of subgroups is likely to be ineffective, various Bayesian designs have been proposed; see (Berry et al., 2011) for an overview of current clinical practice. Dixon and Simon (1991), Simon et al. (1995) and Simon (2002) develop a Bayesian model for subset analysis in clinical trials with binary covariates. Use of Bayesian statistical methods allows for prior specification of the likelihood of qualitative treatment-by-subset interactions. Then, estimates of subgroup-specific treatment effects are computed as a weighted combination of within-subgroup efficacy estimates and overall efficacy estimates. Their models are based on a key assumption of exchangeability among the treatment-by-subset interactions, restricting their appropriateness to the situation where no *a priori* distinction can be made between subgroups relative to treatment effect. The use of hierarchical Bayes modeling means no specification of subjective priors is required, see for example (Spiegelhalter et al., 2004, p. 277). No formal control of false positive error rates is included, so these methods are more suitable for subgroup screening which might be performed in Phase II.

Wathen et al. (2008) propose a hierarchical Bayesian model for Phase II trials where the exchangeability assumption is not appropriate. This might be the case when previous studies indicate that patients can be divided into groups of "good" prognosis and "bad" prognosis. They propose a parametric likelihood (for binary or survival endpoints) that borrows strength across subgroups, and assume informative priors for the baseline and prognostic parameters (unrelated to treatment-by-subgroup interaction parameters). An algorithm is provided to obtain values for the hyperparameters, and simulation is used to calibrate the model to have "good frequentist characteristics." Their simulation results indicate that when subgroup-specific effects are present, the design substantially outperforms other procedures that do not account for these interactions. However, when no interactions are present and the treatment does not achieve its targeted improvement, the design is less likely to stop the trial for futility.

Due to ethical concerns, adaptive randomization can be used in early stages to decrease the probability that patients are assigned to an ineffective treatment arm. Zhou et al. (2008) propose a Phase II Bayesian adaptive randomization design for patients with advanced stage lung cancer. In this procedure, patients are adaptively randomized to one of four treatments according to continuously updated response rate estimates as data is accumulated throughout the trial. Treatments that appear to perform well for a certain biomarker profile have higher randomization rates, and vice versa. If a treatment performs poorly enough, it may be (temporarily) suspended from randomization, with the possibility to be reinstated later as more data has been examined. A simulation study indicates that in addition to identifying effective treatments with a high probability, more patients are treated with treatments that fit their biomarker profile. The design relies on short assessment times for patient biomarker profiles, as well as short outcome-observation times so that decisions are based on up-to-date data. The design does not take into account assay sensitivity and specificity, and the authors point out that the effect of assay performance on the design requires further investigation.

Adaptive seamless Phase II/III designs, see Bretz et al. (2006) and Schmidli et al. (2006), involve joint planning of both Phase II and III with the intention of cutting delays between the two phases, while allowing for flexible adaptation based both on trial data, and on external factors. We discuss these types of designs in more detail in Chapter 2. Zuber et al. (2006) and Brannath et al. (2009) combine Bayesian decision tools with an adaptive seamless Phase II/III design for a targeted therapy in oncology. They propose a three stage design which is intended to demonstrate efficacy of a targeted agent for a full population, or for a prespecified subgroup. They employ a multiple level testing rule that is unaffected by decision rules used at interim analyses, and they allow the trial to be terminated before the final analysis, either due to early rejection or futility. The decision rule is based on predictive power, which combines the uncertainty about current estimates with the variability inherent in future estimates. Adaptive enrichment is not employed in this design. Their findings indicate that the procedure achieves gains in time and reduction in overall sample size compared to more conventional group sequential procedures, since data from exploratory and confirmatory studies (Phase II and III) are combined and no independent Phase II trials are required to confirm sensitivity of the subgroup in question.

1.2.4 Regulatory Guidance

Regulatory issues and good practice considerations for clinical trials are discussed at length in, for example, (US Food and Drug Administration, 1998) and are not detailed in this thesis. Rather, in this section we summarize some of the guidance principles published by the FDA and EMEA specifically regarding the conduct of subgroup analysis in clinical trials. On reliable conclusions from a subgroup analysis, the European Agency for the Evaluation of Medicinal Products (2002) says the following: *"Reliable conclusions from subgroup analyses generally require pre-specification and appropriate statistical analysis strategies. A license may be restricted if unexplained strong heterogeneity is found in important sub-populations, or if heterogeneity of the treatment effect can reasonably be assumed but cannot be sufficiently evaluated for important subgroups."*

As discussed previously in this chapter, there may be interest in examining the relationship between treatment efficacy and the measurement of various baseline covariates. In many studies, such analyses serve a supportive purpose and results should be interpreted with the appropriate caution. In particular, when subgroup analyses are exploratory, they should be clearly identified as such; commonly these analyses are intended to explore the uniformity of treatment effects overall. In the absence of a pre-specified corresponding null hypothesis and an appropriate analysis strategy, claims of subgroup-specific beneficial effects are very unlikely to be accepted (European Agency for the Evaluation of Medicinal Products, 2002; US Food and Drug Administration, 1998).

In a draft recently made available on their website, the US Food and Drug Administration (2010) provides guidance on the conduct of population adaptation based on treatment-effect estimates. Methods that allow modification of eligibility criteria after an interim analysis can be cautiously applied, when there is suggestive evidence of subgroup-specific effects but said evidence is not strong enough to warrant confidently selecting solely this population(s) for a confirmatory trial. Allowing such types of trials introduces difficulties concerning selection bias, and any prospective study plan should clearly demonstrate control of the Type I error rate for all hypotheses of interest. Caution is advised in planning studies that allow population adaptation to be performed multiple times, as "when multiple revisions to the study population are made it may be challenging to obtain adequate estimates of the treatment effect in the populations of interest, or to interpret to what patient population the results apply."

1.3 Thesis Outline

The remainder of the thesis is outlined as follows. In Chapter 2, we present notation that is common to the rest of the thesis, and define important terms that are used in the presentation of our procedures. We also give an overview of technical results and definitions from the literature.

In Chapter 3, we consider the problem of designing a clinical trial when there is one subgroup of particular interest. A number of designs are proposed and compared in a comprehensive numerical example. We first devise a one-stage *adjusted* *fallback procedure* (Section 3.2.1), which sequentially tests treatment efficacy in all populations of interest, accounting for correlation of the test statistics. This procedure is however not adaptive, and improvements can be obtained by allowing one or more interim analyses. Accordingly, we extend the adjusted fallback procedure, introducing the so-called *fallback enrichment procedure* (Section 3.2.2). This design allows for one interim analysis where the target population may be redefined according to an *a priori* ordering.

We next consider a number of more flexible designs, where testing is conducted by use of a prespecified combination rule. Three different procedures are discussed, where the main differences lie in the interim adaptation rules that are employed. We propose a "hybrid Bayesian" procedure which relies on specifying a prior distribution of treatment effects, and uses a utility function to determine the appropriate course of action at the interim analysis. Additionally, designs based on conditional or predictive power are discussed; similar procedures have been proposed in the literature (Brannath et al., 2009; Wang et al., 2009). To investigate the operating characteristics of the aforementioned designs, we conduct a numerical experiment where statistical power is obtained and compared between all procedures based on subgroup prevalence, interim analysis timing, and interaction of treatment effects across subgroups.

In Chapter 4, we extend the adjusted fallback procedure, the fallback enrichment procedure, and the hybrid Bayesian design to account for an arbitrary number of subgroups. The first two of these are mainly intended to deal with a nested population structure, though they could be applied in other situations as well. The hybrid Bayesian design is developed for disjoint subgroups where no *a priori* ordering is evident. The adjusted fallback procedure is a one-stage design, while the other two include one interim analysis and are hence two-stage procedures.

In Chapter 5, we propose a K-stage adaptive group sequential design that allows adaptation of the target population. As developed, the procedure can take place over any number of stages, and after the first stage non-responsive populations may be discarded. Remaining populations are pooled and subsequent analyses test only one hypothesis, i.e. the null hypothesis of no treatment effect for the pooled population. Either nested or disjoint subgroups can be handled, or some combination thereof. A simple bootstrap algorithm is used to account for selection bias in point estimates. The procedure is illustrated through two worked examples, with applications in development of antidepressants and cancer treatment. Finally, a numerical example is conducted to compare this procedure with the fallback enrichment procedure and the hybrid Bayesian design.

The thesis concludes in Chapter 6 with a brief overview of results, and a discussion of potential future research.

Chapter 2

Problem Setup and Theoretical Background

2.1 Setup

In this section we define the scope of the problem that we consider throughout the thesis. Section 2.1.1 gives common notation and various technical preliminaries, while Section 2.1.2 gives correlation identities for test statistics of interest.

2.1.1 Common Notation

Let Ω_0 denote a complete population of interest for a clinical trial. Suppose that ℓ subgroups have been identified, denoted as $\Omega_j \subsetneq \Omega_0$ for $j = 1, \ldots, \ell$. The subgroups are not necessarily disjoint, and some may completely contain others. Define $\mathcal{P} = \{0, 1, \ldots, \ell\}$ as the index set of all populations of interest. For $\mathcal{S} \subseteq \mathcal{P}$, let

$$\Omega_{\mathcal{S}} = \bigcup_{j \in \mathcal{S}} \Omega_j$$

denote the subpopulation consisting only of $\Omega_j, j \in \mathcal{S}$. When $\mathcal{S}$ is a singleton, we omit the curly braces from our notation. We point out two specific population structures that might arise in practice:

1. **Nested subgroups:** In this case we have $\Omega_j \subsetneq \Omega_{j'}$ for $j > j'$. For example, in early stage testing or retrospective analysis of previous trials, two biomarkers are believed to be prognostic or predictive of treatment effect.

For simplicity, suppose these biomarkers are binary and let B_1 and B_2 denote their respective indicators, such that patients are classified for each biomarker as B_j^+ or B_j^-. Preliminary results may indicate a strong effect for patients in groups B_1^+ (defined as Ω_1) and $B_1^+ \cap B_2^+$ (defined as Ω_2), so $\Omega_0 \supseteq \Omega_1 \supseteq \Omega_2$. This defines a sort of "natural ordering" of subgroups; for economic and ethical reasons, it is desirable to prove an overall effect, provided such an effect exist. As a "fallback" option, clinicians might plan to evaluate efficacy in Ω_1 and Ω_2 (in that order) to salvage the trial.

2. **Disjoint subgroups:** Here, $\Omega_j \cap \Omega_{j'} = \varnothing$ for $j \neq j'$, and $\Omega_0 = \bigcup_{j=1}^{\ell} \Omega_j$. As an example, again suppose there are two dichotomous biomarkers of interest, B_1 and B_2. If preliminary analysis does not indicate that any group is necessarily stronger than another, it may be of interest to assess whether subgroups induced by the two biomarkers respond differently to the experimental treatment. If there is no specific subgroup of *a priori* interest, then this setup is reasonable. Further, the assumption of exchangeability among treatment effect parameters is supported, which allows consideration of various hierarchical Bayesian models.

While more complicated population structures are certainly possible, we focus mainly on the two examples given above.

We wish to examine efficacy of an experimental treatment (E), versus a control or placebo (C). Let W_{kj}^E denote the observed response to treatment for subject k in Ω_j, $j \in \mathcal{P}$. Similarly, let W_{kj}^C denote patient response to placebo. Define μ_j^E (μ_j^C) as the mean experimental (placebo) treatment effect in Ω_j. Define $\theta_j = \mu_j^E - \mu_j^C$ as the mean difference in treatment effect between the experimental and control treatments. Let σ_j^2 be the observation variance in Ω_j, and then set

$$\delta_j = \theta_j / \sigma_j = \frac{\mu_j^E - \mu_j^C}{\sigma_j}, \ j \in \mathcal{P}. \tag{2.1}$$

Thus, δ_j denotes the standardized true mean treatment difference. We are interested in evaluating the family of elementary null hypotheses $H_j : \delta_j \leq 0$ versus $H_j^a : \delta_j > 0$, $j \in \mathcal{P}$. Hence H_j corresponds to the null hypothesis of no treatment benefit in population Ω_j.

Let $f_{ij} \in [0,1]$ denote the prevalence of Ω_j in Ω_i for $i, j \in \mathcal{P}$. If $\Omega_i \cap \Omega_j = \varnothing$, then $f_{ij} = f_{ji} = 0$. Note that $f_{ij} = 1$ if $i = j$ or if $\Omega_i \subset \Omega_j$. For $\mathcal{S}_2 \subseteq \mathcal{S}_1 \subseteq \mathcal{P}$, let $f_{\mathcal{S}_1, \mathcal{S}_2}$ denote the prevalence of $\Omega_{\mathcal{S}_2}$ in $\Omega_{\mathcal{S}_1}$. In terms of Ω_0, this can be written as $f_{0,\mathcal{S}_2} / f_{0,\mathcal{S}_1}$. Let n_0 denote the overall sample size for Ω_0. If subgroup observations are stratified according to their relative size, then the sample size for Ω_j for $j \in \mathcal{P} \backslash \{0\}$ is given as $n_j = f_{0j} n_0$, and n_j^E (n_j^C) denotes the number of subjects assigned to the experimental treatment (placebo). If the procedure takes place over a number of different stages, say K total, then we denote the kth stage sample size for Ω_j as n_{kj}.

For $\mathcal{S} \subseteq \mathcal{P}$, let $\delta_{\mathcal{S}}$ denote the standardized treatment effect in $\Omega_{\mathcal{S}}$. We define $\delta_{\mathcal{S}}$ as a weighted average of δ_j where $\Omega_j \subseteq \Omega_{\mathcal{S}}$ and the individual Ω_j are disjoint. That is,

$$\delta_{\mathcal{S}} = \sum_{j \in \mathcal{S}} f_{\mathcal{S},j} \delta_j. \tag{2.2}$$

Note that even if the subgroups of clinical interest are not disjoint, we can easily define an auxiliary collection of subgroups that are disjoint. This new collection of subgroups is then used to compute $\delta_{\mathcal{S}}$ for a larger population $\Omega_{\mathcal{S}}$. The pooled variance for $\mathcal{S}$, $\sigma_{\mathcal{S}}^2$, is also given as the weighted average of σ_j^2 for $j \in \mathcal{S}$.

For $j \in \mathcal{P}$, we define our tests in terms of the standardized statistic Z_j, or in terms of the efficient score Y_j, with corresponding observed Fisher's information level $\mathcal{I}_j$. If $\hat{\theta}_j$ is the estimated effect size for Ω_j, then $Z_j = \hat{\theta}_j \sqrt{\mathcal{I}_j} = Y_j / \sqrt{\mathcal{I}_j}$ and the test statistics are distributed as

$$\begin{aligned} Z_j &= \hat{\theta}_j \sqrt{\mathcal{I}_j} \overset{a}{\sim} \mathcal{N}\left(\theta_j \sqrt{\mathcal{I}_j}, 1\right), \\ Y_j &= Z_j \sqrt{\mathcal{I}_j} \overset{a}{\sim} \mathcal{N}\left(\theta_j \mathcal{I}_j, \mathcal{I}_j\right), \end{aligned} \tag{2.3}$$

where $\overset{a}{\sim} \mathcal{N}$ indicates that the test statistic is asymptotically normally distributed. Note also that $\hat{\theta}_j \overset{a}{\sim} \mathcal{N}(\theta_j, \mathcal{I}_j^{-1})$. The relationship between the standardized statistics Z_j and the efficient scores Y_j means that tests can be specified in terms of either statistic, as rejection rules can be easily converted. Denote $p_j = 1 - \Phi(Z_j)$ as the p-value for H_j, where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal random variable. Procedures presented in Chapters 3 and 4 are given in terms of the standardized statistics Z_j and information levels $\mathcal{I}_j$. The design proposed in Chapter 5 is outlined in terms of the efficient scores Y_j. Results are therefore applicable to the extent that endpoints of interest are asymptotically normally distributed. E.g. observations can be normal, binary, time-to-event, Poisson etc. (Jennison and Turnbull, 2000, Ch. 3). For ease of exposition, most explicit calculations are carried out assuming that observations are normally distributed.

Normal Observations: Suppose that $W_{kj}^E \sim \mathcal{N}\left(\mu_j^E, \sigma_j^2\right)$ and $W_{kj}^C \sim \mathcal{N}\left(\mu_j^C, \sigma_j^2\right)$. We could assume that measurement precision differs depending on whether placebo or experimental treatment is administered; however, for simplicity we assume a common known variance σ_j^2 for Ω_j. If an equal number of subjects is assigned to placebo and experimental treatment ($n_j/2 = n_j^E = n_j^C$), the standardized statistic Z_j is naturally defined as

$$Z_j = \sqrt{\frac{n_j}{4\sigma_j^2}} \left(\bar{W}_j^E - \bar{W}_j^C\right) = \frac{1}{\sqrt{n_j \sigma_j^2}} \left(\sum_{k=1}^{n_j/2} W_{kj}^E - \sum_{k=1}^{n_j/2} W_{kj}^C \right).$$

The information in this case is $\mathcal{I}_j = \frac{n_j}{4\sigma_j^2}$, where n_j is the total number of observations available in Ω_j for the trial. We can represent $\mathcal{I}_j$ in terms of the overall information, $\mathcal{I}_0$, viz.

$$\mathcal{I}_j = \frac{n_j}{4\sigma_j^2} = \frac{f_{0j}n_0}{4\sigma_0^2}\frac{\sigma_0^2}{\sigma_j^2} = f_{0j}\frac{\sigma_0^2}{\sigma_j^2}\mathcal{I}_0.$$

Hence, if measurement precision is equal across all subgroups, then $\mathcal{I}_j = f_{0j}\mathcal{I}_0$. Given observations Z_j and information levels $\mathcal{I}_j$ from disjoint subgroups Ω_j, $j \in \mathcal{P}$, we may combine these to obtain the test statistic $Z_{\mathcal{S}}$ for $\mathcal{S} \subseteq \mathcal{P}$ as follows:

$$Z_{\mathcal{S}} = \sum_{j\in\mathcal{S}} u_{\mathcal{S},j}\sqrt{f_{\mathcal{S},j}} \cdot Z_j, \text{ where } u_{\mathcal{S},j}^2 = \frac{\sigma_j^2}{\sigma_{\mathcal{S}}^2}.$$

Thus, if we again have equal measurement precision across subgroups, $Z_{\mathcal{S}}$ is distributed as $\mathcal{N}\left(\theta_{\mathcal{S}}\sqrt{\mathcal{I}_{\mathcal{S}}}, 1\right)$, where $\mathcal{I}_{\mathcal{S}} = \sum\limits_{j\in\mathcal{S}} \mathcal{I}_j$ is the information for $\Omega_{\mathcal{S}}$.

If the analysis is conducted over K stages, we obtain stage-wise statistics Z_{kj} for $k = 1, \ldots, K$ and $j \in \mathcal{P}$, along with corresponding incremental information levels Δ_{kj}. Cumulative information levels are $\mathcal{I}_{kj} = \sum\limits_{i=1}^{k} \Delta_{ij}$. The Z_{kj} are distributed as $\mathcal{N}\left(\theta_j\sqrt{\Delta_{kj}}, 1\right)$, and the final statistic for Ω_j is written as

$$Z_j = \sum_{k=1}^{K} w_{kj} Z_{kj}, \text{ where } \sum_{k=1}^{K} w_{kj}^2 = 1. \tag{2.4}$$

To ensure that Z_j is normally distributed under H_j, the combination weights w_{kj} are specified before any data is unblinded. It is natural to define w_{kj}^2 as the fraction of planned information accumulation for Ω_j during stage k. However, definition of the combination weights is deliberately vague at this stage, and is made more precise in subsequent chapters as we present various clinical trial designs.

2.1.2 Test Statistic Correlation

If $\Omega_i \cap \Omega_j \neq \varnothing$, $i, j \in \mathcal{P}$, then test statistics Z_i and Z_j will be correlated. In many of the procedures proposed in this thesis, it is essential to know this correlation, or to obtain an estimate. To this end, let $\bar{W}_{i\backslash j}$ and $\bar{W}_{i\cap j}$ ($\sigma^2_{i\backslash j}$ and $\sigma^2_{i\cap j}$) denote observed sample means (observation variances) in $\Omega_i\backslash\Omega_j$ and $\Omega_i \cap \Omega_j$ respectively. Then $\bar{W}_i = f_{ij}\bar{W}_{i\cap j} + (1 - f_{ij})\bar{W}_{i\backslash j}$, and

$$\begin{aligned}\text{Var}\left(\bar{W}_i\right) &= f_{ij}^2\text{Var}\left(\bar{W}_{i\cap j}\right) + (1-f_{ij})^2\text{Var}\left(\bar{W}_{i\backslash j}\right)\\ &= \frac{f_{ij}\sigma^2_{i\cap j}}{n_i} + \frac{(1-f_{ij})\sigma^2_{i\backslash j}}{n_i}.\end{aligned}$$

Therefore,

$$\begin{aligned}Z_i = \frac{\bar{W}_i}{\sqrt{\text{Var}\left(\bar{W}_i\right)}} &= \sqrt{n_i}\frac{f_{ij}\bar{W}_{i\cap j} + (1-f_{ij})\bar{W}_{i\backslash j}}{\sqrt{f_{ij}\sigma^2_{i\cap j} + (1-f_{ij})\sigma^2_{i\backslash j}}}\\ &= \frac{\sqrt{\frac{n_i f_{ij}}{\sigma^2_{i\cap j}}}\bar{W}_{i\cap j}\sqrt{f_{ij}\sigma^2_{i\cap j}} + \sqrt{\frac{n_i(1-f_{ij})}{\sigma^2_{i\backslash j}}}\bar{W}_{i\backslash j}\sqrt{(1-f_{ij})\sigma^2_{i\backslash j}}}{\sqrt{f_{ij}\sigma^2_{i\cap j} + (1-f_{ij})\sigma^2_{i\backslash j}}}\\ &= \frac{Z_{i\cap j}\sqrt{f_{ij}\sigma^2_{i\cap j}} + Z_{i\backslash j}\sqrt{(1-f_{ij})\sigma^2_{i\backslash j}}}{\sqrt{f_{ij}\sigma^2_{i\cap j} + (1-f_{ij})\sigma^2_{i\backslash j}}}\\ &= Z_{i\cap j}\sqrt{r_{ij}} + Z_{i\backslash j}\sqrt{1-r_{ij}},\end{aligned}$$

where

$$r_{ij} = \frac{f_{ij}\sigma^2_{i\cap j}}{f_{ij}\sigma^2_{i\cap j} + (1-f_{ij})\sigma^2_{i\backslash j}}. \tag{2.5}$$

Hence,

$$\begin{aligned}\text{Corr}(Z_i, Z_j) &= \text{Corr}\left(Z_{i\cap j}\sqrt{r_{ij}} + Z_{i\backslash j}\sqrt{1-r_{ij}}, Z_{i\cap j}\sqrt{r_{ji}} + Z_{j\backslash i}\sqrt{1-r_{ji}}\right)\\ &= \sqrt{r_{ij}r_{ji}},\ \ i, j \in \mathcal{P}.\end{aligned}$$

If the populations are nested, i.e. if $\Omega_j \subseteq \Omega_i$, then $f_{ji} = 1$, and

$$\text{Corr}(Z_i, Z_j) = \sqrt{r_{ij}}. \tag{2.6}$$

Sometimes, the simplifying assumption of equal measurement precision can be made, in which case $r_{ij} = f_{ij}$ and $\text{Corr}(Z_i, Z_j) = \sqrt{f_{ij} f_{ji}}$.

2.2 Technical Results and Preliminaries

In this section, we review several important definitions and theoretical results from the literature. We define family-wise error rate and multiple testing principles. Several common multiplicity adjustment methods are reviewed, as well as the more recently developed literature on flexible adaptive designs.

2.2.1 Multiple Comparison Procedures - Fundamentals

When hypothesis testing is conducted, two types of errors can be committed. Falsely rejecting a true null hypothesis is termed a Type I error, while failing to reject a false null hypothesis is referred to as a Type II error. The main regulatory concern when designing a clinical trial is that false positives occur with sufficiently small probability. If an elementary hypothesis is tested such that the probability of a false positive does not exceed α, we say that the test is carried out at *significance level* α. In a family consisting of multiple hypotheses, this is termed as the *comparison-wise* error rate, see (Hochberg and Tamhane, 1987, p. 7) or (Dmitrienko et al., 2010, p. 37).

Family-wise Error Rate

When considering a family of elementary hypotheses, controlling the comparison-wise error rate at level α is not adequate. Since erroneously rejecting at least one true null hypothesis is considered an incorrect conclusion, we must extend the definition of Type I error to the setting of multiple hypotheses. The definition of *family-wise* error rate is stated in many references; we present the definition given in Dmitrienko et al. (2010, p. 37):

Definition 2.1. *Suppose we have a family of hypotheses $H_j : \delta_j \leq 0$, $j \in \mathcal{H}$, and let $\mathcal{T} \subseteq \mathcal{H}$ denote the index set of true hypotheses. Then the family-wise error rate is defined as*

$$\sup FWER := \max_{\mathcal{T}} \sup_{\{\delta_j(\mathcal{T})\}} \mathbb{P}(\textit{Reject at least one } H_j,\ j \in \mathcal{T}),$$

where the supremum is taken over all δ_j satisfying $\delta_j \leq 0$ for $j \in \mathcal{T}$ and $\delta_j > 0$ for $j \notin \mathcal{T}$.

If $\mathcal{T} = \mathcal{H}$, and FWER $\leq \alpha$, then we say that we have *weak* control of the family-wise error rate. In general, it is not reasonable to expect that all null hypotheses are true, and hence control of FWER at level α is enforced for an arbitrary set $\mathcal{T}$. When this is the case, we say that FWER is controlled *strongly.* We note that naively testing each elementary hypotheses at nominal significance level α does not control FWER weakly nor strongly.

Union-Intersection Testing

Union-intersection testing is a heuristic method introduced by Roy (1953), which involves testing of any hypothesis H that can be stated as an intersection of a family of hypotheses. This is commonly used in pharmaceutical applications where there is an interest to demonstrate that at least one null hypothesis is not true. Again, suppose that we have a family $\mathcal{H}$ of null hypotheses $\{H_j\}_{j\in\mathcal{H}}$, along with corresponding alternate hypotheses $\{H_j^a\}_{j\in\mathcal{H}}$. The union-intersection procedure tests the *global* intersection hypothesis against the union of alternative hypotheses, i.e.

$$H := \bigcap_{j\in\mathcal{H}} H_j \quad \text{vs.} \quad H^a := \bigcup_{j\in\mathcal{H}} H_j^a.$$

In terms of a pharmaceutical objective, the procedure is able to answer the question whether all treatments are ineffective (all populations nonresponsive), or at least one treatment is effective (at least one subgroup is responsive). As a union-intersection test involves multiple inferences, a proper multiplicity adjusted procedure is required to carry out this test.

Closure Principle

Although union-intersection tests allow for multiple inferences on a family of hypotheses, inference on individual elementary hypotheses is not considered. The closure principle introduced by Marcus et al. (1976) is an important part of multiple testing theory; most multiple testing procedures employed in pharmaceutical testing are based on this principle (Dmitrienko et al., 2010, Ch. 2.3.3). The closure principle is outlined as follows:

- Let $\{H_j\}_{j\in\mathcal{H}}$ be a finite family of hypotheses and form the closure of this family by taking all nonempty intersections $H_{\mathcal{P}} = \bigcap_{j\in\mathcal{P}} H_j$ for all $\mathcal{P} \subseteq \mathcal{H}$.
- Define an α level test for each intersection hypothesis $H_{\mathcal{P}}$.

- Reject $H_{\mathcal{P}}$ if and only if all intersection hypotheses $H_{\mathcal{P}'}$ with $\mathcal{P}' \supseteq \mathcal{P}$ are rejected using their respective α level tests.

In particular, an elementary hypothesis H_j is rejected if and only if all intersection hypotheses $H_{\mathcal{P}}$ with $j \in \mathcal{P}$ are rejected with their individual α level tests. It is a well known fact that closed testing procedures (CTPs) control FWER in the strong sense. As it is an important result, we state it as a theorem.

Theorem 2.1. *A closed testing procedure provides strong control of the FWER.*

Proof: Let $\mathcal{T} \subseteq \mathcal{H}$ be the index set corresponding to all true null hypotheses. A Type I error is committed if any null hypothesis H_j with $j \in \mathcal{T}$ is rejected. However, since $H_{\mathcal{T}}$ must be rejected in order for any $H_j, j \in \mathcal{T}$, to be rejected, $\bigcup_{j \in \mathcal{T}}$ [Reject H_j] $\subseteq$ [Reject $H_{\mathcal{T}}$] so

$$\mathbb{P}[\text{Reject any } H_j,\ j \in \mathcal{T}] \leq \mathbb{P}[\text{Reject } H_{\mathcal{T}}] \leq \alpha.$$

This completes the proof. ■

If the number of elementary hypotheses, m say, is large, then the closure principle algorithm can be computationally intensive as it requires the testing of 2^m hypotheses. As a result, "shortcut" procedures have been discussed in the literature, most recently by Bretz et al. (2009), in which the authors outline a simple iterative graphical approach for various procedures based on the closure principle.

2.2.2 Common Multiple Testing Procedures

In what follows, we present some commonly used hypothesis testing procedures that are suitable in multiple testing situations. Throughout, we assume that we have a family of m elementary null hypotheses $\{H_j\}_{j\in\mathcal{H}}$, with associated p-values p_j (or test statistics Z_j). Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ denote the ordered p-values, and let $H_{(1)}, \ldots, H_{(m)}$ denote the corresponding sequence of elementary hypotheses.

Bonferroni-Based Methods

The most common (and perhaps simplest) MCP is likely the Bonferroni procedure which is attributed to Sir Ronald Fisher (Hochberg and Tamhane, 1987, p. 3). In this procedure, individual hypotheses H_j are each tested at level α/m which controls the FWER strongly at level α. This can be a very conservative procedure, and many alternatives have been proposed.

Holm (1979) proposed the so-called *Sequentially Rejective Bonferroni Procedure* which starts by ordering all p-values. Then the hypothesis corresponding to the smallest p-value is tested at level α/m. If rejected, the hypothesis corresponding to the next smallest p-value is tested at level $\alpha/(m-1)$, and so on. If at any point a hypothesis is not rejected, no further testing will take place. Holm also considered a weighted version which assigns positive weights to each hypothesis depending on its importance. If p_i and c_i are the p-value and weight corresponding to H_i, respectively, then define the statistic $S_i = p_i/c_i$. The S_i are then ordered and if $S_{(j)} < \alpha \Big/ \sum\limits_{k=j}^{n} c_{(k)}$ then $H_{(j)}$ is rejected. If at any point a null hypothesis is not rejected, no further testing takes place. This is a *step-down* procedure, in that it starts with the most significant p-value and only continues as long as hypotheses are being rejected.

Simes (1986) proposed a modification of the Bonferroni procedure that orders the p-values and compares the smallest to α/m, the next smallest to $2\alpha/m$ and so on. However, only the global hypothesis, $\bigcap_{i \in \mathcal{H}} H_i$, is considered. An application of the closure principle to Simes' procedure was introduced by Hochberg (1988) to allow inference on individual hypotheses. There, for any $i = m, m-1, \ldots, 1$, if

$$p_{(i)} \leq \alpha/(m-i+1)$$

then this procedure rejects all $H_{(i')}$ for which $i' \leq i$. This is a *step-up* procedure, as testing starts with the largest p-value, and rejection of hypothesis $H_{(i)}$ implies the rejection of all hypotheses with smaller p-values. Since Hochberg's procedure is based on the closure principle, it is uniformly more powerful than Holm's procedure. As it is also very simple to carry out, it is a common method of choice in practice.

Improving on Hochberg's procedure, Hommel (1988) proposed a modified application of the closure principle to Simes' procedure to allow inference on individual hypotheses. The procedure is carried out in m steps as follows: In step $i = 1, \ldots, m$, accept $H_{(m-i+1)}$ and go to the next step if $p_{(m-j+1)} > (i-j+1)\alpha/i$ for $j = 1, \ldots, i$. Else, reject $H_{(m-i+1)}$ and all subsequent hypotheses. Hommel's procedure can also be extended to the case of unequally weighted hypotheses. This procedure, while potentially harder to explain to a non-statistician, is more powerful than Hochberg's version of Simes' test (Hommel, 1989).

Example 2.1. *To illustrate the MCPs discussed above, we consider a simple example in which three ($m = 3$) hypotheses H_1, H_2 and H_3 are to be tested. Suppose we have obtained p-values $p_1 = 0.012$, $p_2 = 0.0251$ and $p_3 = 0.0087$, and $\alpha = 0.025$. Ordering these we have $p_{(1)} = p_3$, $p_{(2)} = p_1$ and $p_{(3)} = p_2$. Holm's procedure starts by examining the smallest p-value at level α/m, and since $p_{(1)} = 0.0087 > \alpha/3$, we retain $H_{(1)} = H_3$ and all subsequent hypotheses. Hochberg's procedure starts by accepting $H_{(3)} = H_2$ because $p_{(3)} > \alpha$. However, $p_{(2)} = 0.012 < \alpha/2 = 0.0125$*

so both $H_{(2)} = H_1$ and H_3 are rejected. Similarly, $H_{(3)} = H_2$ is accepted with Hommel's procedure, because $p_{(3)} > \alpha$. However, since $p_{(2)} = 0.012 < \alpha$ (and also smaller than $\alpha/2$), Hommel's procedure rejects both $H_{(2)} = H_1$ and H_3.

Note that Holm's procedure stops testing as soon as we fail to reject a null hypothesis, while both Hommel's and Hochberg's procedures continue testing until one rejection occurs (at which point all further hypotheses are also rejected). For a slightly different example, suppose we had obtained p-values $p_1 = 0.013$, $p_2 = 0.022$, and $p_3 = 0.006$. In this case, Holm's procedure rejects H_3 because $p_{(1)} = p_3 = 0.006 < \alpha/3$. However, H_1 (and hence H_2) is retained as $p_{(2)} = p_1 = 0.013 > \alpha/2$. Both the Hommel and the Hochberg procedure will in this case reject all hypotheses, as $p_{(3)} = p_2 < \alpha$.

Fixed Sequence Procedure

The fixed sequence procedure of Maurer et al. (1995) and Westfall and Krishen (2001) is a straightforward way to test a family of hypotheses. Prior to unblinding data, hypotheses must be ordered according to their clinical importance. That is, H_1 is of most relevance, then H_2 and so on. The procedure rejects H_i, $i = 1, \ldots, m$, if and only if $p_i \leq \alpha$, and all hypotheses prior to H_i were also rejected. Hence, as soon as one hypothesis is retained, all subsequent hypotheses are also retained and the procedure stops. It is easy to see that the fixed sequence procedure controls FWER at level α in the strong sense.

A nice feature of this procedure is that any tested hypotheses is evaluated at the maximum level, i.e. α. On the other hand, as soon as we fail to reject one hypothesis, we are forced to abandon further testing. This can be particularly troublesome if the prespecified ordering is unsuitable with respect to actual parameter values. In such cases, the procedure can be severely underpowered if early hypotheses are unlikely to be rejected.

Example 2.2. *Continuing with Example 2.1, the fixed sequence procedure will test H_1, H_2 and H_3 in that order (suppose this was indeed our pre-specified ordering), and each hypothesis is tested at nominal significance level $\alpha = 0.025$. In this case, H_1 is rejected as $p_1 = 0.012 < \alpha$, but since $p_2 > \alpha$, H_2 is retained. At this point, no further testing may take place, so H_3 is also retained.*

Fallback Procedure

Wiens (2003) proposed the so-called fallback procedure as an alternative to the fixed sequence procedure. As before, hypotheses are ordered beforehand according to clinical importance. The method also requires the prespecification of "local" significance-levels α_j for $j \in \mathcal{H}$, such that $\sum\limits_{j \in \mathcal{H}} \alpha_j = \alpha$. The procedure is then given as follows:

- Test H_1 at level α_1, i.e. rejecting if $p_1 \leq \alpha_1$.
- For $j = 2, \ldots, m$, test H_j at level $\alpha'_j = \sum\limits_{i=k+1}^{j} \alpha_i$, where H_k is the last accepted hypothesis before H_j. If all hypotheses were rejected before H_j, then $k = 0$.

The idea of the fallback procedure is that Type I error only accumulates as hypotheses are rejected. Note that, by setting $\alpha_1 = \alpha$ and $\alpha_2 = \cdots = \alpha_m = 0$, the fallback procedure reduces to the fixed sequence procedure. In a subsequent paper, Wiens and Dmitrienko (2005) proved that the procedure is equivalent to a CTP and hence it controls FWER strongly. They also compared the fallback procedure to Hommel's closure of the weighted Simes test (Hommel, 1988), and showed that neither procedure is more powerful than the other. The fallback procedure with larger early weights (larger α_j early) has an advantage over the weighted Hommel procedure. When weights are all similar, the weighted Hommel procedure may have an advantage.

Example 2.3. *To illustrate the fallback procedure, we continue with Examples 2.1 and 2.2. Suppose we have chosen local significance levels* $\alpha_1 = 0.010$*,* $\alpha_2 = 0.010$ *and* $\alpha_3 = 0.005$*. With the p-values* $p_1 = 0.012$*,* $p_2 = 0.0251$ *and* $p_3 = 0.0087$*, this is a particularly unfortunate weighting scheme as no hypotheses are rejected. To see this, note that* $p_1 > \alpha_1$ *so* H_1 *is retained and* H_2 *is tested at level* α_2*. We have* $p_2 > \alpha_2$*, so* H_2 *is retained and* H_3 *must be tested at level* α_3*. Since* $p_3 > \alpha_3$*,* H_3 *is also retained. If, on the other hand, we had chosen* $\alpha_1 = 0.015$*,* $\alpha_2 = 0$ *(i.e.* H_2 *can only be tested if* H_1 *was rejected) and* $\alpha_3 = 0.010$*, then* H_1 *would be rejected as* $p_1 < \alpha_1$*. However,* $p_2 > \alpha_1 + \alpha_2$ *so* H_2 *is still retained. Finally,* H_3 *is rejected as* $p_3 < \alpha_3$*.*

Additional Comments

The procedures described so far are all defined without any distributional assumptions. The appeal of such procedures is obviously their generality, as they are applicable in a variety of settings. For an overview of parametric multiple testing procedures, see (Dmitrienko et al., 2010, Ch. 2.7).

When test statistics are correlated – as is commonly the case in subgroup analysis – distribution-free testing procedures can be quite conservative. As detailed in Chapter 1, Section 1.2.3, Alosh and Huque (2009) and Song and Chi (2007), propose testing procedures for a single-subgroup setting where correlation is accounted for. Huque and Alosh (2008) also propose the “flexible fixed sequence” procedure (FFS) which relies on exploiting endpoint correlation to improve on the original fallback procedure. In Chapter 3, Section 3.2.1, and Chapter 4, Section 4.1, we introduce the adjusted fallback procedure (AFP) which is developed with a similar objective in mind. The two procedures are different approaches to the same idea, and the main insight obtained here is that, just as the fallback procedure, the AFP is equivalent to a CTP.

2.2.3 Adaptive Designs

Adaptive trial designs have attracted substantial interest over the last decade and a half. They offer the option to make significant changes to a trial protocol during an experiment, either depending on data already observed, or due to other external factors. For example, some doses may be dropped, sample-size can be modified, and certain patient populations may be excluded. Some changes involve a change to the hypotheses of interest, and as a result the final analysis may differ significantly from what was originally planned in the study protocol. Following the work of Bauer and Köhne (1994), Proschan and Hunsberger (1995), and Bauer and Kieser (1999), several statistical procedures have been developed that allow various mid-trial modifications motivated by unblinded data, while ostensibly preserving the integrity of the trial. If applied recklessly, however, adaptations and subsequent analysis can lead to wildly misleading conclusions, inflated Type I error rates, and the introduction of various biases (such as selection bias).

As many important variables particular to a clinical trial are often unknown at the outset, the appeal of adaptive designs is understandable. However, the flexibility involved comes at a price, such as reduction in statistical efficiency, difficulties in interpretation, and concerns about basing data-driven trial modifications on sometimes highly unreliable interim estimates. For instance, designs that allow sample size adjustment based on interim data may end up basing the final analysis on statistics that are not sufficient, and are hence inefficient. The inefficiency of such trials has been discussed at great length; see for example Tsiatis and Mehta (2003), Jennison and Turnbull (2003, 2006b) and Fleming (2006). Burman and Sonesson (2006) showed extreme examples where modifications can lead to rejection of the null hypothesis when the overall effect estimate (in a lower one-sided hypothesis setting) is in truth negative. Concerns about inefficiency and anomalous results has hence led to some scepticism about the legitimacy of data-driven adaptations. Nevertheless, allowing for "sensible" adaptations, preferably listed in the trial protocol and clearly explained in the study results, is still an intriguing option. In an important discussion paper, Gallo (2006) argues that review of accumulating data and subsequent interim decisions should be carried out by a "data monitoring committee," independent (or nearly independent) of the trial sponsor. Gallo further argues that the official study protocol should not list adaptation plans in detail, as this may allow trial sponsors or other observers to infer likely interim estimates from adaptations made mid-trial. Rather, full detail of the adaption protocol should be contained in a separate document, only to be disseminated to the data monitoring committee. Following these recommendations may alleviate some of the concerns regarding the validity of adaptive clinical trials.

To further expedite the drug testing process, so-called confirmatory adaptive seamless Phase II/III designs (see Chapter 1, Section 1.2.3) have been discussed by Bretz et al. (2006); Schmidli et al. (2006), and by Maca et al. (2006). Adaptive seamless designs (ASD) are intended to combine Phases II and III into a larger confirmatory trial, where various data-driven design modifications may be necessary at the end of Phase II. The final analysis combines data from both stages, where proper statistical inference methods are applied to prevent bias and Type I error inflation. Some of the main goals of an ASD, summarized in (Bretz et al., 2006, p. 624), are:

1. Reduce the time to decide on, plan and implement the next clinical phase (reduction of the "white-space" between the two studies);
2. Save costs through the combination of evidences across two studies and thus the need for fewer patients (or, equivalently, increase the information value and the reliability of decision making while maintaining the same sample sizes);
3. Get long-term safety data earlier as a direct consequence of following up the Phase II patients.

The authors show that there are gains in efficiency in a combined Phase II/III trial, relative to the traditional split of separate Phase II and Phase III trials. They caution that, though arbitrary modifications to the trial are possible without inflating Type I error rates, only modest adaptations should be employed in order for the trial conclusions to be credible. In their discussions of ASDs, Jennison and Turnbull (2006a, 2007) argue that conventional group sequential methods should not be overlooked, and point to the work of Stallard and Todd (2003). Therein, a group sequential design is proposed which consists of $K + 1$ stages; the first stage is a treatment-selection phase (Phase II) where the best treatment is selected to proceed, while the remaining K stages employ traditional group sequential stopping rules, testing only the selected treatment.

2.2.4 Adaptive Testing Procedures

The spirit of "truly" adaptive designs is to allow mid-trial modifications that are not specified beforehand, and Type I error rates are protected by using specific combination methods to obtain final test statistics. In this section, we describe some common combination testing procedures that have been proposed in the literature. Let $c \in \mathbb{R}$ be given, and suppose we have p-values $p_1, \ldots, p_K$ for a hypothesis H, collected over K stages. Using a combination function $C(p_1, \ldots, p_K)$ and the decision rule to reject H if $C(p_1, \ldots, p_K) > c$, these procedures will control Type I error regardless of the adaptation rule. A key assumption about the p-values obtained throughout a trial is that they are *p-clud* (Brannath et al., 2002):

Definition 2.2. *The p-values $p_1, \ldots, p_K$ are p-clud if, under H, the distribution of p_1, and the conditional distribution of p_k given previously observed p-values, is stochastically larger than or equal to the uniform distribution on [0,1].*

The authors show that if independent sample units are collected at each stage and tests are applied that control false positive error rate at a prespecified level α, then the obtained p-values will be p-clud regardless of adaptations performed at the interim analyses. Although the assumption that p-values are p-clud is sufficient, a more common assumption (and slightly stronger) is that the p-values are conditionally independent and uniformly distributed on [0,1], as long as H is true. Hence, under H, p-values are also unconditionally independent and uniformly distributed. Below, we describe some commonly used combination functions.

The first approach, attributed to Sir Ronald Fisher, and preferred by Bauer and Köhne (1994), is the "inverse χ^2" method, which rejects H at level α if

$$C(p_1, \ldots, p_k) := -2 \log \left(\prod_{i=1}^{k} p_i \right) > \chi^2_{2k,\alpha},$$

where $\chi^2_{2k,\alpha}$ is the upper α percentile of the χ^2_{2k} distribution. This follows because under the null hypothesis, and conditional on adaptations up to $k-1$, P_k is distributed as $\mathcal{U}(0,1)$. Then, $-2\log(P_k) \sim \text{Exp}(1) \sim \chi^2_2$.

The second approach is the "weighted inverse normal" combination method (Cui et al., 1999; Lehmacher and Wassmer, 1999), which rejects H at level α if $C(p_1,\ldots,p_K) > z_\alpha$, where z_α is the upper α percentile of the standard normal distribution,

$$C(p_1,\ldots,p_k) := \left(\sum_{i=1}^{k} w_i^2\right)^{-1/2} \sum_{i=1}^{k} w_i \Phi^{-1}(1-p_i),$$

and $w_1,\ldots,w_K$ are prespecified combination weights for which $\sum\limits_{k=1}^{K} w_k^2 = 1$. We note that combined test statistics defined in Equation (2.4) are equivalent to the weighted inverse normal combination function defined above. As mentioned in Section 2.1, a natural definition of w_k is the fraction of information accumulated in statistic Z_k (hence p_k), so $w_k = \sqrt{\Delta_k/\mathcal{I}_K}$, where $\mathcal{I}_K$ is the total planned information.

The third approach is based on the *conditional error principle.* In a two-stage setting, we define the conditional Type I error probability as $A(p_1) = \mathbb{P}(\text{Reject } H|p_1)$. First discussed by Proschan and Hunsberger (1995), they let $A(p_1)$ be a monotonic non-decreasing function of first stage results, p_1, such that

$$\alpha_1^* + \int_{\alpha_1^*}^{\beta_1^*} A(p_1)dp_1 = \alpha.$$

Here, H is rejected if $p_1 \leq \alpha_1^*$, or if $\alpha_1^* < p_1 \leq \beta_1^*$ and $p_2 \leq A(p_1)$. Müller and Schäfer (2001) proposed that the conditional error function should be defined in terms of a preplanned hypothesis test. That is, suppose that $\varphi = I(p \leq \alpha)$ is the indicator function for our preplanned test of H. This test is to be carried out at level α after all observations are taken. After performing the desired adaptations at an interim analysis (where we have observed p_1), we redefine the level of the original test, setting it to $\mathbb{E}_H[\varphi|p_1]$. If no adaptations are performed, the originally planned test φ is used at the final analysis.

Chapter 3

Procedures For One Subgroup

In this chapter, we consider the problem of designing a clinical trial when there is one subgroup of interest. We propose a number of novel procedures, and compare these to methods already existing in the literature. Further, we analyze operating characteristics of a targeted trial with respect to various factors, such as subgroup prevalence, interim analysis timing, and treatment effect size in the subgroup and its complement.

3.1 Setup

Let Ω_0 denote the complete population of interest. A targeted population, $\Omega_1 \subsetneq \Omega_0$, has been identified, and we wish to examine efficacy of an experimental treatment, E, versus a control (or best known treatment), C. This will be investigated in both Ω_0 and Ω_1. Let Ω_2 denote the complement of Ω_1, i.e. $\Omega_2 = \Omega_0 \backslash \Omega_1$.

We are concerned with gaining knowledge about the standardized effect sizes $\delta_j = \theta_j/\sigma$, $j = 0, 1$, defined in Equation (2.1). The hypotheses of interest are:

- $H_0 : \delta_0 \leq 0$, the null hypothesis of no treatment effect in the full population, versus $H_0^a : \delta_0 > 0$.
- $H_1 : \delta_1 \leq 0$, the null hypothesis of no treatment effect in the targeted population, versus $H_1^a : \delta_1 > 0$.

These are tested at a prespecified significance level such that FWER is controlled at level α in the strong sense. The hypotheses can of course also be stated in terms of θ_0 and θ_1.

The overall sample size for Ω_0 is denoted as n_0, and as introduced in Chapter 2, f_{0j} denotes the prevalence of Ω_j in Ω_0, where $0 < f_{0j} < 1$, $f_{00} = 1$. Then the sample size for Ω_j is given as $n_j = f_{0j} n_0$, and the stage-wise sample size for Ω_j during stage k is denoted by n_{kj}, $k = 1, 2$ and $j = 0, 1, 2$. We assume that observation variance is equal across all populations. The mean treatment effect for Ω_0 is then given as

$$\delta_0 = f_{01}\delta_1 + (1 - f_{01})\delta_2.$$

We assume that responses are normally distributed, and define our tests in terms of the standardized statistics Z_j which are distributed as $\mathcal{N}\left(\theta_j\sqrt{\mathcal{I}_j}, 1\right)$, where $\mathcal{I}_j = \frac{n_j}{4\sigma^2}$ denotes the observed information. The test statistics Z_0, Z_1 and Z_2 are hence jointly distributed as

$$\begin{pmatrix} Z_0 \\ Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left\{\begin{pmatrix} \theta_0\sqrt{\mathcal{I}_0} \\ \theta_1\sqrt{f_{01}\mathcal{I}_0} \\ \theta_2\sqrt{f_{02}\mathcal{I}_0} \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{f_{01}} & \sqrt{f_{02}} \\ \sqrt{f_{01}} & 1 & 0 \\ \sqrt{f_{02}} & 0 & 1 \end{pmatrix}\right\}$$

Given Z_1 and Z_2, we can write $Z_0 = \sqrt{f_{01}}Z_1 + \sqrt{(1 - f_{01})}Z_2$. We also define p-values $p_j = 1 - \Phi(Z_j)$ where Φ is the cumulative distribution function of a standard normal random variable. For $\eta \in [0, 1]$, define C_η as the upper η-percentile of the standard normal distribution. That is, $\eta = \mathbb{P}[\mathcal{N}(0, 1) > C_\eta]$.

For trials that are divided into two separate stages, let $t \in (0, 1)$ denote the timing of the interim analysis. If $\mathcal{I}_j$ denotes the cumulative observed information for Ω_j, we have interim information levels $\Delta_{1j} = t\mathcal{I}_j$ for the first stage, $\Delta_{2j} = (1 - t)\mathcal{I}_j$ for the second stage, and $\mathcal{I}_j = \Delta_{1j} + \Delta_{2j}$. Stage-wise test statistics Z_{kj}, for stage k and population Ω_j, $k = 1, 2$ and $j = 0, 1, 2$, are approximately distributed as $\mathcal{N}(\theta_j\sqrt{\mathcal{I}_{kj}}, 1)$, and final statistics are written as

$$Z_j = \sum_{k=1}^{2} w_{kj} Z_{kj}, \text{ where } w_{1j}^2 + w_{2j}^2 = 1.$$

The weights w_{kj} are assigned before the trial, typically chosen such that w_{1j}^2 is the interim analysis time t. Note that

$$w_{1j} = \sqrt{t} = \sqrt{\Delta_{1j}/\mathcal{I}_j} \text{ and } w_{2j} = \sqrt{1-t} = \sqrt{\Delta_{2j}/\mathcal{I}_j}, \tag{3.1}$$

so the value of w_{kj}^2 represents the fraction of information accumulated during stage k. If enrichment is envisaged at the interim analysis, more observations are taken from Ω_1 during the second stage than was originally planned. Hence it may seem reasonable to use different weights that are adjusted for the new (and different) information levels. Wang et al. (2009) refer to this as *empirical data weights*. In this case the weights will depend on the adaptation decision made at interim, so the final test statistic Z_1 is not necessarily a standard normal random variable under H_1. As a result, the Type I error probability may be inflated, and the FDA requires strong control FWER to be enforced strictly (US Food and Drug Administration, 1998). However, our adaptive procedures only allow very limited adaptations at interim, and hence the inflation may be relatively small. In Section 3.3, we examine the effects of using empirical data weights on Type I error probabilities. Adjusted combination weights for Ω_1 are given as

$$\widetilde{w}_{11} = \sqrt{\tilde{t}} = \sqrt{\frac{tf_{01}}{tf_{01} + (1-t)}} \tag{3.2}$$

and

$$\widetilde{w}_{21} = \sqrt{1-\tilde{t}} = \sqrt{\frac{1-t}{tf_{01} + (1-t)}}. \tag{3.3}$$

3.2 Procedures

In this section we detail a number of clinical trial designs that can be used in the setup discussed above. Some of these procedures, e.g. AFP (see Section 3.2.1) and FE (see Section 3.2.2), rely on the availability of the joint distribution of test statistics to compute exact boundaries of the acceptance region. Other methods, i.e. CP, HPP and HUT (see Sections 3.2.3 and 3.2.4), employ flexible adaptation rules (Bretz et al., 2006; Schmidli et al., 2006), and must hence use combination techniques and the closure principle (see Section 2.2) at the final analysis stage. Here, we use Simes' method to test intersection hypotheses, the weighted inverse normal combination rule to obtain final stage test statistics, and apply the closure principle.

3.2.1 Adjusted Fallback Procedure (One Stage)

We improve the fallback procedure introduced by Wiens (2003) by accounting for correlation between test statistics. Let α_0, α_1 be specified beforehand, such that $\alpha_0 + \alpha_1 = \alpha$, the desired FWER. The α_i represent *local* significance levels that will be used to test H_0 and H_1. Assume also that we have an adjusted local significance level $\tilde{\alpha}_1$ for H_1. Below, we explain how to obtain this value. The adjusted fallback procedure (AFP) is defined as follows:

- *Step 1:* Test H_0 at level α_0, rejecting if $p_0 \leq \alpha_0$.
- *Step 2:*
  - If H_0 was rejected in Step 1, test H_1 at level $\alpha_0 + \alpha_1 = \alpha$, rejecting if $p_1 \leq \alpha$.

– If H_0 was not rejected, test H_1 at level $\tilde{\alpha}_1$, rejecting if $p_1 \leq \tilde{\alpha}_1$.

The AFP improves on the original fallback procedure by using a level $\tilde{\alpha}_1 \in [\alpha_1, \alpha]$ to test H_1 in the event that H_0 was not rejected. When there is only one subgroup of interest, AFP is a special case of a procedure proposed by Alosh and Huque (2009). However, in Chapter 4, we present the AFP for the case where there are multiple subgroups under consideration.

Obtaining $\tilde{\alpha}_1$

Define $\xi \in \mathbb{R}$, satisfying

$$\alpha - \alpha_0 = \mathbb{P}[Z_0 < C_{\alpha_0}, Z_1 \geq \xi \mid H_0 \cap H_1]. \tag{3.4}$$

Set $x = \mathbb{P}[Z_1 \geq \xi | H_1]$ and finally $\tilde{\alpha}_1 = \min(x, \alpha)$ so $\tilde{\alpha}_1 \leq \alpha$. In what follows, let f_{Z_0,Z_1} denote the bivariate density function for Z_0 and Z_1, and f_Z the marginal density for Z. Similarly, let F_{Z_0,Z_1} and F_Z denote the cumulative distribution functions for the bivariate density and the marginal density, respectively. Then,

$$\begin{aligned}
\alpha - \alpha_0 &= \int_{-\infty}^{C_{\alpha_0}} \int_{\xi}^{\infty} f_{Z_0,Z_1}(z_0, z_1; \rho) dz_0 \ dz_1 \\
&= \int_{-\infty}^{C_{\alpha_0}} f_{Z_0}(z_0) dz_0 - \int_{-\infty}^{C_{\alpha_0}} \int_{-\infty}^{\xi} f_{Z_0,Z_1}(z_0, z_1; \rho) dz_0 \ dz_1 \\
&= F_{Z_0}(C_{\alpha_0}) - F_{Z_0,Z_1}(C_{\alpha_0}, \xi; \rho) \\
&= 1 - \alpha_0 - F_{Z_0,Z_1}(C_{\alpha_0}, \xi; \rho)
\end{aligned}$$

Thus ξ can be obtained via a numerical search, such that it satisfies $1 - \alpha = F_{Z_0,Z_1}(C_{\alpha_0}, \xi; \rho)$. The value for $\tilde{\alpha}_1$ will depend on the value chosen for α_0, as well as the correlation $\rho = \sqrt{f_{01}}$ between Z_0 and Z_1.

Control of the FWER

We have three null configurations: $(\delta_0 = 0, \delta_1 = 0), (\delta_0 \neq 0, \delta_1 = 0)$ and $(\delta_0 = 0, \delta_1 \neq 0)$. Define the events

$$\begin{aligned} R_0 &:= [Z_0 \geq C_{\alpha_0}], \\ R_{10} &:= [Z_1 \geq C_\alpha \mid Z_0 \geq C_{\alpha_0}], \\ R_{11} &:= [Z_1 \geq C_{\tilde{\alpha}_1} \mid Z_0 < C_{\alpha_0}]. \end{aligned}$$

A Type I error can occur in one of the three following ways:

$$\begin{aligned} &A_1 := R_0 \cap R_{10} \text{ (both are true and both are rejected)} \\ &A_2 := R_0 \cap \bar{R}_{10} \text{ (H_0 true and rejected, H_1 not rejected)} \\ &A_3 := \bar{R}_0 \cap R_{11} \text{ (H_0 not rejected but H_1 is true and rejected)} \end{aligned}$$

Note that $A_1 \cup A_2 = R_0$. Under $(\delta_0 = 0, \delta_1 = 0)$, we have

$$\begin{aligned} \mathbb{P}[\text{At least one Type I error}] &= \mathbb{P}[A_1 \cup A_2 \cup A_3 | H_0 \cap H_1] \\ &= \mathbb{P}[R_0 | H_0] + \mathbb{P}[\bar{R}_0 \cap R_{11} | H_0 \cap H_1] \\ &\leq \alpha_0 + \mathbb{P}[Z_0 < C_{\alpha_0}, Z_1 \geq \xi \mid H_0 \cap H_1], \quad \text{since } C_{\tilde{\alpha}_1} \geq \xi \\ &= \alpha_0 + (\alpha - \alpha_0) = \alpha, \quad \xi \text{ satisfies (3.4).} \end{aligned}$$

Next note that $\tilde{\alpha}_2 \leq \alpha$ and hence $[Z_1 \geq C_{\tilde{\alpha}_2}] \subseteq [Z_1 \geq C_\alpha]$. Therefore, under $(\delta_0 \neq 0, \delta_1 = 0)$,

$$\begin{aligned} \mathbb{P}[\text{At least one Type I error}] &= \mathbb{P}[A_1 \cup A_3 \mid H_1] \\ &= \mathbb{P}\left[(R_0 \cap R_{10}) \cup (\bar{R}_0 \cap R_{11}) \mid H_1\right] \\ &\leq \mathbb{P}[R_{10} \cup R_{11} \mid H_1] = \mathbb{P}[Z_1 \geq C_\alpha \text{ or } Z_1 \geq C_{\tilde{\alpha}_2} \mid H_1] \\ &= \mathbb{P}[Z_1 \geq C_\alpha \mid H_1] = \alpha. \end{aligned}$$

Finally, under $(\delta_0 = 0, \delta_1 \neq 0)$, the event $A_1 \cup A_2 = R_0$ leads to a Type I error, and the probability of this is clearly $\alpha_0 \leq \alpha$. Hence the above procedure controls the FWER at level α in the strong sense. Note that the construction of the AFP does not require that σ_j are equal for all j.

Rejection Probabilities

We proceed assuming that the overall study for Ω_0 (carried out at level α_0) is powered at $1-\beta$ for detecting the effect $\delta_0 = \frac{\theta_0}{\sigma}$, where σ^2 is the pooled variance for the whole population. Since half of our observations are taken from the control group, i.e. $n_0/2 = n_0^E = n_0^C$, we have

$$\delta_0 = \sqrt{\frac{4}{n_0}}(C_{1-\alpha_0} + C_{1-\beta}).$$

Denote $\delta_1 = \eta\delta_0$ where $\eta \geq 0$. Non-centrality parameters for Z_0 and Z_1 are as follows:

$$\lambda_0 = \mathbb{E}(Z_0 \mid \delta_0 = \delta) = \theta_0\sqrt{\mathcal{I}_0} = \sqrt{\frac{n_0}{4}}\delta_0 = (C_{1-\alpha_0} + C_{1-\beta}), \text{ and}$$
$$\lambda_1 = \mathbb{E}(Z_1) = \theta_1\sqrt{\mathcal{I}_1} = \sqrt{\frac{n_1}{4}}\delta_1 = \eta\sqrt{f_{01}}\sqrt{\frac{n_0}{4}}\delta = \eta\sqrt{f_{01}}\lambda_0.$$

Therefore we can compute λ_0 by specifying α_0 and β, and λ_1 by specifying f_{01} and η. For short-hand notation, set $a_0 = C_{\alpha_0} - \lambda_0$, $a = C_\alpha - \lambda_1$ and $a_1 = C_{\tilde{\alpha}_1} - \lambda_1$. The power of the subgroup study is then

$$\begin{aligned}
&\text{Power}(\eta, f_{01}, \alpha_0, \delta_0) = \mathbb{P}[A_1 \cup A_3 \mid \eta, \delta_0] \\
&= \mathbb{P}[Z_0 > C_{\alpha_0}, Z_1 > C_\alpha \mid \eta, \delta_0] + \mathbb{P}[Z_0 \leq C_{\alpha_0}, Z_1 > C_{\tilde{\alpha}_1} \mid \eta, \delta_0] \\
&= 1 - \Phi(a_0) - \Phi(a) + \Phi(a_0, a; \rho) + \Phi(a_0) - \Phi(a_0, a_1; \rho).
\end{aligned}$$

We comment on the performance of the AFP in Section 3.3. In Chapter 4, we also generalize the method to allow consideration of an arbitrary number of subgroups.

3.2.2 Adjusted Fallback Procedure with Enrichment (Two Stages)

We can improve on the AFP by combining the fallback approach with adaptive enrichment. The resulting procedure, abbreviated as FE, is outlined here for the case of one subgroup, and generalized to consider multiple subgroups in Section 4.2. We extend the AFP to allow an interim analysis, at which point three decisions are entertained. First, we may decide to stop for futility and accept both hypotheses. Second, we can proceed to the second stage sampling from the complete population just as planned. Third, we can enrich and sample only from Ω_1 during the second stage. As before, the user specifies local significance levels α_0 and α_1 that sum to α. Further, $\gamma_0, \gamma_1 \in [0, 1]$ are specified for use in the interim analysis. In the case that enrichment is carried out, we define the final test statistic for H_1 as

$$\widetilde{Z}_1 = \widetilde{w}_{11} Z_{11} + \widetilde{w}_{21} \widetilde{Z}_{21},$$

where $\widetilde{w}_{kj}$ are given in equations (3.2) and (3.3), and $\widetilde{Z}_{21} \sim \mathcal{N}(\theta_1 \sqrt{\mathcal{I}_{20}}, 1)$. The procedure is now given:

I.1 If $Z_{10} \geq C_{\gamma_0}$ then go to II.1.

I.2 If $Z_{10} < C_{\gamma_0}$ and $Z_{11} \geq C_{\gamma_1}$ then go to II.2.

I.3 If $Z_{10} < C_{\gamma_0}$ and $Z_{11} < C_{\gamma_1}$ then terminate the trial for futility.

II.1 Take second stage sample from the full population, Ω_0, and carry out the fallback procedure (Wiens, 2003) on H_0 and H_1 using local significance levels α_0 and α_1.

II.2 Take second stage sample from Ω_1 only. If $\widetilde{Z}_1 \geq C_{\tilde{\alpha}_1}$ then reject H_1. H_0 is not tested.

Control of the FWER

The procedure is constructed such that FWER $\leq \alpha$ under any null configuration. Analogous to Section 3.2.1, let $F_X(\cdot)$ denote the cumulative distribution function of a random variable X, $F_{X,Y}(\cdot,\cdot;\psi)$ the joint bivariate CDF of X and Y with dependence parameter ψ, and $F_{X,Y,Z}(\cdot,\cdot,\cdot;\Sigma)$ the joint three-dimensional CDF of X, Y and Z with correlation matrix Σ. Define the following events:

$$\begin{aligned}
S_0 &:= [Z_{10} \geq C_{\gamma_0}],\\
S_1 &:= [Z_{11} \geq C_{\gamma_1}],\\
R_0 &:= [Z_0 \geq C_{\alpha_0} \mid Z_{10} \geq C_{\gamma_0}],\\
R_{10} &:= [Z_1 \geq C_{\alpha} \mid Z_{10} \geq C_{\gamma_0}, Z_0 \geq C_{\alpha_0}],\\
R_{11} &:= [Z_1 \geq C_{\alpha_1} \mid Z_{10} \geq C_{\gamma_0}, Z_0 < C_{\alpha_0}],\\
\widetilde{R}_{11} &:= [\widetilde{Z}_1 \geq C_{\tilde{\alpha}_1} \mid Z_{10} < C_{\gamma_0}, Z_1 \geq C_{\gamma_1}].
\end{aligned}$$

When rejecting a hypothesis, one of the following conclusions must be reached:

$$\begin{aligned}
A_1 &:= S_0 \cap R_0 \cap R_{10} \quad \text{(No enrichment, reject both hypotheses)},\\
A_2 &:= S_0 \cap R_0 \cap \bar{R}_{10} \quad \text{(No enrichment, reject } H_0 \text{ only)},\\
A_3 &:= S_0 \cap \bar{R}_0 \cap R_{11} \quad \text{(No enrichment, reject } H_1 \text{ only)},\\
A_4 &:= \bar{S}_0 \cap S_1 \cap \widetilde{R}_{11} \quad \text{(Enrichment, reject } H_1).
\end{aligned}$$

Let $(\delta_0 = 0, \delta_1 = 0)$, then (all probabilities taken under $H_0 \cap H_1$)

$$\begin{aligned}
&\mathbb{P}[\text{At least one Type I error}] = \mathbb{P}[A_1 \cup A_2 \cup A_3 \cup A_4]\\
&= \mathbb{P}\big[\underbrace{(S_0 \cap R_0 \cap R_{10}) \cup (S_0 \cap R_0 \cap \bar{R}_{10})}_{=S_0 \cap R_0} \cup (S_0 \cap \bar{R}_0 \cap R_{11}) \cup (\bar{S}_0 \cap S_1 \cap \widetilde{R}_{11})\big]\\
&= \underbrace{\mathbb{P}[S_0 \cap R_0]}_{=(i)} + \underbrace{\mathbb{P}[(S_0 \cap \bar{R}_0 \cap R_{11})]}_{=(ii)} + \underbrace{\mathbb{P}[\bar{S}_0 \cap S_1 \cap \widetilde{R}_{11}]}_{=(iii)}.
\end{aligned}$$

Referring to Section 2.1.2, we know that $\mathrm{Corr}(Z_{10}, Z_0) = w_{10}$, $\mathrm{Corr}(Z_{10}, Z_1) = w_{11}\sqrt{f_{01}}$, $\mathrm{Corr}(Z_0, Z_1) = \sqrt{f_{01}}$, $\mathrm{Corr}(Z_{10}, \widetilde{Z}_1) = \widetilde{w}_{11}\sqrt{f_{01}}$ and $\mathrm{Corr}(Z_{11}, \widetilde{Z}_1) = \widetilde{w}_{11}$. Hence define the correlation matrices

$$\Sigma_1 = \begin{pmatrix} 1 & w_{10} & w_{11}\sqrt{f_{01}} \\ w_{10} & 1 & \sqrt{f_{01}} \\ w_{11}\sqrt{f_{01}} & \sqrt{f_{01}} & 1 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 1 & \sqrt{f_{01}} & \widetilde{w}_{11}\sqrt{f_{01}} \\ \sqrt{f_{01}} & 1 & \widetilde{w}_{11} \\ \widetilde{w}_{11}\sqrt{f_{01}} & \widetilde{w}_{11} & 1 \end{pmatrix}.$$

Now,

$$(i) = \mathbb{P}[Z_{10} \geq C_{\gamma_0}, Z_0 > C_{\alpha_0}] = \gamma_0 + \alpha_0 - 1 + F_{Z_{10},Z_0}(C_{\gamma_0}, C_{\alpha_0}; w_{10}),$$

and

$$\begin{aligned} (ii) &= \mathbb{P}[Z_{10} > C_{\gamma_0}, Z_0 \leq C_{\alpha_0}, Z_1 > C_{\alpha_1}] \\ &= 1 - \alpha_0 - F_{Z_{10},Z_0}\left(C_{\gamma_0}, C_{\alpha_0}; w_{10}\right) - F_{Z_0,Z_1}\left(C_{\alpha_0}, C_{\alpha_1}; \sqrt{f_{01}}\right) \\ &\qquad + F_{Z_{10},Z_0,Z_1}\left(C_{\gamma_0}, C_{\alpha_0}, C_{\alpha_1}; \Sigma_1\right). \end{aligned}$$

Finally,

$$\begin{aligned} (iii) &= \mathbb{P}[Z_{10} < C_{\gamma_0}, Z_{11} \geq C_{\gamma_1}, \widetilde{Z}_1 \geq C_{\tilde{\alpha}_1}] \\ &= 1 - \gamma_0 - F_{Z_{10},Z_{11}}\left(C_{\gamma_0}, C_{\gamma_1}; \sqrt{f_{01}}\right) - F_{Z_{10},\widetilde{Z}_1}\left(C_{\gamma_0}, C_{\tilde{\alpha}_1}; \widetilde{w}_{11}\sqrt{f_{01}}\right) \\ &\qquad + F_{Z_{10},Z_{11},\widetilde{Z}_1}\left(C_{\gamma_0}, C_{\gamma_1}, C_{\tilde{\alpha}_1}; \Sigma_2\right). \end{aligned}$$

Adding these up, we get

$$\begin{aligned} (i) + (ii) + (iii) \;=\; & 1 - F_{Z_0,Z_1}\left(C_{\alpha_0}, C_{\alpha_1}; \sqrt{f_{01}}\right) + F_{Z_{10},Z_0,Z_1}\left(C_{\gamma_0}, C_{\alpha_0}, C_{\alpha_1}; \Sigma_1\right) \\ & -F_{Z_{10},Z_{11}}\left(C_{\gamma_0}, C_{\gamma_1}; \sqrt{f_{01}}\right) - F_{Z_{10},\widetilde{Z}_1}\left(C_{\gamma_0}, C_{\tilde{\alpha}_1}; \widetilde{w}_{11}\sqrt{f_{01}}\right) \\ & +F_{Z_{10},Z_{11},\widetilde{Z}_1}\left(C_{\gamma_0}, C_{\gamma_1}, C_{\tilde{\alpha}_1}; \Sigma_2\right). \end{aligned} \tag{3.5}$$

To obtain $\tilde{\alpha}_1$, use a numerical search for $C_{\tilde{\alpha}_1}$ in Equation (3.5) to ensure that $(i)+(ii)+(iii) = \alpha$. Then, $\tilde{\alpha}_1 = \mathbb{P}[Z_1 \geq C_{\tilde{\alpha}_1} | H_1]$. Note that if $\gamma_1 = 1$ (no stopping for futility), then $C_{\gamma_1} = -\infty$ and the terms involving γ_1 will equal zero.

When $(\delta_0 = 0, \delta_1 \neq 0)$, a Type I error is committed if $A_1 \cup A_2 = S_0 \cap R_0$ occurs. But $S_0 \cap R_0 \subset R_0$ and $\mathbb{P}[R_0|H_0] = \alpha_0 < \alpha$. Hence Type I error probability is bounded above by α. Finally, suppose that $(\delta_0 \neq 0, \delta_1 = 0)$. Then a Type I error is committed when $A_1 \cup A_3 \cup A_4$ occurs. Note that for a sufficiently large sample size, A_3 and A_4 are dominated by A_1, which implies that

$$\begin{aligned} \mathbb{P}[A_1 \cup A_3 \cup A_4 \mid H_2] &\approx \mathbb{P}[S_0 \cap R_0 \cap R_{10} \mid H_1] \\ &\leq \Pr[R_{10} \mid H_1] = \alpha. \end{aligned}$$

We have thus shown that, asymptotically, FWER is strongly protected at level α.

Rejection Probabilities

As in Section 3.2.1, we can compute rejection probabilities numerically for the FE design. Again we suppose that the study is powered at level $1 - \beta$ to detect an effect $\delta_0 = \frac{\theta_0}{\sigma}$ in Ω_0, where σ^2 is the pooled variance for the overall population. The effect in Ω_1 is given by $\delta_1 = \eta\delta_0$. The non-centrality parameter $\lambda_0 = \mathbb{E}[Z_0|\delta]$ is readily obtained by noting that

$$\begin{aligned} 1 - \beta &= \mathbb{P}[A_1 \cup A_2 \mid \delta] \\ &= \mathbb{P}[S_0 \cap R_0 \mid \delta] \\ &= \mathbb{P}[Z_{10} \geq C_{\gamma_0}, Z_0 \geq C_{\alpha_0} \mid \delta] \\ &= 1 - F_{Z_{10}}(C_{\gamma_0}) - F_{Z_0}(C_{\alpha_0}) + F_{Z_{10},Z_0}(C_{\gamma_0}, C_{\alpha_0}; w_{10}) \\ &= 1 - \Phi(C_{\gamma_0} - w_{10}\lambda_0) - \Phi(C_{\alpha_0} - \lambda_0) + \Phi(C_{\gamma_0} - w_{10}\lambda_0, C_{\alpha_0} - \lambda_0; w_{10}). \end{aligned}$$

By specifying $\alpha_0, \beta, \gamma_0$ and $w_{10}^2 = t$, we can solve the above equation numerically to obtain $\lambda_0 \equiv \lambda_0(\alpha_0, \beta, \gamma_0, t)$. This gives us a known value for $\mathbb{E}[Z_0|\delta] = \sqrt{\frac{n_0}{4}}\delta_0$, and

$$\lambda_1 = \mathbb{E}(Z_1) = \theta_1\sqrt{\mathcal{I}_1} = \sqrt{\frac{n_1}{4}}\delta_1 = \eta\sqrt{f_{01}}\lambda_0.$$

When enrichment is envisaged,

$$\widetilde{\lambda}_1 = \mathbb{E}(\widetilde{Z}_1) = \sqrt{\frac{tf_{01}n_0}{4} + \frac{(1-t)n_0}{4}}\delta_1 = \eta\sqrt{(tf_{01}+1-t)}\lambda_0.$$

Note that, because $f_{01} \in [0,1]$,

$$\begin{aligned} f_{01} &= tf_{01} + (1-t)f_{01} \\ &\leq tf_{01} + 1 - t \\ \Rightarrow \lambda_1 &\leq \widetilde{\lambda}_1, \end{aligned}$$

which signifies the potential increase in power to reject H_1 when enrichment is carried out. For notational convenience, set $a_0 = C_{\alpha_0} - \lambda_0$, $a_1 = C_{\alpha_1} - \lambda_1$, $g_0 = C_{\gamma_0} - w_{10}\lambda_0$, $g_1 = C_{\gamma_1} - w_{11}\lambda_1$, $a = C_\alpha - \lambda_1$ and $\tilde{a} = C_{\tilde{\alpha}_1} - \widetilde{\lambda}_1$. The power of the subgroup study is then

$$\begin{aligned} &\text{Power}(\alpha_0, \gamma_0, \gamma_1, f_{01}, t, \eta, \delta_0) = \mathbb{P}[A_1 \cup A_3 \cup A_4 \mid \eta, \delta_0] \\ &\quad = \mathbb{P}[S_0 \cap R_0 \cap R_{10} \mid \eta, \delta_0] + \mathbb{P}[S_0 \cap \bar{R}_0 \cap R_{11} \mid \eta, \delta_0] + \mathbb{P}[\bar{S}_0 \cap S_1 \cap \widetilde{R}_{11} \mid \eta, \delta_0], \end{aligned}$$

where

$$\begin{aligned} \mathbb{P}[S_0 \cap R_0 \cap R_{10} \mid \eta, \delta_0] &= \mathbb{P}[Z_{10} \geq C_{\gamma_0}, Z_0 \geq C_{\alpha_0}, Z_1 \geq C_\alpha \mid \eta, \delta_0] \\ &= 1 - \Phi(g_0) - \Phi(a_0) + \Phi(g_0, a_0; w_{10}) + \Phi\left(g_0, a; w_{11}\sqrt{f_{01}}\right) \\ &\quad + \Phi\left(a_0, a; \sqrt{f_{01}}\right) - \Phi\left(g_0, a_0, a; \Sigma_1\right) \end{aligned}$$

and Σ_1 is defined in the previous section. Next,

$$\begin{aligned} \mathbb{P}[S_0 \cap \bar{R}_0 \cap R_{11} \mid \eta, \delta_0] &= \mathbb{P}[Z_{10} \geq C_{\gamma_0}, Z_0 < C_{\alpha_0}, Z_1 \geq C_{\alpha_1} \mid \eta, \delta_0] \\ &= \Phi(a_0) - \Phi(g_0, a_0; w_{10}) - \Phi\left(a_0, a_1; \sqrt{f_{01}}\right) \\ &\quad + \Phi(g_0, a_0, a_1; \Sigma_1), \end{aligned}$$

and

$$\begin{aligned}\mathbb{P}[\bar{S}_0 \cap S_1 \cap \widetilde{R}_{11} \mid \eta, \delta_0] &= \mathbb{P}[Z_{10} < C_{\gamma_0}, Z_{11} \geq C_{\gamma_1}, \widetilde{Z}_1 \geq C_{\tilde{\alpha}_1} \mid \eta, \delta_0] \\ &= \Phi(g_0) - \Phi\left(g_0, g_1; \sqrt{f_{01}}\right) - \Phi\left(g_0, \tilde{a}; \widetilde{w}_{10}\sqrt{f_{01}}\right) \\ &\quad +\Phi\left(g_0, g_1, \tilde{a}; \Sigma_2\right).\end{aligned}$$

Note that $\Phi(\cdot, \cdot; \psi)$ is the cumulative distribution function for a bivariate standard normal density with correlation coefficient ψ, and $\Phi(\cdot, \cdot, \cdot; \Sigma)$ is the three-dimensional CDF, with correlation matrix Σ. Σ_2 was defined in the previous section.

Sample Size and Efficiency

When computing power for the procedures AFP and FE, we obtain the non-centrality parameter λ_0 after specifying desired power, $1 - \beta$, given some value of $\delta_0 = \theta_0/\sigma$. However, due to the fact that FE allows an interim analysis, the required sample size necessary to power the study at the desired level is not the same in general. Consider the AFP for one subgroup. Knowing λ_0 and δ_0, we can re-arrange the well known formula for sample size to obtain the required observed Fisher's information $\mathcal{I}_{AFP}$

$$\lambda_0 = \sqrt{\frac{n_0}{4}}\delta_0 = (Z_{1-\alpha_0} + Z_{1-\beta}) \Rightarrow \left(\frac{Z_{1-\alpha_0} + Z_{1-\beta}}{\theta_0}\right)^2 = \frac{n_0}{4\sigma^2} =: \mathcal{I}_{AFP}.$$

For the FE procedure we simply use the value obtained numerically for $\lambda_0 = \lambda(\alpha_0, \beta, \gamma_0, t)$ and set

$$\mathcal{I}_{FE} = \left(\frac{\lambda(\alpha_0, \beta, \gamma_0, t)}{\theta_0}\right)^2.$$

Define $\mathcal{E}_{FE} := \mathcal{I}_{FE}/\mathcal{I}_{AFP}$ to compare the relative efficiency of FE with that of AFP. Then

$$\mathcal{E}_{FE} = \frac{\mathcal{I}_{FE}}{\mathcal{I}_{AFP}} = \frac{n_{FE}}{n_{AFP}}$$

Table 3.1: FE required sample size as a percentage of that of AFP, given various values of γ_0, with $t = 1/2$.

γ_0	0.1	0.2	0.3	0.4	0.5	0.6
$\mathcal{E}_{FE}$	1.2686	1.0937	1.0378	1.0155	1.0061	1.0022

so $\mathcal{E}_{FE}$ gives the sample size required for FE as a percentage of the sample size required for AFP. Note that sample size requirements for H_0 in the FE procedure are not affected by γ_1, the futility parameter. Hence values of $\mathcal{E}_{FE}$ will be the same regardless of whether we allow early stopping for futility or not (in Section 4.2.1, we discuss the effect of early stopping for futility on expected sample size). Table 3.1 shows $\mathcal{E}_{FE}$ for various γ_0 values with $t = 1/2$. We see that for $\gamma_0 \geq 0.3$, the increase in sample size is minimal or less than 4%. As γ_0 gets smaller, however, the difference in necessary sample size starts to increase quite a bit. Under H_0 and with $\gamma_0 = 0.1$, the probability of enrichment is equal to 0.9. Although this probability will decrease as δ_0 increases, the likelihood that enrichment is envisaged remains quite substantial with such a low value for γ_0. Therefore a larger sample size is required to give sufficient power to test H_0.

3.2.3 Adaptive Design with Conditional Power (Two Stages)

As done by Wang et al. (2009), we can specify an adaptive design where all decisions at the interim analysis are based on conditional power, i.e. the probability that we reject the respective hypotheses, conditional on first stage results. We call this design CP. Let

$$\mathrm{CP}_{\Omega_j}(t, z_{1j}, \delta) = \mathbb{P}[\text{Reject } H_j \mid Z_{1j} = z_{1j}, \delta_j = \delta],\ j = 0, 1,$$

be the conditional power for H_j, given interim analysis time t, first stage results $Z_{1j} = z_{1j}$, and true effect size δ. For Ω_0, we have

$$\begin{aligned}
\mathrm{CP}_{\Omega_0}(t, z_{10}, \delta) &= \mathbb{P}\left[Z_0 \geq C_\alpha \mid Z_{10} = z_{10}, \delta_0 = \delta\right] \\
&= \mathbb{P}\left[Z_{20} \geq \frac{-C_\alpha + \sqrt{t} \cdot z_{10}}{\sqrt{1-t}} \mid \delta_0 = \delta\right] \\
&= \Phi\left(\frac{-C_\alpha + \sqrt{t} \cdot z_{10} + (1-t)\sqrt{\mathcal{I}_0} \cdot \theta}{\sqrt{1-t}}\right).
\end{aligned}$$

The choice of δ for $\mathbb{E}(Z_{2j}) = \sqrt{\frac{n_{2j}}{\sigma}}\delta_j$ impacts the conditional power, and hence subsequent decisions. Two possible choices are to use the first stage estimate $\hat{\delta}_j$, or the planned effect size (clinically significant effect size). If we use the first stage estimate, $\hat{\theta}_{10} = z_{10}/\sqrt{\mathcal{I}_{10}}$, CP_{Ω_0} evaluates to

$$\begin{aligned}
\mathrm{CP}_{\Omega_0}(t, z_{10}, \hat{\delta}_{10}) &= \Phi\left(\frac{-C_\alpha + \sqrt{t} \cdot z_{10} + (1-t)\sqrt{\mathcal{I}_0} \cdot z_{10}/\sqrt{\mathcal{I}_{10}}}{\sqrt{1-t}}\right) \\
&= \Phi\left(\frac{1}{w_{20}}\left[-C_\alpha + w_{10}z_{10} + w_{20}^2 z_{10}/w_{10}\right]\right) \\
&= \Phi\left(\frac{1}{w_{20}}\left[-C_\alpha + z_{10}/w_{10}\right]\right).
\end{aligned}$$

For Ω_1, we get

$$\begin{aligned}
\mathrm{CP}_{\Omega_1}(t, z_{11}, \delta) &= \mathbb{P}\left[\widetilde{Z}_1 \geq C_\alpha \mid Z_{11} = z_{11}, \delta_1 = \delta\right] \\
&= \Phi\left(\frac{1}{\sqrt{1-\tilde{t}}}\left[-C_\alpha + \sqrt{\tilde{t}} \cdot z_{11} + \sqrt{(1-t)\mathcal{I}_0}\sqrt{1-\tilde{t}} \cdot \theta\right]\right),
\end{aligned}$$

where $\tilde{t}$ is defined in Equation (3.1). By using $\hat{\theta}_{11} = z_{11}/\sqrt{\mathcal{I}_{11}}$, we get

$$\begin{aligned}
\mathrm{CP}_{\Omega_1}(t, z_{11}, \hat{\delta}_{11}) &= \Phi\left(\frac{1}{\widetilde{w}_{21}}\left[-C_\alpha + \widetilde{w}_{11}z_{11} + \widetilde{w}_{21}^2 z_{11}/\widetilde{w}_{11}\right]\right) \\
&= \Phi\left(\frac{1}{\widetilde{w}_{21}}\left[-C_\alpha + z_{11}/\widetilde{w}_{11}\right]\right).
\end{aligned}$$

The choices mentioned for δ when computing CP_{Ω_i} both have drawbacks. First, the estimated effect size after stage one can be highly variable due to a small sample size. To see this, note that

$$\mathrm{Var}(\hat{\theta}_{11}) = \mathcal{I}_{11}^{-1} = \frac{1}{f_{01}t\mathcal{I}_0}.$$

Now if, for example, $f_{01} = t = 1/4$, the variance of $\hat{\theta}_{11}$ is sixteen times the variance of $\hat{\theta}_0$, the final estimate of θ_0. On the other hand, using the clinically significant effect size can lead to poor results if the experimental treatment is believed to be more efficacious than it truly is, as using too large a value for δ_0 when computing CP_{Ω_0} may result in overly optimistic decisions at interim.

Let $L_0^{CP}, L_1^{CP} \in [0, 1]$. The interim assessment for the CP design is as follows:

- If $\text{CP}_{\Omega_0}(t, Z_{10}, \hat{\delta}_{10}) \geq L_0^{CP}$, then proceed to stage two with the full population.
- Else, if $\text{CP}_{\Omega_1}(t, Z_{11}, \hat{\delta}_{11}) \geq L_1^{CP}$, then proceed to stage two sampling from Ω_1 only.
- Else, abandon the trial and accept both hypotheses H_0 and H_1.

The constants L_0^{CP} and L_1^{CP} should be carefully chosen to ensure that "sensible" decisions are made at interim. To this end, extensive simulations should be run over a variety of plausible configurations of θ_1 and θ_2, and choices for L_0^{CP} and L_1^{CP} made depending on desired operating characteristics.

3.2.4 Hybrid Designs (Two Stages)

As discussed in Section 2.2.3, adaptive seamless designs (Bretz et al., 2006; Schmidli et al., 2006) allow arbitrary modifications mid-trial while preserving the Type I error rate. In particular, we can then use Bayesian computational tools at interim analysis points, which necessitates the specification of prior distributions on all parameters of interest. As $Z_j \sim \mathcal{N}(\theta_j \sqrt{\mathcal{I}_j}, 1)$ the unknown components of the parameters of our test statistics are θ_j, $j = 1, 2$. In what follows, we assume that the parameters (θ_1, θ_2) are exchangeable in the joint distribution so $f(\theta_1, \theta_2)$ is invariant to index permutations. For the treatment effect parameters, and for the hyperparameter ν, we impose the following prior distributions:

$$\theta_1, \theta_2 \overset{iid}{\sim} \mathcal{N}(\nu, \tau^2), \text{ and } \nu \sim \mathcal{N}(\phi, \omega^2).$$

We note that since Ω_1 has been identified beforehand, the exchangeability assumption may not be appropriate. However, it can be viewed as the "pessimist" modeling assumption that Ω_1 is no more likely to be responsive than Ω_2. As $\theta_0 = f_{01}\theta_1 + (1 - f_{01})\theta_2$, a prior on θ_0 has implicitly been specified. While one could also impose prior distributions on the dispersion parameters τ and ω, we treat these as known. Similarly, we will specify a value for ϕ. The assumption that $\mathbb{E}(\theta_j) = \nu$ for all j models the potential relationship between effect sizes in individual subgroups, while setting τ large means there is strong *a priori* belief in the presence of heterogeneity. Given observations $z_{1j} = \hat{\theta}_{1j}\sqrt{\mathcal{I}_{1j}}$, posterior distributions for θ_j and ν are (derived in Section 4.3)

$$\theta_j \mid z_{1j}, \nu, \tau^2 \sim \mathcal{N}\left\{\hat{\theta}_{1j} - \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}\left(\hat{\theta}_{1j} - \nu\right), \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}\tau^2\right\}, \; j = 1, 2$$

$$\nu \mid z_j, \tau^2, \phi, \omega^2 \sim \mathcal{N}\left\{\sigma^2_{\nu|z}\left(\sum_{j=1}^{2} \frac{\hat{\theta}_{1j}}{\mathcal{I}_{1j}^{-1} + \tau^2} + \frac{\phi}{\omega^2}\right), \sigma^2_{\nu|z}\right\}$$

where

$$\sigma^2_{\nu|z} = \left(\sum_{j=1}^{2} \frac{1}{\mathcal{I}_{1j}^{-1} + \tau^2} + \frac{1}{\omega^2}\right)^{-1}.$$

Computations based on the posterior distribution of θ_j require a value for ν. For this, we use its estimated posterior expected value. That is,

$$\mu_{\theta_j|z} := \mathbb{E}\left[\theta_j \mid z_{1j}, \nu, \tau^2\right] = \hat{\theta}_{1j} - \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}(\hat{\theta}_{1j} - \hat{\nu})$$

where

$$\hat{\nu} = \sigma^2_{\nu|z}\left(\sum_{j=1}^{2} \frac{\hat{\theta}_{1j}}{\mathcal{I}_{1j}^{-1} + \tau^2} + \frac{\phi}{\omega^2}\right).$$

Utility-Based Interim Analysis

We first propose a hybrid design that relies on the use of a utility function at the interim analysis. The procedure is abbreviated as HUT. We specify a loss function to quantify the cost of available decisions, depending on the true value of the parameters θ_j. Let $\mathcal{P}_1 = \{1, 2\}$ denote the index set of available populations at the onset of the study. Fix $\theta^-, \theta^+ \in \mathbb{R}$ such that $\theta^- \leq \theta^+$, and suppose the following rules hold:

1. If $\theta_j \leq \theta^-$, we always want to accept H_j, and incur a cost c_1 when Ω_j is chosen;
2. If $\theta_j \geq \theta^+$, we always want to reject H_j, and incur a cost c_2 when Ω_j is not chosen;
3. If $\theta_j \in (\theta^-, \theta^+)$, no costs are incurred either way.[1]

Ω_j is said to be *chosen* if $\mathcal{P}_2$ (the index set of populations chosen for stage two) contains j. Let $k^{HUT} = c_2/c_1$. Hence if $k^{HUT} < 1$, we will tend to eliminate populations from consideration unless early results are very good. This might be an appropriate setting if subsequent stages will either be very time-consuming and/or expensive. On the other hand, if $k^{HUT} > 1$ there is a higher cost for false negatives, and we will refrain from unnecessarily eliminating populations. This is desirable if early estimates are highly variable.

At the interim analysis, we choose a decision $d \subset \mathcal{P}_1$. This decision can be to abandon the trial for futility ($d = \varnothing$), proceed with Ω_1 only ($d = \{1\}$), or to

[1] Rule 3 describes what is often referred to as an "indifference region," see for example (Bechhofer et al., 1995, p. 8) for further details.

proceed with the full population ($d = \{1, 2\}$). The decision is based on the set $\mathcal{P}_2$ associated with the largest value of our utility function:

- $d = \varnothing$ has utility $-\mathbb{P}[\theta_1 \geq \theta^+ | z_{11}] - k^{HUT} \cdot \mathbb{P}[\theta_2 \geq \theta^+ | z_{12}]$.
- $d = \{1\}$ has utility $-\mathbb{P}[\theta_1 \leq \theta^- | z_{11}] - k^{HUT} \cdot \mathbb{P}[\theta_2 \geq \theta^+ | z_{12}]$.
- $d = \{1, 2\}$ has utility $-\mathbb{P}[\theta_1 \leq \theta^- | z_{11}] - k^{HUT} \cdot \mathbb{P}[\theta_2 \leq \theta^- | z_{12}]$.

In Chapter 4, Section 4.3, we outline this method for an arbitrary number of populations. We also give a detailed derivation of the utility function.

Predictive Probabilities

Given the interim data, we can compute the predictive probability that various hypotheses will be rejected, see (Brannath et al., 2009). The predictive distributions of second-stage parameter estimates take into account both the variability in the second-stage observations, and the uncertainty about first-stage estimates.

We first derive the predictive distributions for Z_0 and Z_1, given the first stage data. Let $\mu_{\theta_j|z} = \mathbb{E}[\theta_j | z_{1j}, \nu, \tau^2]$ and $\sigma^2_{\theta_j|z} = \mathrm{Var}[\theta_j | z_{1j}, \tau^2]$, $j = 1, 2$. Note that

$$\mu_{\theta_0|z} = f_{01}\mu_{\theta_1|z} + (1 - f_{01})\mu_{\theta_2|z}$$
$$\text{and } \sigma^2_{\theta_0|z} = f_{01}^2\sigma^2_{\theta_1|z} + (1 - f_{01})^2\sigma^2_{\theta_2|z}.$$

Now, with $Z_{1j} = z_{1j}$,

$$\begin{aligned}
Z_j &= w_{1j} z_{1j} + w_{2j} Z_{2j}, \quad \text{where } Z_{2j} \sim \mathcal{N}(\theta_j \sqrt{\mathcal{I}_{2j}}, 1) \\
&= w_{1j} z_{1j} + w_{2j} \left[\theta_j \sqrt{\mathcal{I}_{2j}} + \varepsilon \right], \quad \text{where } \varepsilon \sim \mathcal{N}(0,1) \\
&= w_{1j} z_{1j} + w_{2j} \left[\mu_{\theta_j|z} \sqrt{\mathcal{I}_{2j}} + \underbrace{(\theta_j - \mu_{\theta_j|z})}_{\sim \mathcal{N}(0, \sigma^2_{\theta_j|z})} \sqrt{\mathcal{I}_{2j}} + \varepsilon \right] \\
&= \mathcal{N} \left\{ w_{1j} z_{1j} + w_{2j} \mu_{\theta_j|z} \sqrt{\mathcal{I}_{2j}}, \ w_{2j}^2 \left(\mathcal{I}_{2j} \sigma^2_{\theta_j|z} + 1 \right) \right\}.
\end{aligned}$$

Denote $v_j^2 = w_{2j}^2 \left(\mathcal{I}_{2j} \sigma^2_{\theta_j|z} + 1 \right)$ and let

$$\eta_j = \frac{1}{v_j} \left(w_{1j} z_{1j} + w_{2j} \mu_{\theta_j|z} \sqrt{\mathcal{I}_{2j}} \right), \ j = 0, 1.$$

Then the predictive distribution for (Z_0, Z_1) is

$$\begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} \sim \mathcal{N} \left\{ \begin{pmatrix} \eta_0 \\ \eta_1 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{f_{01}} \\ \sqrt{f_{01}} & 1 \end{pmatrix} \right\}.$$

If enrichment is envisaged, the predictive distribution for $\widetilde{Z}_1$ is

$$\widetilde{Z}_1 \sim \mathcal{N} \left\{ \widetilde{w}_{11} z_{11} + \widetilde{w}_{21} \mu_{\theta_1|z} \sqrt{\mathcal{I}_{20}}, \ \widetilde{w}_{2j}^2 \left(\mathcal{I}_{20} \sigma^2_{\theta_1|z} + 1 \right) \right\},$$

where $\widetilde{w}_{11}$ and $\widetilde{w}_{21}$ are defined in Equations (3.2) and (3.3), respectively. Next, let

$$\begin{aligned}
\pi_0 &= \mathbb{P}[\text{Reject } H_0 \text{ or } H_1 \text{ if continue with } \Omega_0] \\
\pi_1 &= \mathbb{P}[\text{Reject } H_1 \text{ if proceed with } \Omega_1 \text{ only}] \\
\pi_2 &= \mathbb{P}[\text{Reject } H_2 \text{ if continue with } \Omega_0],
\end{aligned}$$

where the probabilities are taken under the predictive distributions above. Then,

$$\begin{aligned}
\pi_0 &= \mathbb{P}[Z_0 \geq C_\alpha \text{ or } Z_1 \geq C_\alpha] \\
&= 1 - \Phi \left(C_\alpha - \eta_0, C_\alpha - \eta_1; \sqrt{f_{01}} \right)
\end{aligned}$$

and

$$\pi_1 = \mathbb{P} \left[\widetilde{Z}_1 \geq C_\alpha \right] = \Phi \left(-C_\alpha + \widetilde{\eta}_1 \right).$$

with $\widetilde{\eta}_1 = \left(\widetilde{w}_{11}z_{11} + \widetilde{w}_{21}\mu_{\theta_1|z}\sqrt{\mathcal{I}_{20}}\right)/\widetilde{v}_1$ and $\widetilde{v}_1^2 = \widetilde{w}_{2j}^2\left(\mathcal{I}_{20}\sigma^2_{\theta_1|z} + 1\right)$. π_2 is computed in similar fashion. Note that π_0 is only an upper bound on the predictive probability of rejection when continuing with the full population. This is because we do not account for multiplicity adjustments which will be carried out in the final analysis.

Let $L_0^{HPP}, L_1^{HPP}, L_2^{HPP} \in [0,1]$. At interim, the predictive probability (HPP) design is outlined as follows:

- If $\pi_0 \geq L_0^{HPP}$, $\pi_1 \geq L_1^{HPP}$ and $\pi_2 \geq L_2^{HPP}$, then continue to the second stage sampling from the full population.
- If $\pi_0 \geq L_0^{HPP}$ and $\pi_1 < L_1^{HPP}$, then also continue sampling from the full population.
- If $\pi_0 \geq L_0^{HPP}$, $\pi_1 \geq L_1^{HPP}$ but $\pi_2 < L_2^{HPP}$ then enrich and continue with Ω_1 only.
- If $\pi_0 < L_0^{HPP}$ and $\pi_1 \geq L_1^{HPP}$ then enrich and continue with Ω_1 only.
- If $\pi_0 < L_0^{HPP}$ and $\pi_1 < L_1^{HPP}$ then abandon the trial and accept all hypotheses.

The HPP design will only proceed with the full population if there is a sufficiently high probability of a positive finding at the trial conclusion. Note that if the first-stage results in Ω_2 are poor, then the design may choose to enrich even if the overall results are good. This is desirable as in such cases the positive overall effect is likely driven by very good findings in Ω_1. As with the CP design, the constants L_j^{HPP} need to be strategically chosen to ensure desirable operating characteristics at the interim analysis.

3.3 Numerical Study

To compare the designs outlined in Section 3.2, we conduct a Monte Carlo simulation under multiple states of nature, and for a variety of procedure-specific parameter specifications. One issue of interest is to compare operating characteristics of the designs previously described. Additionally, we consider a number of questions that naturally arise when carrying out clinical trials with population-specific hypothesis testing. For example,

1. What is the influence of the prevalence of Ω_1 in Ω_0 (f_{01}) on rejection probabilities for H_0 and H_1?
2. How important is the timing of an interim analysis? Does it affect H_0 and H_1 similarly, or is there a trade-off?
3. For what types of values of (θ_1, θ_2) are adaptive designs preferable to fixed designs, and vice-versa?
4. How do procedures specifically tailored to subgroup testing perform compared to standard p-value adjustment procedures, or group sequential methods?
5. How robust are the adaptive designs to selecting the correct course of action at the interim analysis?
6. For an adaptive design, are transparency and ease of exposition preferable to a seemingly well performing "black box?"

Simulations were run for values of $\theta_1 \in [0, 30]$, with $\theta_2 \in \{0, 10, 20\}$ and we set $\theta^* = 20$ as the clinically significant treatment effect. We use $f_{01} \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$ and $t \in \left\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\right\}$. The total information $\mathcal{I}_{\text{total}}$ is set as 0.0271, which guarantees power equal to 0.9 ($\beta = 0.1$) in a 2-stage trial to test H_0 only (no subgroup analysis) at level $\alpha = 0.025$ (O'Brien and Fleming, 1979). Included with our results, are simulations from a 2-endpoint, 2-stage O'Brien & Fleming design adjusted for multiplicity (abbreviated as OBFm) using a weighted Bonferroni correction for the individual endpoints, see for example (Jennison and Turnbull, 2000, Ch. 15.2). We also include results from a non-adaptive one stage design that uses the Hochberg-Simes (HS) method to test individual hypotheses.

For AFP and FE, we use $\alpha_0 \in \{0.02, 0.015\}$ and $\alpha_1 \in \{0.005, 0.01\}$. For the multiple-endpoint OBFm design, we use the same α-levels as in AFP, in addition to Bonferroni-levels 0.025/2. For FE, we use $\gamma_0 \in \{0.3, 0.4, 0.5\}$ and $\gamma_1 \in \{0.5, 0.75, 1\}$. For the adaptive methods CP and HPP, we use the following parameter values: $L_0^{CP} \in \{0.2, 0.4, 0.6, 0.8\}$, $L_1^{CP} \in \{0, 0.2, 0.4, 0.6, 0.8\}$, $L_0^{HPP}, L_1^{HPP} \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, $L_2^{HPP} \in \{0.1, 0.25, 0.5\}$. Finally, for the HUT design we let $\theta^+ \in \{0, 10, 20\}$ and $k^{HUT} \in \{0.5, 1, 1.5\}$.

For an initial analysis, Tables 3.2 and 3.3 show simulated rejection- and interim decision probabilities for all procedures described above, using $t = 1/4$ and 1/2, respectively. Treatment effect parameters, (θ_1, θ_2), equal $(0, 0)$, $(20, 0)$ and $(20, 20)$, with $f_{01} = 1/4$. First we see that, relative to traditional methods such as HS or OBFm, procedures that allow enrichment (FE or CP/HPP/HUT) can substantially improve power for H_1. Adaptive procedures also perform reasonably well when $(\theta_1, \theta_2) = (20, 20)$, with FE marginally outperforming others.

When the full null hypothesis is true, $(\theta_1, \theta_2) = (0, 0)$, all procedures control Type I error at the required level. Note that the adaptive procedures use empirical data weights, and are still within specified limits for α for the cases considered. Stops for futility range from rare (2–3% for OBFm), to quite common (55–65%

for CP/HPP). For adaptive procedures, these decisions are highly dependent on predetermined decision parameters such as L_0^{HPP} and L_1^{HPP}. E.g. for CP, with $L_0^{CP} = L_1^{CP} = 0.6$ and $t = 1/2$, futility stops occur with 87% probability. This is certainly desirable under the complete null hypothesis, but for other values of θ_i (and same L_i^{CP}), CP stops too often for futility (23% when $(\theta_1, \theta_2) = (20, 20)$ and $t = 1/2$). Entries in the tables are obtained from $L_0^{CP} = 0.4$ and $L_1^{CP} = 0.2$.

It is quite clear that for $t = 1/2$, interim decisions are "better" than for $t = 1/4$. We should expect this, as $t = 1/2$ implies that more information is accumulated during the first stage than for $t = 1/4$. A more interesting question is whether delaying the interim analysis leads to better decisions at the end of the trial. Consider the case $(\theta_1, \theta_2) = (20, 0)$. In this case, enrichment is the proper decision at interim, and at the final analysis we should reject H_1 only (though reaching a positive result for Ω_0 is not necessarily "wrong"). For $t = 1/4$, HPP and HUT enrich roughly 47% of the time and both reach a positive result for Ω_1 roughly 45% of the time. When $t = 1/2$, enrichment is more likely (51% and 56% respectively) but now HPP has lower power for H_1. HUT, on the other hand, has similar power for H_1 and reaches a positive result with slightly higher probability than for $t = 1/4$ (52% vs. 51%). FE sees a slightly larger power reduction for H_1 when t is increased for reasons detailed in Section 3.3.1. Overall, there is only a slight difference between the adaptive designs in power performance.

If $(\theta_1, \theta_2) = (20, 20)$, the correct interim decision is to proceed to the second stage using the full population. Enrichment, while undesirable, is not as disastrous as abandoning the trial for futility. Again, as expected, increasing t has a positive effect on interim decisions for all adaptive procedures. We also see that a positive result is reached with greater frequency, in particular for methods such as CP, HPP and HUT. With $t = 1/2$, adaptive procedures are closer to fixed procedures such as FE or AFP, as HUT reaches a positive conclusion with 83% probability vs. 88% for FE and AFP. With $t = 1/4$, this difference is almost doubled.

In conclusion, early findings discussed above indicate that t should not be too small. Most adaptive designs achieve similar power for H_1 (when $(\theta_1, \theta_2) = (20, 0)$) regardless of whether $t = 1/4$ or $t = 1/2$, but do much better for H_0 (when $(\theta_1, \theta_2) = (20, 20)$) when $t = 1/2$. In subsequent sections, we inspect in greater detail the effects of t and f_{01}, the performance of adaptive designs at interim, and the performance of all designs at the final stage. The analysis of t and f_{01} is limited to selected procedures.

Table 3.2: Simulated probabilities for three configurations of (θ_1, θ_2), $t = 1/4$ and $f_{01} = 1/4$. All procedures are described in Chapter 3. Parenthesized value in OBFm row indicates "early rejection" of H_0.

Effect Size	Procedure	Final Decision				Interim Decision		
(θ_1, θ_2)		None	H_1 Only	H_0	H_0 or H_1	Futility	Enrichment	Full Population
(0,0)	AFP	0.975	0.0100	0.0150	0.0250	N/A	N/A	N/A
	HS	0.9759	0.0096	0.0145	0.0241	N/A	N/A	N/A
	OBFm	0.9766	0.0103	0.0132	0.0235	0.0176	(0.0013)	0.9810
	FE	0.9750	0.0119	0.0131	0.0250	0.2096	0.3904	0.4000
	CP	0.9812	0.0106	0.0081	0.0188	0.5607	0.2466	0.1926
	HPP	0.9814	0.0121	0.0066	0.0187	0.5569	0.3146	0.1286
	HUT	0.9808	0.0111	0.0081	0.0192	0.5026	0.2625	0.2348
(20,0)	AFP	0.6969	0.2141	0.0890	0.3031	N/A	N/A	N/A
	HS	0.6909	0.1987	0.1105	0.3091	N/A	N/A	N/A
	OBFm	0.7000	0.2161	0.0840	0.3000	0.0025	(0.0112)	0.9863
	FE	0.4975	0.4238	0.0787	0.5025	0.1058	0.3821	0.5121
	CP	0.5164	0.4114	0.0722	0.4836	0.2679	0.4085	0.3235
	HPP	0.4790	0.4568	0.0642	0.5210	0.2522	0.4756	0.2721
	HUT	0.4880	0.4462	0.0658	0.5120	0.2216	0.4619	0.3165
(20,20)	AFP	0.1237	0.0071	0.8691	0.8763	N/A	N/A	N/A
	HS	0.1344	0.0042	0.8613	0.8656	N/A	N/A	N/A
	OBFm	0.1405	0.0086	0.8509	0.8595	0.0002	(0.0659)	0.9339
	FE	0.1281	0.0510	0.8209	0.8719	0.0048	0.0607	0.9346
	CP	0.2231	0.0806	0.6963	0.7769	0.1252	0.0935	0.7813
	HPP	0.2514	0.1297	0.6188	0.7486	0.1666	0.1531	0.6803
	HUT	0.2014	0.0893	0.7093	0.7986	0.0926	0.1024	0.8051

Table 3.3: Simulated probabilities for three configurations of (θ_1, θ_2), $t = 1/2$ and $f_{01} = 1/4$. All procedures are described in Chapter 3. Parenthesized value in OBFm row indicates "early rejection" of H_0.

Effect Size	Procedure	Final Decision				Interim Decision		
(θ_1, θ_2)		None	H_1 Only	H_0	H_0 or H_1	Futility	Enrichment	Full Population
(0,0)	AFP	0.9750	0.0100	0.0150	0.0250	N/A	N/A	N/A
	HS	0.9760	0.0098	0.0142	0.0240	N/A	N/A	N/A
	OBFm	0.9767	0.0103	0.0130	0.0233	0.0314	(0.0014)	0.9673
	FE	0.9750	0.0103	0.0147	0.0250	0.2096	0.3904	0.4000
	CP	0.9795	0.0112	0.0093	0.0205	0.6680	0.2275	0.1044
	HPP	0.9803	0.0107	0.0090	0.0197	0.6560	0.2064	0.1375
	HUT	0.9790	0.0113	0.0097	0.0210	0.5514	0.2637	0.1849
(20,0)	AFP	0.6969	0.2141	0.0890	0.3031	N/A	N/A	N/A
	HS	0.6906	0.1997	0.1097	0.3094	N/A	N/A	N/A
	OBFm	0.6980	0.2086	0.0934	0.3020	0.0020	(0.0252)	0.9728
	FE	0.5700	0.3425	0.0875	0.4300	0.0759	0.3444	0.5797
	CP	0.4962	0.4230	0.0809	0.5038	0.2465	0.5040	0.2495
	HPP	0.5075	0.4170	0.0755	0.4925	0.2393	0.5075	0.2533
	HUT	0.4783	0.4468	0.0749	0.5217	0.1619	0.5573	0.2808
(20,20)	AFP	0.1237	0.0071	0.8691	0.8763	N/A	N/A	N/A
	HS	0.1340	0.0042	0.8618	0.8660	N/A	N/A	N/A
	OBFm	0.1395	0.0084	0.8521	0.8605	0.0000	(0.2035)	0.7965
	FE	0.1222	0.0126	0.8652	0.8778	0.0004	0.0144	0.9852
	CP	0.1834	0.0467	0.7699	0.8166	0.0817	0.0614	0.8569
	HPP	0.1778	0.0406	0.7816	0.8222	0.0659	0.0503	0.8838
	HUT	0.1664	0.0413	0.7923	0.8336	0.0390	0.0524	0.9086

3.3.1 Influence of Interim Analysis Timing

We begin by examining the effect of differently timed interim analyses. Intuitively, an early interim analysis (small t) suffers from inexact estimates due to a relatively small sample size in the first stage. On the other hand, there should be more to gain from enrichment, as the second stage offers a potentially larger sample size increase for Ω_1. Early results from the previous section indicated that $t \in [\frac{1}{4}, \frac{1}{2}]$ may be appropriate. Figure 3.1 shows empirical probabilities for interim analysis decisions for the HUT procedure, using various values of t. It is clear in these plots that using larger values of t increases the probability of making the correct decision at interim. This result is not surprising, as a longer first stage leads to more accurate estimates at interim.

An argument in favor of small t might be that shortening the first stage leaves more room for error in the trial planning process. That is, should we discover that enrichment is necessary, a shorter first stage allows us to sample more patients for Ω_1 during the second stage, and hence the likelihood of a positive outcome is increased. To investigate this, the impact of t on HUT power performance is shown in Figure 3.2, which displays various rejection probabilities using both $\theta_2 = 0$ and $\theta_2 = 20$. When $\theta_2 = 0$ (plot (a)), we see that when $\theta_1 \geq 20$, power is largest for $t = 1/2$. Using $t = 1/4$ is close for smaller values of θ_1 while power for $t = 3/4$ is dominated until $\theta_1 \approx 25$. It is clear from the plot that using a longer first stage, i.e. $t = 3/4$, is helpful for interim decisions but does not necessarily lead to high power at the end. On the other hand, power for $t = 1/4$ is dominated by that of $t = 1/2$ for moderate-to-high values of θ_1. The behavior is similar in plot (c), as $t = 1/2$ results in highest probability of a positive outcome. As θ_1 grows, power for $t = 3/4$ is greater than power for $t = 1/4$.

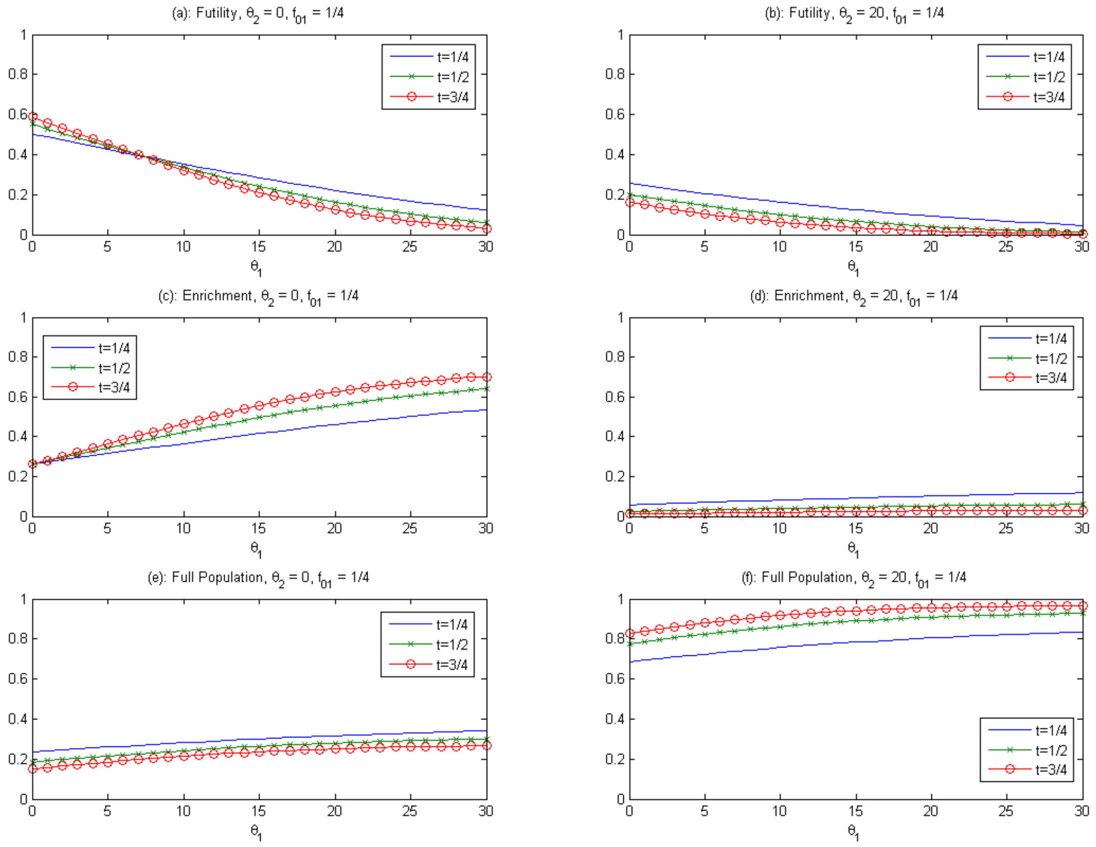


Figure 3.1: Empirical interim analysis probabilities for the HUT procedure. The left and right columns use $\theta_2 = 0$ and 20 respectively. Plots (a) and (b) show the probability of stopping for futility, (c) and (d) show enrichment probabilities, and (e) and (f) show the probability of proceeding with the full population. $f_{01} = 1/4$ throughout.

If $\theta_2 = 20$, we are interested in power for H_0. In plot (b), we see that power appears to increase with t, as expected. If $\theta_2 = 20$, enrichment should not be carried out at interim, and larger t reduce the probability of erroneously restricting sampling to Ω_1 in stage two. Referring to plot (d), we can see that increasing t past $1/2$ results in no tangible gain in power for H_0 or H_1. Power for $t = 1/4$ is always dominated by that of $t = 1/2$ or $t = 3/4$.

In light of the results observed in Figure 3.2, it appears that a sensible choice for the interim analysis is close to the middle of the trial ($t = 1/2$). An early analysis ($t = 1/4$), while allowing for a greater benefit of enrichment in stage two,

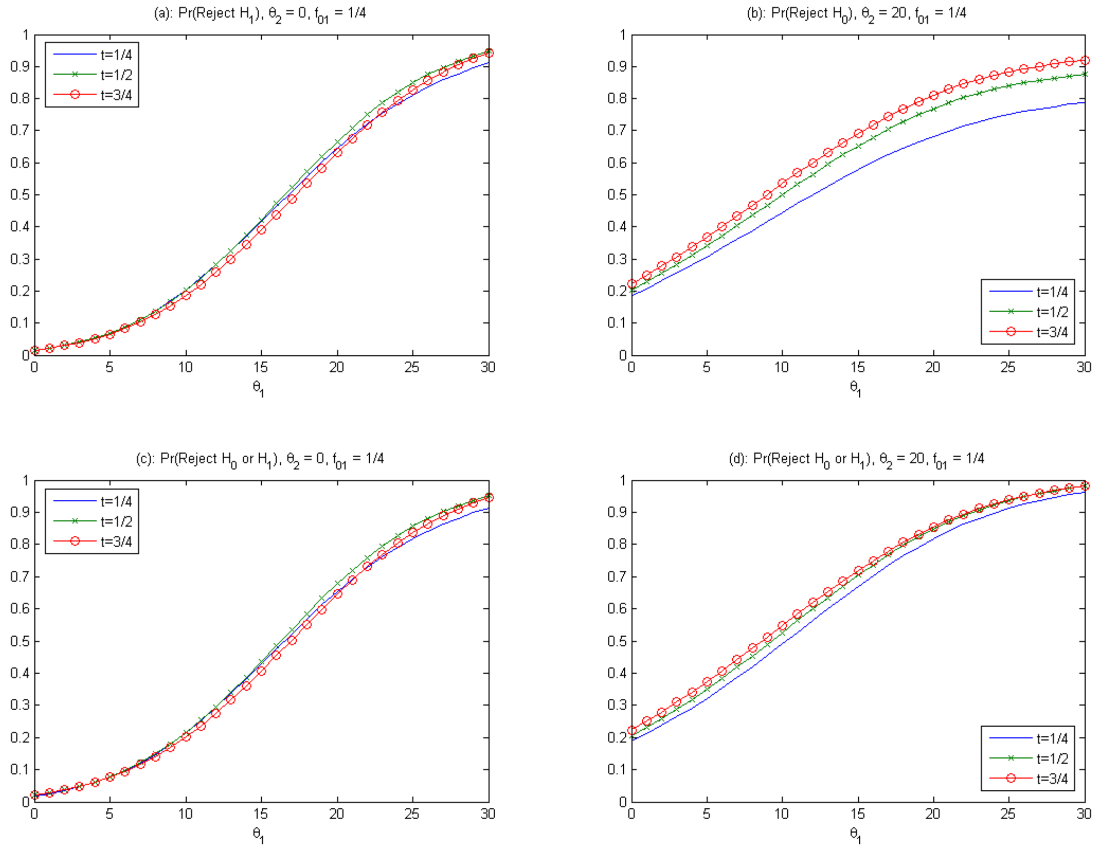


Figure 3.2: Empirical rejection probabilities for the HUT procedure. The left and right columns use $\theta_2 = 0$ and 20 respectively. Plots (a) and (b) show the probability of rejecting H_1 and H_0, respectively. Plots (c) and (d) show the probability of a positive outcome. $f_{01} = 1/4$ throughout.

does not result in highest power for either H_1 nor H_0. Similarly, setting t too large has a detrimental effect on the probability of a positive study, particularly in the case that $\theta_2 = 0$. We note that differences in power are generally quite small, and hence logistical reasons may drive the choice of interim analysis timing, rather than the desire for best power performance. Similar plots for HPP and CP were examined (not shown), and these exhibited the same characteristics as those described here in reference to the HUT procedure.

We also examined results for larger values of f_{01}. When $f_{01} \geq 1/2$, the difference between $t = 1/2$ and $3/4$ is smaller, but in both cases, power is always better

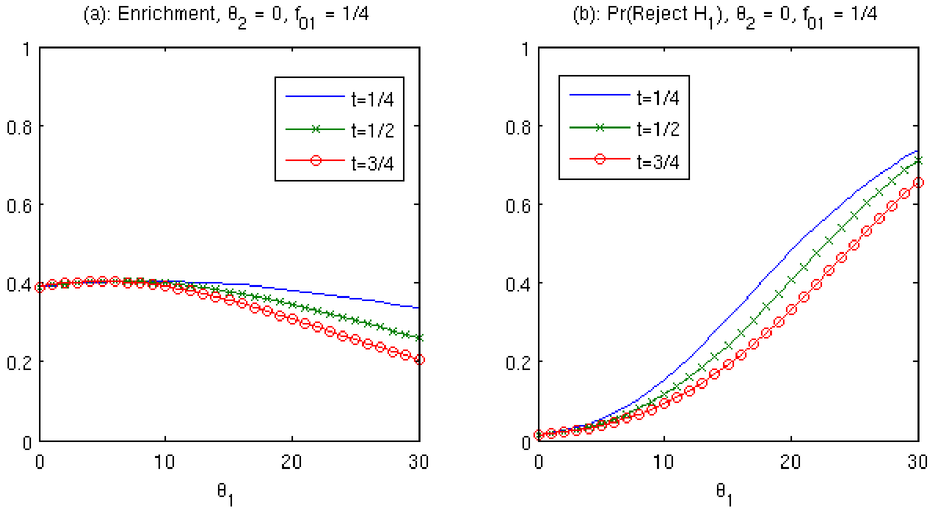


Figure 3.3: Empirical probabilities for FE. Plot (a) shows enrichment probability, and plot (b) shows power for H_1. Throughout, $\theta_2 = 0$ and $f_{01} = 1/4$.

than for $t = 1/4$. In fact, the difference between power for $t = 1/4$ and higher t is more pronounced in this setting. We conclude that procedures employing flexible adaptation rules (similar to those examined here) will likely benefit from using $t \approx 1/2$, though thorough simulations should always be conducted for the particular setting at hand.

It is also of interest to inspect the effect of t on FE. This procedure, while allowing for some adaptation at interim, is essentially fixed at the outset of the trial and rejection regions are known. Figure 3.3 shows enrichment and H_1 rejection probabilities when $\theta_2 = 0$ for $t = 1/4, 1/2$ and $3/4$. In plot (a) we see that, contrary to what we observed above, enrichment is more likely for small values of t. This can be explained by considering the design of FE: at the interim analysis, Z_{10} is checked and, if large enough, the trial proceeds with the full population. The non-centrality parameter of Z_{10}, $\mathbb{E}\left[Z_{10}\right] = w_{10}\lambda_0 = \theta_0\sqrt{t\mathcal{I}_{\max}}$, is increasing in t (if $\theta_1 > 0$), and hence a longer first stage will decrease the likelihood of enrichment, even when $\theta_2 = 0$. Plot (b) shows further consequences of this characteristic, as power for H_1 is at its largest when $t = 1/4$.

The observation made from Figure 3.3 may motivate a modification of the FE design. Indeed, we can add a "check" for Z_{12} before allowing the trial to proceed with the full population. However, an appealing characteristic of FE is its simplicity. Adding decision rules will make computation of the critical boundaries more complicated, in particular when the number of subgroups is increased. In addition, some care must be taken in the determination of how small a value for Z_{12} is "small enough." In Chapter 5, we propose a multi-stage design that only eliminates subgroups that show early signs of non-responsiveness. This design may offer a reasonable compromise to the weakness inherit in the FE design.

3.3.2 Influence of Subgroup Prevalence

It is clear that power for Ω_1 is increasing in f_{01} (Alosh and Huque, 2009). It is therefore of greater interest to inspect the effect that f_{01} has on power for Ω_0. Figure 3.4 shows rejection probabilities for various values of f_{01}, as obtained for the HUT procedure in our simulations. In plots (a) and (c), we see that for $\theta_2 = 0$, as f_{01} increases so does the probability that H_0 is rejected, as well as the probability of a positive outcome in the study. This is expected, since the effect is only positive in Ω_1 it naturally becomes easier to detect an overall effect if Ω_1 is large.

Plots (b) and (d) show that the same does not hold when $\theta_2 = 20$. Power for Ω_0 is now maximized when $f_{01} = 1/4$, as seen in plot (b). The reason is likely that a larger Ω_1 will dilute the estimate of θ_0 when θ_1 is small and $\theta_2 = 20$. In such cases, a small sample size from Ω_1 is actually beneficial. In plot (d) we see that, if θ_1 is much smaller than 20, a smaller subgroup is again better for the probability of a positive outcome. However, as θ_1 exceeds clinical significance, a larger subgroup becomes advantageous.

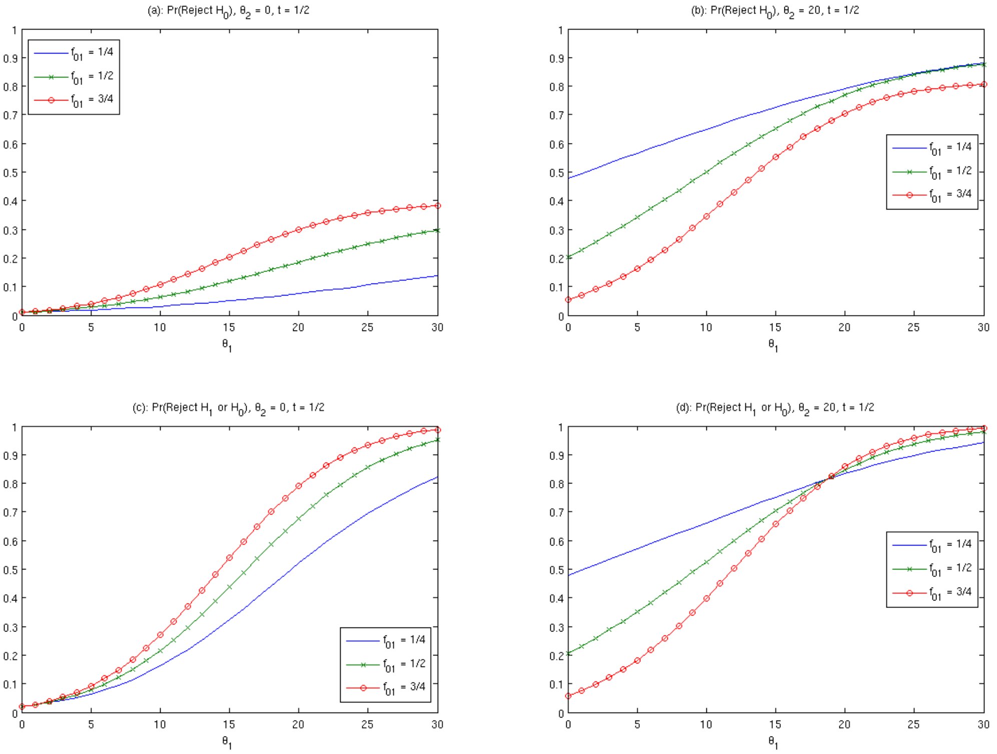


Figure 3.4: Empirical rejection probabilities for the HUT procedure. The left and right columns use $\theta_2 = 0$ and 20 respectively. Plots (a) and (b) show the probability of rejecting H_0. Plots (c) and (d) show the probability of a positive outcome. $t = 1/2$ throughout.

Figure 3.5 is set up as Figure 3.4, showing the same probabilities, only for FE. In plot (a) we see that, as expected, larger values of f_{01} increase the probability of rejecting H_0 when $\theta_2 = 0$. The reason for this is the design of the FE procedure, as explained in Section 3.3.1. When $\theta_2 = 0$, $\mathbb{P}[\text{Reject } H_0 \text{ or } H_1]$ also increases with f_{01}. Plots (b) and (d) show similar behavior as for the HUT procedure. That is, small f_{01} is beneficial when $\theta_1 < 20$, and large f_{01} is preferred for $\theta_1 \geq 20$.

In conclusion, we have observed that if $\theta_2 = 0$, then H_0 is likelier to be rejected if f_{01} is large. However, if $\theta_2 = 20$, a small f_{01} is results in better power performance for H_0. Examination of enrichment behavior reveals that in their current form, FE and CP are better suited to handle small subgroups. This can be fixed, for

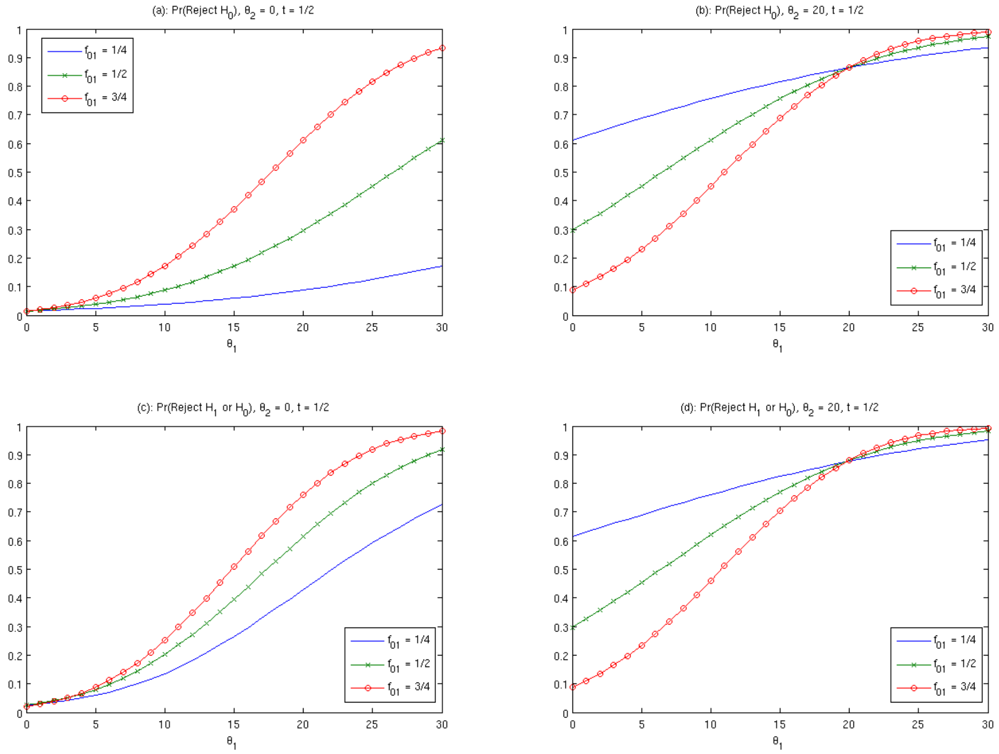


Figure 3.5: Empirical rejection probabilities for FE. The left and right columns use $\theta_2 = 0$ and 20 respectively. Plots (a) and (b) show the probability of rejecting H_0. Plots (c) and (d) show the probability of a positive outcome. $t = 1/2$ throughout.

instance by requiring that $[p_{10} \leq \gamma_0, p_{21} \leq \gamma_2]$ in order for FE to continue without enrichment. For CP, if $CP_{\Omega_0}(t, z_{10}, \hat{\delta}_{10}) \geq L_0^{CP}$, then we could force a check to see if $CP_{\Omega_2}(t, z_{12}, \hat{\delta}_{12})$ is large enough to warrant continuation without enrichment.

3.3.3 Interim Analysis Decisions

In this section we investigate the robustness of the adaptive designs to selecting the correct course of action for the second stage. Figure 3.6 shows empirical interim decision probabilities for FE, CP, HPP and HUT, where the left and right columns are based on $\theta_2 = 0$ and $\theta_2 = 20$ respectively, and $t = 1/2$. For example, when $f_{01} = 1/4$ and $(\theta_1, \theta_2) = (20, 0)$, we see from the left column that HUT will enrich,

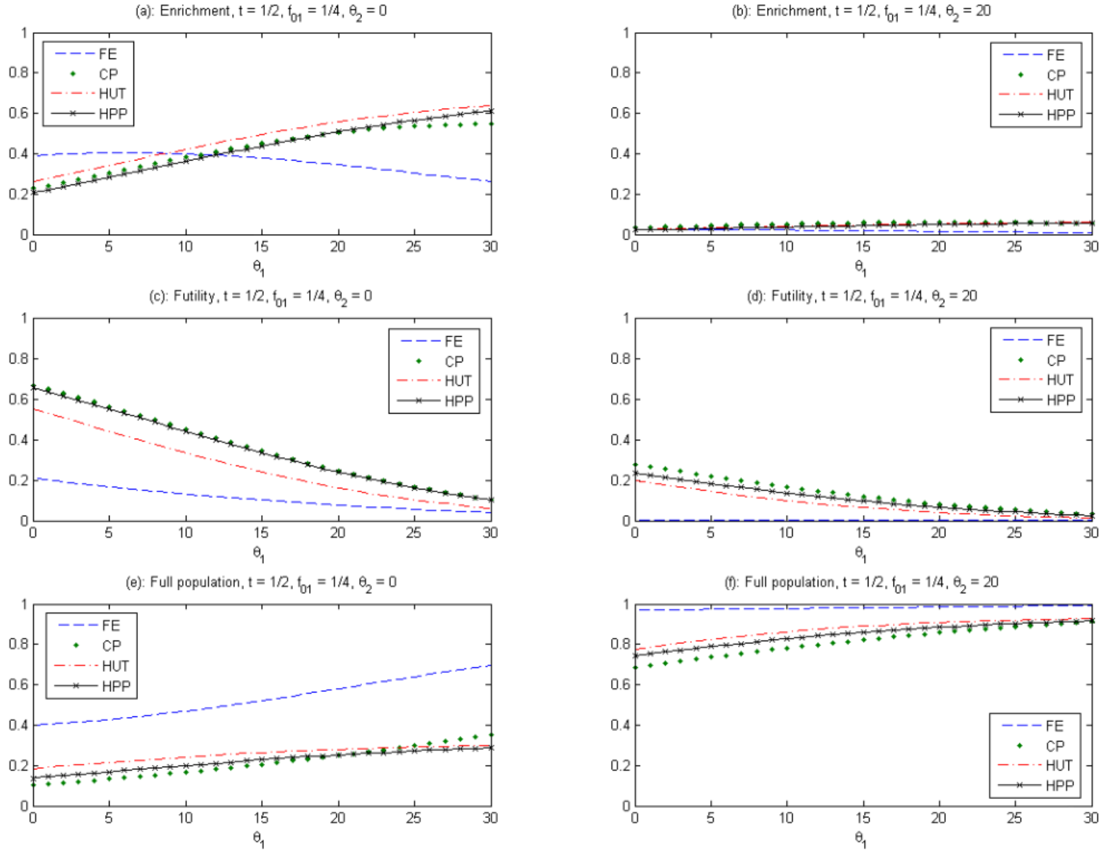


Figure 3.6: Empirical probabilities for interim analysis decisions. The left column corresponds to $\theta_2 = 0$, and the right column to $\theta_2 = 20$. Plots (a) and (b) show enrichment probabilities, plots (c) and (d) show the trial abandonment probabilities, and plots (e) and (f) show the probability that a study continues with the full population. $f_{01} = 1/4$ and $t = 1/2$ throughout.

stop for futility and continue with the full population roughly 56%, 16% and 28% of the time, respectively.

When $\theta_2 = 0$, and θ_1 is small, stopping for futility is the appropriate decision at interim. In plot (c), we see that CP and HPP have the highest likelihood of early stopping (66% when $\theta_1 = 0$) though, when $\theta_1 = 20$, they still stop for futility with probability 0.24. While these probabilities may seem high, we keep in mind that $f_{01} = 1/4$, so the number of observations accumulated for Ω_1 at interim is relatively low. When $\theta_2 = 20$, stopping for futility is much less likely. In that case, CP has the largest probability, 0.08.

In plot (a), we observe that HUT achieves the highest enrichment probability throughout, though HPP and CP are quite close. The enrichment probability for FE actually decreases when θ_1 increases, but this is due to the aforementioned design issue, whereby a large value for θ_1 starts to affect the estimate for θ_0. When $\theta_2 = 20$ (plot (b)), enrichment is very unlikely for all procedures, for all considered values of θ_1. At $\theta_1 = 20$, enrichment probability is < 0.05 for all procedures, which is desirable as the correct decision is now to proceed sampling from the full population.

When $\theta_2 = 0$ (plot (e)), CP, HPP and HUT are unlikely to proceed with the full population, while FE does this with a substantially higher probability. As said before, we might improve the interim decision making for FE by enforcing a check for efficacy in Ω_2. When $\theta_2 = 20$ (plot (f)), all procedures proceed with the full population with high probability. CP is slightly worse here than the other procedures, but the difference is very small.

Figure 3.7 shows the same plots as Figure 3.6, only with $f_{01} = 1/2$. The overall behavior is quite similar though there are a few things of note. First, when $\theta_2 = 0$, enrichment probability does not appear to change much compared to that for $f_{01} = 1/4$ (see plot (a), Figure 3.6). Intuitively, we might have expected that a larger Ω_1 would increase the chance of enrichment. A possible explanation is that higher prevalence of Ω_1 implies increased correlation between $\hat{\theta}_1$ and $\hat{\theta}_0$. We also note that, for $\theta_2 = 20$ and $\theta_1 < 20$, futility stops are a bit more common when $f_{01} = 1/2$ than for $f_{01} = 1/4$. Likewise, the probability to proceed with the full population is slightly decreased as compared to Figure 3.6. The effect is most noticeable when θ_1 is small; when $\theta_1 \geq 20$ futility probabilities are roughly the

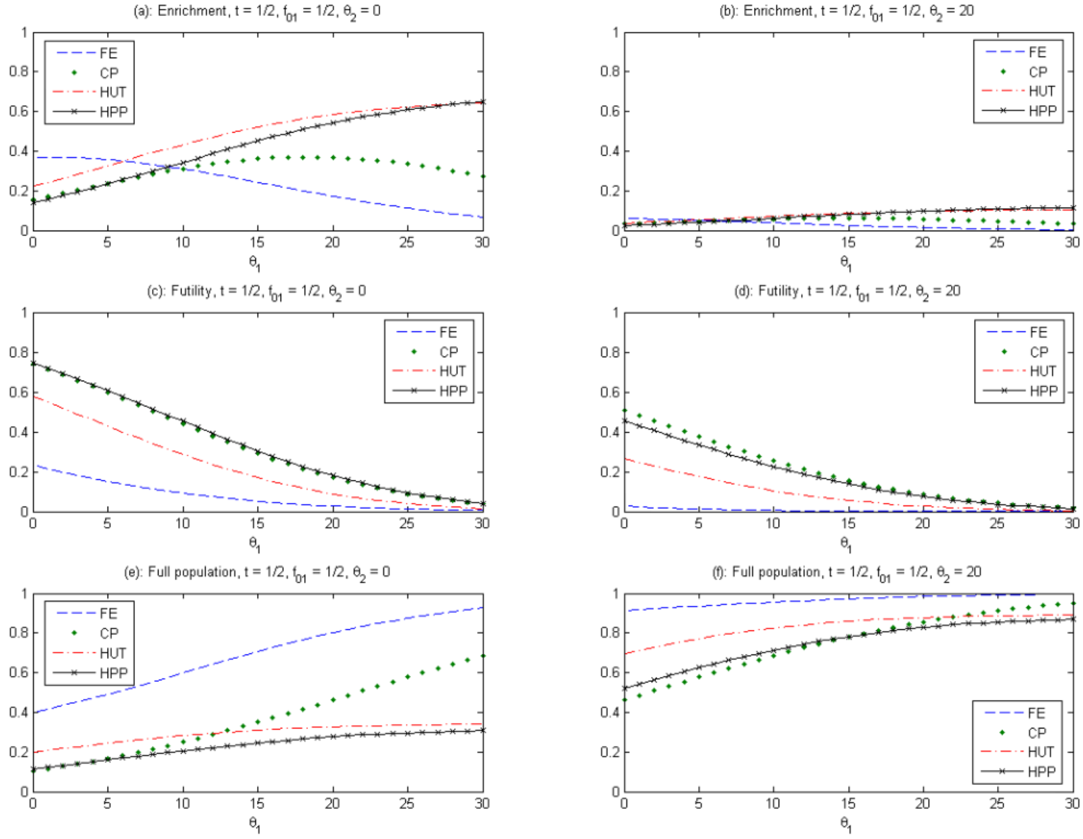


Figure 3.7: Empirical probabilities for interim analysis decisions. The left column corresponds to $\theta_2 = 0$, and the right column to $\theta_2 = 20$. Plots (a) and (b) show enrichment probabilities, plots (c) and (d) show the trial abandonment probabilities, and plots (e) and (f) show the probability that a study continues with the full population. $f_{01} = 1/2$ and $t = 1/2$ throughout.

same as in Figure 3.6.

Throughout this section, we chose the procedure-specific parameter values that appeared (in a rough analysis) to work best for that particular procedure. Overall, HUT and HPP perform better than CP and FE at the interim analysis, though the latter two procedures could be improved as previously remarked. There is no substantial difference in the performance of the two hybrid Bayesian methods. In general, our procedures perform reasonably well at the interim analysis, particularly in light of the fact that with the value chosen for $\mathcal{I}_{\text{total}}$, interim estimates of θ_j are quite variable (standard deviations of $\hat{\theta}_{10}$ and $\hat{\theta}_{11}$ approximately equal to

8.6 and 12.1, respectively).

3.3.4 Empirical Power

First, we examine the impact of using *empirical data weights* on Type I error. See Section 2.1.1 for discussion on these weights. Type I error probabilities were examined across all configurations for $\theta_1, \theta_2, f_{01}$ and t, as well as for all parameter values for the procedures in question (CP, HPP and HUT). When futility checks were enforced, Type I error was not greater than $\alpha = 0.025$ for any of the tested configurations. The largest Type I error probability was 0.0246 for HPP with $(\theta_1, \theta_2) = (0, 20)$, $f_{01} = 1/4$, $t = 1/4$, using procedure-specific parameters $L_0^{HPP} = 0.35$, $L_1^{HPP} = 0.2$ and $L_2^{HPP} = 0.1$. If, on the other hand, futility stops are removed, we see some increase in Type I error. For example, the CP method with $(\theta_1, \theta_2) = (0, 20)$ will reject H_1 erroneously with maximum probability 0.0264 (for our configurations). Wang et al. (2009) report similar results for empirical data weights and note that in general the Type I error inflation is essentially negligible. We do note that current FDA regulations insist on strict α-protection (US Food and Drug Administration, 1998).

Figure 3.8 shows empirical rejection probabilities for all procedures under consideration, with $f_{01} = 1/4$. Since FE and CP seem to favor small t, and HUT and HPP favor large t, we use $t = 1/3$. One-stage procedures are shown in blue, and two-stage procedures are shown in red. In plot (a), we see that the probability of rejecting H_1 is highest for the hybrid procedures HUT and HPP, as well as CP. FE is close, with a maximum difference appearing to be less than 5%. While one-stage procedures AFP and HS are expected to perform poorly relative to two-stage procedures, we note that the multiplicity adjusted OBFm procedure does just as

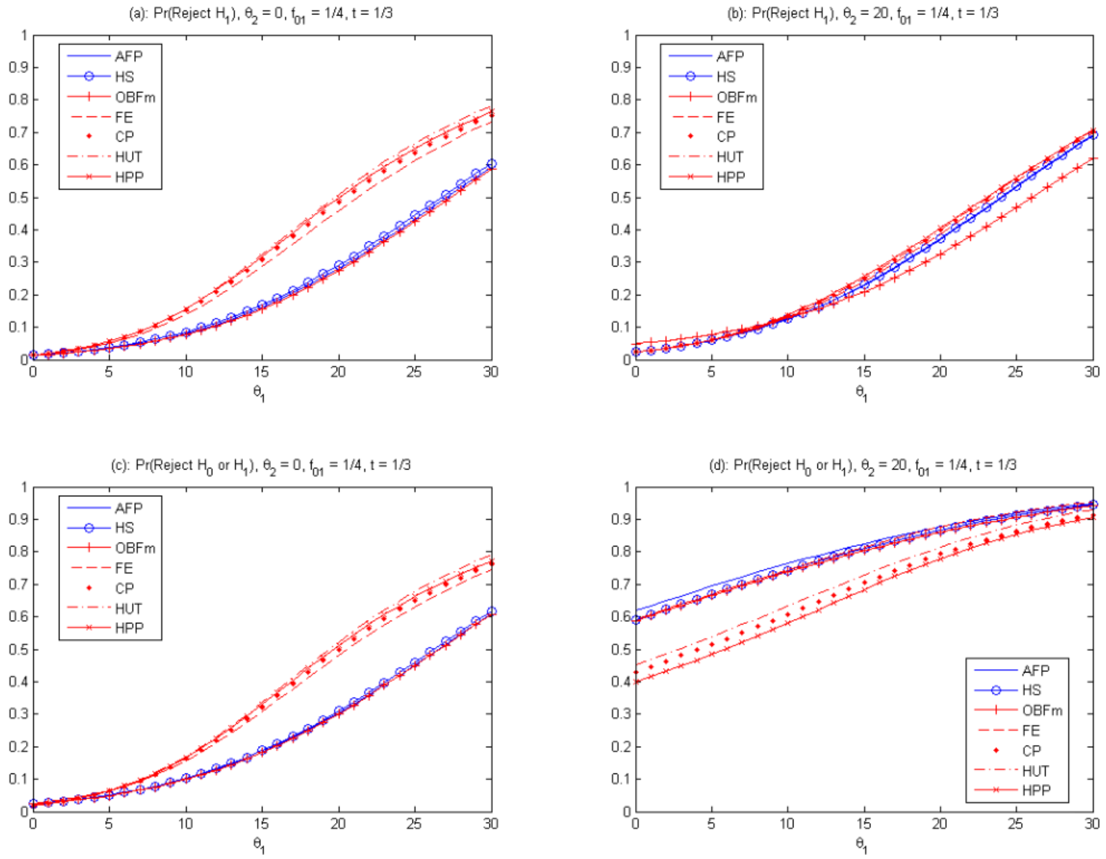


Figure 3.8: Empirical power for all procedures, using $f_{01} = 1/4$. One-stage procedures are shown in blue, and two-stage procedures in red. Plots (a) and (c) use $\theta_2 = 0$, and plots (b) and (d) use $\theta_2 = 20$. Throughout, $t = 1/3$.

poorly. This can partly be explained by the fact that test statistics Z_0 and Z_1 are correlated, which is known to make procedures based on p-value adjustments conservative. When $\theta_2 = 20$ (plot (b)), most procedures perform in a similar manner. Adaptive procedures are still more powerful than AFP and HS, though the difference is not great. For $\theta_1 \geq 10$, OBFm is dominated by all other procedures.

Plot (c) shows the probability of a *positive result* when $\theta_2 = 0$. Patterns are similar to those of plot (a); two-stage procedures, sans OBFm, dominate one-stage procedures. HUT and HPP seem to perform best, though CP and FE are quite close. Finally, plot (d) shows the probability of a positive result with $\theta_2 =$

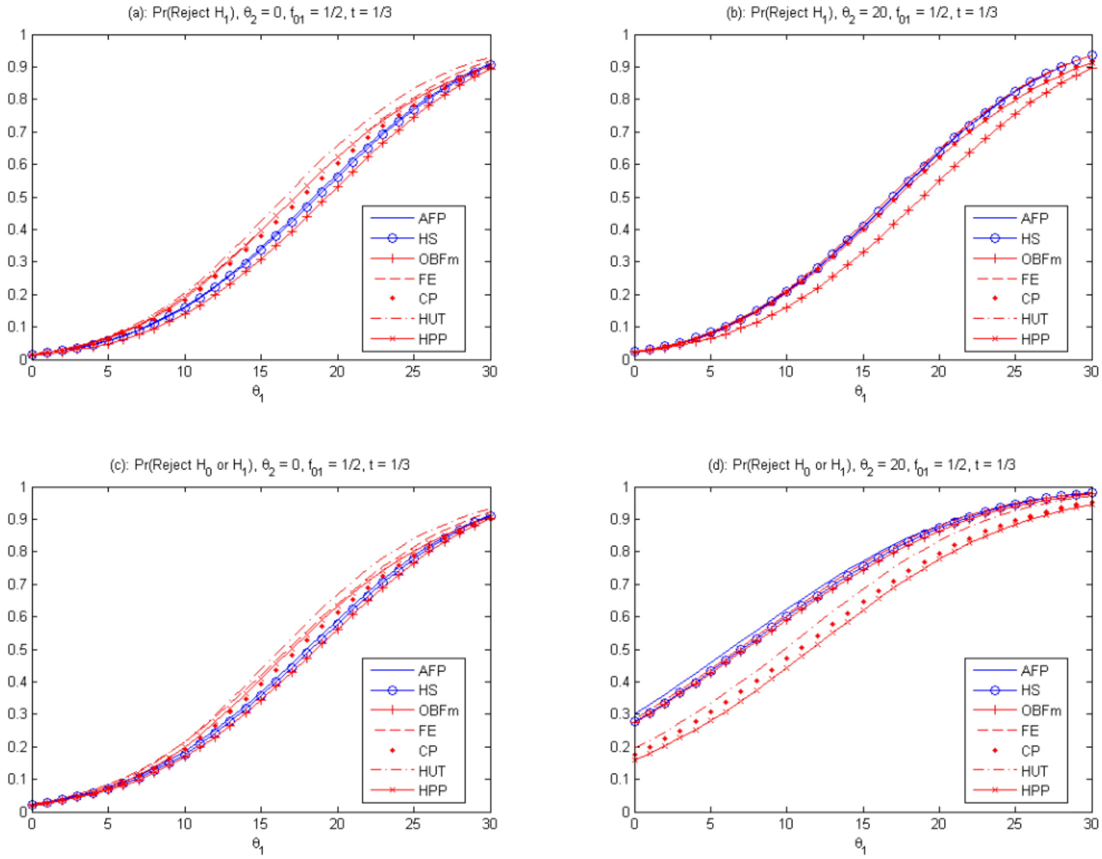


Figure 3.9: Empirical power for all procedures, using $f_{01} = 1/2$. One-stage procedures are shown in blue, and two-stage procedures in red. Plots (a) and (c) use $\theta_2 = 0$, and plots (b) and (d) use $\theta_2 = 20$. Throughout, $t = 1/3$.

20. We see that fallback-based procedures, AFP and FE, have highest rejection probabilities, while HS and OBFm are quite close. We note that CP, HPP and HUT do not compare favorably in this scenario, and they are dominated for all values of θ_1 that were considered. Out of the three, HUT appears to perform best.

We show results for $f_{01} = 1/2$ in Figure 3.9. In plots (a) and (c) we see similar results as when $f_{01} = 1/4$, though power curves are now closer. HUT still dominates other procedures when $\theta_2 = 0$, but FE and HPP are very close. We also note that OBFm is still dominated by all other procedures. Plots (b) and (d) show results for $\theta_2 = 20$. In (b), we see that AFP, FE, HS and HUT are most powerful, and the difference between these four procedures is essentially negligible. In plot

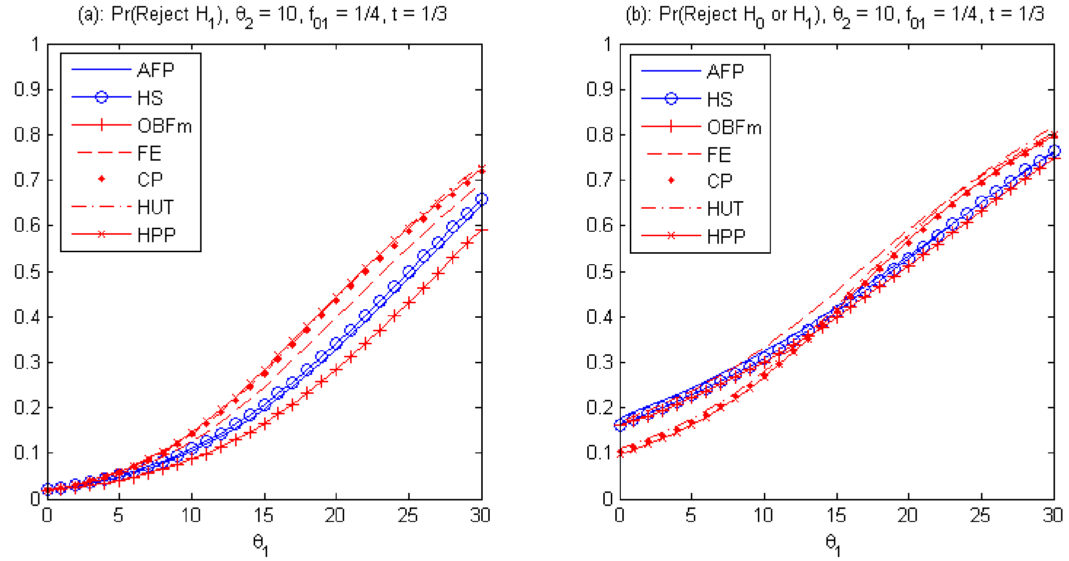


Figure 3.10: Empirical power for all procedures, using $f_{01} = 1/4$. One-stage procedures are shown in blue, and two-stage procedures in red. Plot (a) shows the probability that H_1 is rejected, while plot(b) shows the probability that the study yields a positive result. Here, $\theta_2 = 10$ and $t = 1/3$.

(d) we see that rejection probabilities for either H_0 or H_1 are highest for AFP and FE across all values of θ_1. As when $f_{01} = 1/4$, HUT, HPP and CP are dominated in this scenario, though HUT comes the closest to matching other procedures.

To summarize, adaptive methods such as HUT, HPP and CP perform well when there is only an effect in the subgroup Ω_1. FE can get very close, in particular when t is chosen to be reasonably small (e.g. 1/4 or 1/3). When $\theta_2 = 20$, adaptive methods such as HUT, HPP and CP do not perform as well, and AFP and FE seem to achieve highest probability of rejecting H_0 (or H_1). The gain (in power) for adaptive methods is greater when the subgroup is smaller, as evidenced by comparing Figures 3.8 and 3.9. Finally, in Figure 3.10, we include two plots of empirical power when $\theta_2 = 10$. For these plots, we use $f_{01} = 1/4$ and, as before, $t = 1/3$. In plot (a), which shows rejection probabilities for H_1, HUT, HPP and CP are roughly equal, and perform better than other procedures. Plot (b) shows the probability of a positive result (reject H_0 or H_1), and in this case, FE achieves the best performance for most values of θ_1. As θ_1 increases, adaptive procedures achieve performance roughly equal to FE.

In the scenarios analyzed above, it appears that HUT and HPP can be quite useful for detecting effects in a small subgroup. However, when there is an overall positive effect, FE outperforms other methods. No procedure dominates all others across all scenarios, so experimenters must choose a procedure appropriate for the specifics of a clinical trial. FE is appealing due to its transparency and, to a certain degree, simplicity. It performs very well when there is a positive overall effect, and can still detect effects only present in the subgroup. HUT and HPP are more powerful when only a subgroup effect is present, but suffer when θ_2 is close to clinical significance. Of the two, HUT may be preferable as it requires very little parameter tuning. In particular, if there is more than one subgroup it can be difficult to choose appropriate values for L_i^{HPP}, whereas HUT just requires specification of k^{HUT} and θ^+.

We conclude this chapter by summarizing the six questions that were raised in the beginning of Section 3.3, along with lessons learned in our numerical analysis.

1. What is the influence of the prevalence of Ω_1 in Ω_0 (f_{01}) on rejection probabilities for H_0 and H_1?

   As we expected, power performance for Ω_1 increases with f_{01}. If there is no effect in the subgroup complement, i.e. $\theta_2 = 0$, then probability of rejecting H_0 increases with f_{01}. When there is a significant effect in Ω_2, i.e. $\theta_2 = 20$, then rejection probabilities for H_0 decrease with f_{01}.

2. How important is the timing of an interim analysis? Does it affect H_0 and H_1 similarly, or is there a trade-off?

   Correct decisions at interim are made with greater consistency as t increases.

Setting t large (e.g. equal to 3/4) yields high power for H_0, but results in reduced power performance for H_1. The difference in power for H_1 depending on t is not great, and logistical reasons may drive the choice for interim timing, rather than desired "optimal" power performance.

3. For what types of values of (θ_1, θ_2) are adaptive designs preferable to fixed designs, and vice-versa?

   The FE procedure, as well as adaptive procedures CP, HUT and HPP are useful for detecting a treatment effect that is confined to the subgroup Ω_1. When $\theta_2 = 0$, these designs have substantially higher power for H_1 than the fixed procedures we considered. The multiplicity-adjusted OBF design does poorly in this setting. When both subgroups are effective however, fixed designs achieve highest power, and the FE procedure is close. Adaptive designs see a power reduction on the order of 5–10%.

4. How do procedures specifically tailored to subgroup testing perform compared to standard p-value adjustment procedures, or group sequential methods?

   The FE procedure performs quite well overall (for both H_0 and H_1), and can be improved as has been discussed. This improvement will likely reduce FE power for H_0. The AFP has low power to detect an effect only existing in one subgroup, but does well when there is an overall effect. As mentioned in the previous point, CP and hybrid Bayesian designs are primarily strong when $\theta_2 = 0$.

5. How robust are the adaptive designs to selecting the correct course of action at the interim analysis?

   Qualitatively, the utility-based design (HUT) was found to make correct decisions at interim with higher probability than the other two-stage designs.

HPP also performed quite well at interim, but CP and FE were seen to favor the overall population when θ_1 was large and $\theta_2 = 0$. These issues – which can be fixed – suggest that the decision to restrict sampling should be based on apparent lack of efficacy in particular subgroups, rather than an overall statistic. In Chapter 5, we propose an adaptive group sequential procedure that takes this approach.

6. For an adaptive design, are transparency and ease of exposition preferable to a seemingly well performing “black box?”

   In general, procedures that are easy to explain and implement are used with greater frequency in practice (Dmitrienko et al., 2010, Ch. 2.6). Hence, procedures such as FE or HPP have real merit. For HPP in particular, predictive probabilities can be intuitively explained to participants that are not statistically inclined so interim decisions will be easy to understand. On the other hand, the HUT procedure is likely difficult to explain, and “blindly” following a utility function at interim can make participants nervous if they resulting course of action seems unintuitive. Carefully determining the loss/gain function so as to reflect practical implications of the decisions that can be made might address these concerns. Doing so would be desirable, as the HUT design was seen to perform relatively well in comparison to our other procedures.

Chapter 4

Multi-Subgroup Procedures

In this chapter, we extend some of the procedures introduced in Chapter 3 to allow consideration of any number of subgroups. Reviewing notation introduced in Chapter 2, let $\Omega_0, \Omega_1, \ldots, \Omega_\ell$, $\ell > 1$, denote the populations of interest, and $\Omega_j \subsetneq \Omega_0$ for all $j = 1, \ldots, \ell$. For each Ω_j, we have a corresponding null hypothesis $H_j : \delta_j \leq 0$. The subgroups are not necessarily disjoint, and some may completely contain others. Procedures outlined in Sections 4.2 and 4.3 take place over two separate stages. Let $\mathcal{P}_i$ denote the index set of populations under consideration for stage i. Thus, $\mathcal{P}_1 = \{0, 1, \ldots, \ell\}$ and $\mathcal{P}_2 \subseteq \mathcal{P}_1$.

Recall that f_{ij} was defined as the prevalence of Ω_j in Ω_i. Now, for $\mathcal{S} \subseteq \mathcal{P}_i$, and

$$\Omega_{\mathcal{S}} := \bigcup_{j \in \mathcal{S}} \Omega_j,$$

we define $f_{\mathcal{S}}$ as the prevalence of $\Omega_{\mathcal{S}}$ in Ω_0. If $0 \in \mathcal{S}$, then $f_{\mathcal{S}} = 1$. Finally, we note that for $j \in \mathcal{P}_2$, the prevalence of Ω_j during stage 2 is $f_{0j}/f_{\mathcal{P}_2}$.

4.1 Adjusted Fallback Procedure (One Stage)

We suppose that the $H_j, j = 0, \ldots \ell$ are naturally ordered in the sense that we wish to establish efficacy for H_0 first, then H_1 and so on. This ordering is imposed *a priori*, i.e. before any data analysis takes place. Define local significance levels $\alpha_j \in [0, 1]$ for $j = 0, \ldots, \ell$, requiring that

$$\alpha = \sum_{j=0}^{\ell} \alpha_j$$

where α is the desired FWER. As in the one-subgroup case, we also require *adjusted* significance levels $\tilde{\alpha}_j, j = 1, \ldots, \ell$ ($\tilde{\alpha}_0 = \alpha_0$). For now, we take these as known, and computational details are given in Section 4.1.1. Let T_j denote the test statistic corresponding to H_j, and let p_j be its associated p-value. Denote $\rho_{ij} = \text{Corr}(T_i, T_j)$ for $0 \leq i, j \leq \ell$. As discussed in Chapter 1, we assume treatment effects in opposite directions do not occur, so $\rho_{ij} \geq 0$ for all i, j. The general procedure is given in Algorithm 1. Note that in the original fallback procedure (Wiens, 2003), H_i is tested at level α_{i-1} if H_{i-1} was not rejected (if $m_i^* = i - 1$). Hence, if $\tilde{\alpha}_i \geq \alpha_i$ we have uniform improvement in power to reject H_i. See Section 4.1.1 for further details.

```
Input: $\alpha_i$, $\tilde{\alpha}_i$ and $p_i$ for $i = 0, \ldots, \ell$
$\alpha'_0 = \alpha_0$
if $p_0 \leq \alpha'_0$ then
    Reject $H_0$
end
for $i = 1$ to $\ell$ do
    if $p_j \leq \alpha'_j$ for all $j = 0, 1, \ldots, i-1$ then
        $m_i^* = -1$
    else
        $m_i^* = \max\left\{j \in \{0, \ldots, i-1\} : p_j > \alpha'_j\right\}$
    end
    $\alpha'_i = \begin{cases} \sum_{k=m_i^*+1}^{i} \alpha_k & \text{if } m_i^* < i-1 \\ \tilde{\alpha}_i & \text{if } m_i^* = i-1 \end{cases}$
    if $p_i \leq \alpha'_i$ then
        Reject $H_i$
    end
end
```

Algorithm 1: Testing algorithm for the adjusted fallback procedure.

Example 4.1. *Suppose we have one subgroup with $\alpha_0 = 0.02$ and $\alpha_1 = 0.005$. Further, $\tilde{\alpha}_1 = 0.012$. Suppose we observe p-values $p_0 = 0.025$ and $p_1 = 0.01$. The original fallback procedure rejects neither hypothesis ($p_0 > 0.02$ and $p_1 > 0.005$), but the AFP does reject H_1 since $p_2 = 0.01 < 0.012 = \tilde{\alpha}_2$.*

4.1.1 Obtaining $\tilde{\alpha}_i$

In this section, we outline how to obtain the adjusted levels $\tilde{\alpha}_i$, given values of ρ_{ij} for $1 \leq i, j \leq \ell$. The idea is to exploit the correlation structure in order to test H_i at a level higher than α_i in the case that H_{i-1} was not rejected. Recall that, in Section 3.2.1 we saw that for $\ell = 1$, $\tilde{\alpha}_1$ may be easily obtained. When $\ell = 2$, $\tilde{\alpha}_2$ can also be derived analytically but more work is involved, and the algebra becomes increasingly tedious as more subgroups are added. A more succinct way of specifying computational details for $\tilde{\alpha}_i$ is therefore desirable. We propose using a so-called *decision matrix*, used by Wiens and Dmitrienko (2005) to define closed testing procedures.

Definition 4.1. *Given an intersection hypothesis H, define the $(\ell+1)$-dimensional decision vector $\boldsymbol{v}(H)$ as*

$$\boldsymbol{v}(H) = (v_0(H), v_1(H), \ldots, v_\ell(H)),$$

for given $v_i(H) \in [0, 1]$, $i = 0, 1, \ldots, \ell$. Next, for a collection of intersection hypotheses $\{H_j\}$, $j = 1, \ldots, J$, define the $J \times (\ell + 1)$ dimensional decision matrix $\boldsymbol{V}$ as

$$\boldsymbol{V} = \begin{pmatrix} \vdots \\ \boldsymbol{v}(H_j) = (v_0(H_j), \ldots, v_\ell(H_j)) \\ \vdots \end{pmatrix}.$$

Definition 4.2. *Given a decision matrix $\boldsymbol{V}$, a particular intersection hypothesis H_j is tested as follows. Obtain the corresponding decision vector $\boldsymbol{v}(H_j)$, and for each elementary hypothesis a p-value p_i, $i = 0, \ldots, \ell$. If there exists $i \in \{0, 1, \ldots, \ell\}$ such that $p_i \leq v_i(H_j)$, then H_j is rejected.*

Clearly, one may use a decision matrix to carry out a closed testing procedure: reject the elementary hypothesis H_i if each H containing H_i was rejected, using the approach given in Definitions 4.1 and 4.2. If each intersection hypothesis is tested at level α, that is, if for each H,

$$\mathbb{P}\left[\exists i \in \{0, 1, \ldots, \ell\} \text{ such that } p_i \leq v_i(H) \mid H\right] \leq \alpha,$$

then Theorem 2.1 implies that FWER is controlled strongly at level α. We next give a simple example to illustrate how a decision matrix is defined, and how it may be used to test naturally ordered hypotheses.

Example 4.2. *Suppose $\ell = 1$, and we are therefore testing the two hypotheses H_0 and H_1. We let $\alpha_0 = 0.02$ and $\alpha_1 = 0.005$ so $\alpha_1 + \alpha_2 = 0.025$. With two hypotheses, we need to test the intersection hypotheses $H_{11} = H_0 \cap H_1$, $H_{10} = H_0$ and $H_{01} = H_1$. Applying the closure principle, H_0 is rejected if and only if H_{11} and H_{10} are rejected. Likewise, to reject H_1, we must reject H_{11} and H_{01}. Using the specified local significance levels, the decision matrix is given as*

$$\boldsymbol{V} = \begin{pmatrix} 0.02 & 0.005 \\ 0.02 & 0 \\ 0 & 0.025 \end{pmatrix}.$$

Hence, H_{11} is rejected if either $p_0 \leq 0.02$ or $p_1 \leq 0.005$, H_{10} is rejected if $p_0 \leq 0.02$, and H_{01} is rejected if $p_1 \leq 0.025$. If $p_0 = 0.015$ and $p_1 = 0.021$, then both H_0 and H_1 are rejected. However, if $p_0 = 0.021$ and $p_1 = 0.004$ then only H_1 can be rejected. In the proof of Theorem 4.1, we use the decision matrix to show that applying the closure principle is equivalent to the adjusted fallback procedure.

We now explain how the $\tilde{\alpha}_i$ are obtained, proceeding inductively. Suppose $\ell = 1$. The intersection hypotheses are then $H_{11} = H_0 \cap H_1$, $H_{10} = H_0$ and

$H_{01} = H_1$. We have

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{v}(H_{11}) \\ \boldsymbol{v}(H_{10}) \\ \boldsymbol{v}(H_{01}) \end{pmatrix} = \begin{pmatrix} \alpha_0 & \tilde{\alpha}_1 \\ \alpha_0 & 0 \\ 0 & \alpha_0 + \alpha_1 \end{pmatrix}.$$

Here, $\tilde{\alpha}_1$ is set such that $\mathbb{P}[\text{Reject } H_{11} | H_{11}] = \alpha_0 + \alpha_1$. Note that this requires distributional assumptions for T_0 and T_1, as well as a given value for ρ_{01}. Also note that the value for $\tilde{\alpha}_1$ is exactly that of Section 3.2.1, i.e. for the one-subgroup AFP. Since $\ell = 1$, $\alpha_0 + \alpha_1 = \alpha$, and this defines a closed testing procedure. In the proof of Theorem 4.1, we show that this closed testing procedure is equivalent to the AFP.

Now, let $\ell \geq 2$. We have $\alpha_0, \ldots, \alpha_\ell$ and ρ_{ij} for all $0 \leq i, j \leq \ell$. We have seen how to obtain $\tilde{\alpha}_1$. Suppose further, that we have $\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_j$ for $1 \leq j < \ell$. Then $\tilde{\alpha}_{j+1}$ is obtained as follows. There are 2^{j+1} existing intersection hypotheses (obtained from $H_0, H_1, \ldots, H_j$, and counting $H_{00\cdots 0}$). Each such hypothesis will be part of two new intersection hypotheses; one will contain H_{j+1}, and one will not. For those that do not contain H_{j+1}, $\boldsymbol{v}(H)$ will be as before, only adding a zero in the $(j+1)th$ place. Denote the intersection hypotheses that contain H_{j+1} as $\tilde{H}_1, \ldots, \tilde{H}_K$, where $K = 2^{j+1}$. For $i = 1, \ldots, K$, let m_i^* denote the largest number of an elementary hypothesis in $\tilde{H}_i$, that is smaller than $j + 1$. Then, for each of the $\tilde{H}_i$, the $(j+1)th$ slot of $\boldsymbol{v}(\tilde{H}_i)$ is set as

$$v_{j+1}\left(\tilde{H}_i\right) = \begin{cases} \sum\limits_{k=m_i^*+1}^{j+1} \alpha_k & \text{if } m_i^* < j \\ \xi_i & \text{if } m_i^* = j \end{cases}$$

Each ξ_i is determined such that

$$\begin{aligned} \mathbb{P}\left[\text{Reject } \tilde{H}_i \mid \tilde{H}_i\right] &= \mathbb{P}\left[\exists k \in \{0, 1, \ldots, j+1\} \text{ such that } p_k \leq v_k(\tilde{H}_i) \mid \tilde{H}_i\right] \\ &= \sum_{k=0}^{j+1} \alpha_k. \end{aligned}$$

In Section 4.1.3, we outline a computational approach using iterated integrals as developed by Armitage et al. (1969), which allows for easy computation of ξ_i even if many subgroups are involved. Finally, each ξ_i is replaced with $\tilde{\alpha}_{j+1} := \min\limits_{i=1,\dots,K} \xi_i$, which guarantees that

$$\sup_{i=1,\dots,K} \mathbb{P}\left[\text{Reject } \tilde{H}_i \mid \tilde{H}_i\right] \leq \sum_{k=1}^{j+1} \alpha_k.$$

As the original fallback procedure can be stated using decision matrices without adjusted α-levels, and since it controls FWER strongly, we note that all adjusted α levels $\tilde{\alpha}_j$ will be at least as large as their unadjusted counterparts α_j. Hence, the adjusted fallback procedure attains uniform improvement in power over the original fallback procedure.

4.1.2 Control of the FWER

The main result of this section is Theorem 4.1, which proves that the adjusted fallback procedure is equivalent to a closed testing procedure (CTP). Since CTPs control the FWER strongly (i.e. under any null hypothesis configuration), the same holds for the adjusted fallback procedure. The proof of Theorem 4.1 follows closely the proof given by Wiens and Dmitrienko (2005), with additional arguments where necessary due to differences in the two procedures.

Theorem 4.1. *The adjusted fallback procedure, defined in Algorithm 1, is a closed testing procedure.*

Proof: Fix $\ell \geq 1$. We will show that the AFP rejects an arbitrary elementary hypothesis H_j, $j = 0, 1, \dots, \ell$, if and only if a CTP rejects H_j. The proof is by induction in j. We are given $\alpha_0, \dots, \alpha_\ell$, and $\tilde{\alpha}_1, \dots, \tilde{\alpha}_\ell$ have been obtained using the method outlined in Section 4.1.1. The case of $j = 0$ is trivial, so consider the case of $j = 1$, and the resulting intersection hypotheses H_{11}, H_{10} and H_{01}. The three decision vectors are

$$\begin{aligned} \boldsymbol{v}(H_{11}) &= (\alpha_0, \tilde{\alpha}_1) \\ \boldsymbol{v}(H_{10}) &= (\alpha_0, 0) \\ \boldsymbol{v}(H_{01}) &= (0, \alpha_0 + \alpha_1). \end{aligned}$$

The AFP rejects H_0 if $p_0 \leq \alpha_0$. In a CTP, $p_0 \leq \alpha_0$ implies H_{11} and H_{10} are rejected, and hence H_0 is rejected. If AFP rejects H_1 there are two cases to consider:

(i) AFP rejects H_1 with $p_0 \leq \alpha_0$ and $p_1 \leq \alpha_0 + \alpha_1$. In this case, a CTP rejects H_{11} because $p_0 \leq \alpha_0$, while H_{01} is rejected as $p_1 \leq \alpha_0 + \alpha_1$. Hence H_1 is also rejected by a CTP.

(ii) AFP rejects H_1 with $p_0 > \alpha_0$ and $p_1 \leq \tilde{\alpha}_1$. In this case, a CTP rejects both H_{11} and H_{01} as $p_1 \leq \tilde{\alpha}_1$.

We also see, with similar ease, that the AFP and a CTP will accept H_0 and H_1 under the same circumstances. Hence we have proved that the AFP and a CTP will reach the same conclusion when two hypotheses are considered.

Induction Hypothesis: Suppose now that for j hypotheses, $1 < j < \ell$, the AFP corresponds exactly to the decision rule defined by using $\boldsymbol{V}$ (a CTP). That is, for any $i \leq j$, AFP rejects H_i if and only if a CTP rejects H_i.

We now add hypothesis H_{j+1} to consideration, possibly H_ℓ. We must consider three new types of intersection hypotheses: The ones that contain H_{j+1} among others, the ones that do not contain H_{j+1}, and that consisting solely of H_{j+1}. We need to show that rejections (and non-rejections) made using the decision matrix $\boldsymbol{V}$ for all elementary hypotheses $H_0, H_1, \ldots, H_{j+1}$ correspond exactly to those made by the adjusted fallback procedure. Note that adding H_{j+1} does not change the value of $v_i(H)$ for $i \leq j$, nor does it affect the AFP for those hypotheses. Therefore, we need only consider two cases: AFP rejects H_{j+1}, and AFP accepts H_{j+1}. The induction hypothesis assumes equivalence for all intersection hypotheses not containing H_{j+1}, so we focus only on those that do contain H_{j+1}.

Suppose that the AFP does not reject H_{j+1}. Let

$$H = \left(\bigcap_{i \leq j \,:\, p_i > \alpha'_i} H_i \right) \cap H_{j+1},$$

so H is the intersection hypothesis that contains H_{j+1}, as well as all elementary hypotheses before H_{j+1} that were not rejected by the AFP, given our natural ordering of hypotheses. Next, let H_i be an arbitrary hypothesis in H, and let $i^* = \max\{k = 0, 1, \ldots, i-1 : H \text{ contains } H_k\}$ such that H_{i^*} is the last hypothesis in H, before H_i. Note that if $i = 0$, then we set $i^* = -1$ and all steps below still hold. Further, note that H_{i^*} was not rejected by the AFP. Now,

$$\begin{aligned} &\text{if } i^* + 1 = i \text{ then } v_i(H) = \tilde{\alpha}_i; \\ &\text{else, if } i^* + 1 < i \text{ then } v_i(H) = \sum_{k=i^*+1}^{i} \alpha_k. \end{aligned}$$

Since H_{i^*} was the last hypothesis before H_i that was not rejected, the levels used by the AFP were $\alpha'_i = \tilde{\alpha}_i$ if $i^*+1 = i$, and $\alpha'_i = \sum_{k=i^*+1}^{i} \alpha_k$ if $i^*+1 < i$. In either case, $v_i(H) = \alpha'_i$ and hence $v_i(H) = \alpha'_i$ for all H_i contained in H. Since the AFP did not reject any of those hypotheses (by the definition of H), we know that $p_i > \alpha'_i$ for all H_i in H. Hence $p_i > v_i(H)$ for all i, so H is not rejected. In order to reject H_{j+1}, a CTP needs to reject all intersection hypotheses containing H_{j+1}, and since H was not rejected, a CTP does not reject H_{j+1}.

Suppose now that the AFP rejects H_{j+1}. Let H be an arbitrary intersection hypothesis containing H_{j+1}, and let $k^* = \max\{k = 0, 1, \ldots, j : p_k > \alpha'_k\}$, so H_{k^*} is the last hypothesis before H_{j+1} that was not rejected by the AFP. If all were rejected, then set $k^* = -1$ (note that H need not contain H_{k^*}). Next, let $k^{**} = \min\{k = k^* + 1, \ldots, \ell : H \text{ contains } H_k\}$ so $H_{k^{**}}$ is the first hypothesis after H_{k^*} that is contained in H. If necessary, $H_{k^{**}} = H_{j+1}$, so $H_{k^{**}}$ always exists. Also note that the definitions of H_{k^*} and $H_{k^{**}}$ imply that $H_{k^{**}}$ was rejected by the AFP. Next let $m^* = \max\{k = 0, 1, \ldots, k^{**} - 1 : H \text{ contains } H_k\}$, so H_{m^*} is the last elementary hypothesis in H, before $H_{k^{**}}$. Then,

$$v_{k^{**}}(H) = \begin{cases} \sum\limits_{i=m^*+1}^{k^{**}} \alpha_i & \text{if } m^* + 1 < k^{**} \\ \tilde{\alpha}_{k^{**}} & \text{if } m^* + 1 = k^{**} \end{cases}$$

Also, recalling that H_{k^*} was the last hypothesis before $H_{k^{**}}$ that was not rejected, we have

$$\alpha'_{k^{**}} = \begin{cases} \sum\limits_{i=k^*+1}^{k^{**}} \alpha_i & \text{if } k^* + 1 < k^{**} \\ \tilde{\alpha}_{k^{**}} & \text{if } k^* + 1 = k^{**} \end{cases}$$

Now, H_{m^*} is the last hypothesis in H before $H_{k^{**}}$, and $H_{k^{**}}$ is the first hypothesis in H after H_{k^*}. If H_{k^*} is in H, then $H_{m^*} = H_{k^*}$, and if not, then $m^* < k^*$. Hence $m^* \leq k^*$ and $v_{k^{**}}(H) \geq \alpha'_{k^{**}}$. If $k^{**} = j + 1$, then by assumption $H_{k^{**}}$ is rejected by the AFP. Hence $p_{j+1} \leq \alpha'_{k^{**}} \leq v_{k^{**}}(H)$, so H is rejected. On the other hand, if $k^{**} < j + 1$, then the induction hypothesis implies that $H_{k^{**}}$ was rejected by a CTP for the first j hypotheses. Hence, $p_{k^{**}} \leq v_{k^{**}}(H)$ and H is rejected. Since H was arbitrary, and contains H_{j+1}, a CTP will reject H_{j+1}. This concludes the proof. ■

4.1.3 Computational Details

In this section we explain a computationally convenient way to obtain adjusted levels $\tilde{\alpha}_k$ for the adjusted fallback procedure. Following the steps given in Section 4.1.1, obtaining the adjusted level $\tilde{\alpha}_k$ for the elementary hypothesis H_k involves computing a k-dimensional multivariate normal integral. Evaluating such integrals can be challenging, particularly in higher dimensions. However, if a certain structure is imposed, e.g. nested or disjoint subgroups, the computation may be reduced to evaluating $k-1$ univariate normal integrals. This is accomplished by use of recursive formulae in the style of Armitage et al. (1969).

We have $\ell+1$ populations, the overall population Ω_0 and the subgroups $\Omega_1, \dots, \Omega_\ell$. We assume that these populations are nested, i.e. $\Omega_i \supseteq \Omega_j$ when $1 \leq i \leq j \leq \ell$. Associated with each of these populations is an elementary null hypothesis (lower-sided) H_i, and a local significance level $\alpha_i \in [0,1]$, $i = 0, \dots, \ell$. Note that $\sum_i \alpha_i = \alpha$ where α is the desired FWER. From Section 3.2.1 we know how to obtain $\tilde{\alpha}_1$. Suppose then that we have obtained adjusted levels $\tilde{\alpha}_i$ for $i = 1, \dots, k-1 < \ell$ and need to compute $\tilde{\alpha}_k$. We detail how the computation is carried out for the hypothesis $H_0 \cap \cdots \cap H_k$; the basic technique is the same for other intersection hypotheses.

We will find ξ to solve

$$\begin{aligned}
\sum_{i=0}^{k} \alpha_i &= \mathbb{P}\left[p_0 \leq \alpha_0 \text{ or } p_1 \leq \tilde{\alpha}_1 \text{ or } \dots p_{k-1} \leq \tilde{\alpha}_{k-1} \text{ or } p_k \leq \xi \;\middle|\; \bigcap_{i=0}^{k} H_i \right] \\
&= 1 - \mathbb{P}\left[T_0 < C_{\alpha_0}, T_1 < C_{\tilde{\alpha}_1}, \dots, T_{k-1} < C_{\tilde{\alpha}_{k-1}}, T_k < C_\xi \;\middle|\; \bigcap_{i=0}^{k} H_i \right] \\
&=: 1 - \zeta_k(\alpha_0, \tilde{\alpha}_1, \dots, \tilde{\alpha}_{k-1}, \xi; \boldsymbol{\theta} = \mathbf{0})
\end{aligned}$$

where $\boldsymbol{\theta}$ and $\mathbf{0}$ are $\mathbb{R}^{k+1}$ vectors. $\boldsymbol{\theta}$ is the mean vector of $[T_0, \dots, T_k]'$ and $\mathbf{0}$ is a vector of all zeros. Solving the function

$$\zeta_k(\alpha_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_{k-1}, \xi; \mathbf{0}) = 1 - \sum_{i=0}^{k} \alpha_i$$

for ξ involves evaluating the $k+1$-dimensional integral

$$\int_{-\infty}^{C_{\alpha_0}} \int_{-\infty}^{C_{\tilde{\alpha}_1}} \cdots \int_{-\infty}^{C_{\tilde{\alpha}_{k-1}}} \int_{-\infty}^{C_\xi} f(t_0, \ldots, t_k; \boldsymbol{\theta}) \prod_{i=0}^{k} dt_i \tag{4.1}$$

where f is the joint distribution of the test statistics $T_0, \ldots, T_k$, with mean parameter $\boldsymbol{\theta}$. We proceed assuming that these are normally distributed and standardized to have variance equal to one. That is, we are working with the standardized statistics $Z_0, Z_1, \ldots, Z_k$ as defined in Equation (2.3). Furthermore, we assume that measurement precision is equal across subgroups, so from Equation (2.6), $\rho_{ij} := \mathrm{Corr}(Z_i, Z_j) = \sqrt{f_{ij}}$ for $0 \leq i < j \leq \ell$, where $f_{ij} \in [0, 1]$ is the prevalence of Ω_j in Ω_i.

As populations are nested, Z_j depends on $Z_{j+1}, \ldots, Z_k$ only through Z_{j+1}. Hence, the conditional distribution of Z_j, given $Z_{j+1} = z_{j+1}, \ldots, Z_k = z_k$ depends only on z_{j+1}. Recall that $\mathcal{I}_j$ is the observed information for Ω_j. Define θ_j^* as the treatment effect size in $\Omega_j \backslash \Omega_{j+1}$, for $j = 0, 1, \ldots, k-1$, and similarly let $\Delta_j^* = \mathcal{I}_j - \mathcal{I}_{j+1}$ be the observed information for $\Omega_j \backslash \Omega_{j+1}$. Then, $Z_k \sim \mathcal{N}(\theta_k \sqrt{\mathcal{I}_k}, 1)$, and for $j = k-1, \ldots, 0$, increments are distributed as

$$Z_j \sqrt{\mathcal{I}_j} - Z_{j+1} \sqrt{\mathcal{I}_{j+1}} \sim \mathcal{N}\left(\theta_j^* \Delta_j^*, \Delta_j^*\right),$$

where increments are independent of $Z_k, \ldots, Z_{j+1}$. Hence, the conditional density of Z_j, given $Z_{j+1} = z_{j+1}, \ldots, Z_k = z_k$ is equal to

$$f_j(z_j \mid z_{j+1}; \boldsymbol{\theta}) = \sqrt{\frac{\mathcal{I}_j}{\Delta_j^*}} \varphi\left(\frac{z_j \sqrt{\mathcal{I}_j} - \theta_j^* \Delta_j^* - z_{j+1} \sqrt{\mathcal{I}_{j+1}}}{\sqrt{\Delta_j^*}}\right). \tag{4.2}$$

Now we can rewrite the joint density of $Z_0, Z_1, \ldots, Z_k$ as

$$f(z_0, \ldots, z_k; \boldsymbol{\theta}) = f_0(z_0 | z_1; \boldsymbol{\theta}) f_1(z_1 | z_2; \boldsymbol{\theta}) \cdots f_{k-1}(z_{k-1} | z_k; \boldsymbol{\theta}) f_k(z_k; \boldsymbol{\theta})$$

so (4.1) becomes

$$\int_{-\infty}^{C_\xi} f_k(z_k;\boldsymbol{\theta}) \int_{-\infty}^{C_{\tilde{\alpha}_{k-1}}} f_{k-1}(z_{k-1}|z_k;\boldsymbol{\theta}) \cdots \int_{-\infty}^{C_{\tilde{\alpha}_1}} f_1(z_1|z_2;\boldsymbol{\theta}) \int_{-\infty}^{C_{\alpha_0}} f_0(z_0|z_1;\boldsymbol{\theta}) \prod_{i=0}^{k} dz_i$$

which we write recursively as

$$\zeta_k(\alpha_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_{k-1}, \xi; \boldsymbol{\theta}) = \int_{-\infty}^{C_\xi} e_{k-1}(u;\boldsymbol{\theta}) f_k(u;\boldsymbol{\theta}) du \tag{4.3}$$

where the conditional densities of Z_j are given in Equation (4.2), and

$$e_j(x;\boldsymbol{\theta}) = \int_{-\infty}^{C_{\alpha_j}} e_{j-1}(u;\boldsymbol{\theta}) f_j(u|Z_{j+1} = x;\boldsymbol{\theta}) du, \; j = 1, \ldots, k-1,$$

with $e_0 \equiv 1$. Details on how to evaluate the integrals in (4.3) may for example be found in (Jennison and Turnbull, 2000, Ch. 19). When populations are not nested or disjoint, the above method is not feasible. However, Genz and Bretz (2002) have developed powerful algorithms that enable quick computation of multivariate normal and t probabilities with as many as twenty dimensions.

4.2 Adjusted Fallback Procedure with Enrichment (Two Stages)

We extend the FE procedure introduced in Section 3.2.2. For ease of exposition, we first consider the case of $\ell = 2$ subgroups, and then give the general design for $\ell > 2$ subgroups. Though the setup of the procedure does not require assumptions on the structure of the patient populations, it is mainly intended for the case of nested subgroups, i.e. $\Omega_0 \supseteq \Omega_1 \supseteq \cdots \supseteq \Omega_\ell$.

4.2.1 FE for $\ell = 2$ Subgroups

Let $\ell = 2$, $\mathcal{P}_1 = \{0, 1, 2\}$, and let α denote the desired FWER. As in Section 4.1, we specify local significance levels α_0, α_1 and α_2 such that $\alpha_0 + \alpha_1 + \alpha_2 = \alpha$. Equivalently, define weights $c_i \in [0, 1], i = 0, 1, 2$ that sum to 1, and let $\alpha_i = c_i\alpha$. Below we describe how adjusted significance levels $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ are obtained (using the important representation $\alpha_i = c_i\alpha$), but for now these are considered known. Also, stage-one decision parameters $\gamma_i, i = 0, 1, 2$ are specified, all taking values in $[0, 1]$.

The procedure takes place over two stages where, at stage i and for population Ω_j, we obtain the standardized test statistic T_{ij}, or equivalently the p-value p_{ij}. After the first stage, results are inspected and prespecified decision rules are followed to determine whether testing should proceed using the full population, or whether we should enrich to certain subgroups. As for the adjusted fallback procedure of Section 4.1, we assume that there is a natural ordering on the hypotheses under consideration. Greatest emphasis is placed on rejecting H_0, then H_1 and finally H_2. Hence, populations are eliminated sequentially, so $\mathcal{P}_2$ can be set as $\{0, 1, 2\}, \{1, 2\}$ or $\{2\}$. That is, Ω_0 is eliminated first, retaining only $\Omega_1 \cup \Omega_2$. Or, we can eliminate Ω_0 and Ω_1, analyzing only Ω_2 at the final analysis. $\mathcal{P}_2$ can also equal $\varnothing$ if we abandon the trial for futility.

If no enrichment takes place, $\mathcal{P}_2 = \mathcal{P}_1$, the standardized test statistic for Ω_j after the second stage is given as (combination weights are defined below)

$$T_j^{(0)} = w_{1j}^{(0)} T_{1j} + w_{2j}^{(0)} T_{2j}^{(0)},\ j = 0, 1, 2.$$

If we decide to restrict testing to Ω_1 and Ω_2 ($\mathcal{P}_2 = \{1, 2\}$), we get

$$T_j^{(1)} = w_{1j}^{(1)} T_{1j} + w_{2j}^{(1)} T_{2j}^{(1)},\ j = 1, 2.$$

and, finally, if only H_2 is tested ($\mathcal{P}_2 = \{2\}$), we get $T_2^{(2)} = w_{12}^{(2)} T_{12} + w_{22}^{(2)} T_{22}^{(2)}$. Here, the superscript k in $T_j^{(k)}$ corresponds to the smallest index of population Ω_k that is tested after the second stage. Having computed $T_j^{(k)}$, corresponding p-values $p_j^{(k)}$ can of course also be obtained.

This design uses empirical data weights, discussed in Section 2.1.1, and is designed to protect FWER strongly at level α while doing so. This is in contrast to adaptive procedures like CP, HPP and HUT, which will experience Type I error inflation if empirical data weights are used. When no enrichment takes place, weights are just as defined in Chapter 3, Section 3.1:

$$w_{1j}^{(0)} = w_{1j} = \sqrt{t} \text{ and } w_{2j}^{(0)} = w_{2j} = \sqrt{1-t},\ j = 0, 1, 2.$$

If $\mathcal{P}_2 = \{1, 2\}$, then for $j \in \mathcal{P}_2$,

$$\begin{aligned} w_{1j}^{(1)} &= \left(\frac{t f_{0j}}{t f_{0j} + (1-t) f_{\mathcal{P}_2, j}} \right)^{1/2}, \\ w_{2j}^{(1)} &= \left(\frac{(1-t) f_{\mathcal{P}_2, j}}{t f_{0j} + (1-t) f_{\mathcal{P}_2, j}} \right)^{1/2}. \end{aligned}$$

where $f_{\mathcal{P}_i, j} = f_{0j}/f_{\mathcal{P}_i}$ for $i = 1, 2$. Finally, for $\mathcal{P}_2 = \{2\}$, we have

$$w_{12}^{(2)} = \left(\frac{t f_{02}}{t f_{02} + 1 - t} \right)^{1/2} \text{ and } w_{22}^{(2)} = \left(\frac{1-t}{t f_{02} + 1 - t} \right)^{1/2}.$$

It is evident that all observations are given equal weight, and resulting test statistics $T_j^{(k)}$ are functions of sufficient statistics.

The procedure is now given, where the first stage analysis proceeds as follows:

I.0 If $p_{10} \le \gamma_0$, then go to Stage II.0;

I.1 Else, if $p_{11} \le \gamma_1$, then go to Stage II.1;

I.2 Else, if $p_{12} \le \gamma_2$, then go to Stage II.2;

Else, abandon the trial and accept all hypotheses.

Step I.2 represents a futility check so if none of the results are promising enough the trial will be abandoned. Of course, γ_2 may be set equal to 1, forcing the second stage to take place. After taking the second stage observations, the following analysis is carried out:

II.0 Carry out the original fallback procedure of Wiens (2003), using levels $\alpha_i, i = 0, 1, 2$.

II.1 Test H_1 at level $\tilde{\alpha}_1$. If H_1 is rejected, test H_2 at level $\tilde{\alpha}_1 + \tilde{\alpha}_2$. If H_1 is accepted, test H_2 at level $\tilde{\alpha}_2$.

II.2 Test H_2 at level $\tilde{\alpha}_2$.

The fallback procedure was reviewed in Section 2.2.2, but is briefly recalled here: If $p_0 \leq \alpha_0$ then H_0 is rejected. In general for H_i, let H_{m^*} be the last hypothesis before H_i that was rejected, with $m^* = 0$ if all have been rejected. Then H_i $(i > m^*)$ is tested at level $\alpha_{m^*} + \cdots + \alpha_i$.

Depending on the value chosen for $\gamma_2 \in [0, 1]$, early stopping may be allowed due to futility. Specifically, if $0 < \gamma_0, \gamma_1, \gamma_2 < 1$, there is a positive probability that stage two is not carried out if early results are not satisfactory. The total sample size, N, say, therefore becomes random, and

$$N = tn + (1-t)n \cdot I\left\{T_{1j} \geq C_{\gamma_j} \text{ for some } j \in \mathcal{P}_1\right\},$$

where $I(A)$ is the indicator function of an event A. Hence, the expected sample size becomes

$$\begin{aligned}\mathbb{E}(N) &= tn + (1-t)n\left[1 - \mathbb{P}\left(T_{10} < C_{\gamma_0}, T_{11} < C_{\gamma_1}, T_{12} < C_{\gamma_2}\right)\right] \\ &= n\left[t + (1-t)\left(1 - F_{T_{10},T_{11},T_{12}}\left(C_{\gamma_0}, C_{\gamma_1}, C_{\gamma_2}\right)\right)\right] \\ &\leq n,\end{aligned}$$

with equality only if $\gamma_j = 1$ for some $j = 0, 1, 2$.

Example 4.3. *Suppose all null hypotheses are true. Let $\gamma_0 = 0.4$, $\gamma_1 = 0.5$ and $\gamma_2 = 0.6$ and $t = 0.5$. Suppose that $f_{01} = 0.5$ and $f_{02} = 0.25$. Using correlation identities derived in Section 2.1.2, we can evaluate the CDF given above to see that $\mathbb{E}(N) = 0.86{\cdot}n$. If $t = 0.25$, $\mathbb{E}(N) = 0.79{\cdot}n$. We do note that savings in sample size are only possible when stopping the trial for futility (no early rejection). However, if early results indicate that efficacy is very unlikely, then the trial sponsor will likely want to abandon the trial and allocate resources elsewhere.*

4.2.2 Obtaining $\tilde{\alpha}_i$

Let $F_{X,Y,Z}(x, y, z)$ and $F_{X,Y,Z,W}(x, y, z, w)$ denote the CDFs of the random variables X, Y, Z and X, Y, Z, W, respectively. Then, under the complete null hypothesis $H_0 \cap H_1 \cap H_2$, it can be shown that (proof given in Section 4.2.6)

$$\begin{aligned}\mathbb{P}[\text{Type I Error}] = 1 &- F_{T_0^{(0)},T_1^{(0)},T_2^{(0)}}(C_{\alpha_0}, C_{\alpha_1}, C_{\alpha_2}) \\ &+ F_{T_{10},T_0^{(0)},T_1^{(0)},T_2^{(0)}}(C_{\gamma_0}, C_{\alpha_0}, C_{\alpha_1}, C_{\alpha_2}) \\ &- F_{T_{10},T_1^{(1)},T_2^{(1)}}(C_{\gamma_0}, C_{\tilde{\alpha}_1}, C_{\tilde{\alpha}_2}) + T_{T_{10},T_{11},T_1^{(1)},T_2^{(1)}}(C_{\gamma_0}, C_{\gamma_1}, C_{\tilde{\alpha}_1}, C_{\tilde{\alpha}_2}) \\ &- F_{T_{10},T_{11},T_{12}}(C_{\gamma_0}, C_{\gamma_1}, C_{\gamma_2}) - F_{T_{10},T_{11},T_2^{(2)}}(C_{\gamma_0}, C_{\gamma_1}, C_{\tilde{\alpha}_2}) \\ &+ F_{T_{10},T_{11},T_{12},T_2^{(2)}}(C_{\gamma_0}, C_{\gamma_1}, C_{\gamma_2}, C_{\tilde{\alpha}_2}).\end{aligned} \tag{4.4}$$

In Section 2.1.2, we derived correlation identities needed to specify correlation matrices in the above equation. Note that, if we do not allow stopping for futility, then $\gamma_2 = 1$, so $C_{\gamma_2} = -\infty$ and the two terms that involve γ_2 will evaluate to zero.

In order to obtain $\tilde{\alpha}_i, i = 1, 2$, we proceed as follows. Set

$$r_i = \frac{c_i}{\sum_{j=1}^{2} c_j}, \ i = 1, 2,$$

and replace $\tilde{\alpha}_i$ in Equation (4.4) with $r_i\xi_0$. Then, use a numerical search[1] to find ξ_0 when $\mathbb{P}$[Type I Error] is set equal to α. Next, consider an auxiliary system consisting only of Ω_0 and Ω_2. Using parameters $\gamma_0' = \gamma_0$, $\gamma_2' = \gamma_2$, $\alpha_0' = \alpha_0$ and $\alpha_2' = \alpha_1 + \alpha_2$, we can follow steps given in Section 3.2.2 (see Equation (3.5)) to obtain the adjusted significance level, ξ_1, say, that ensures strong protection of FWER at level α in the auxiliary system. In the notation of Section 3.2.2 this is actually $\tilde{\alpha}_1$, but here it is more convenient to use ξ_1, as will become apparent when considering general $\ell > 2$. Now, set $\xi^- = \min\{\xi_0, \xi_1\}$, and $\tilde{\alpha}_i = r_i\xi^-$ for $i = 1, 2$. Note that, since $\xi_1 \leq \alpha$, this ensures that $\tilde{\alpha}_1 + \tilde{\alpha}_2 \leq \alpha$, and $\tilde{\alpha}_2 \leq \xi_1$. Hence critical values for the full system are larger than those obtained from the auxiliary system, which consists only of Ω_0 and Ω_2. By construction, the procedure has Type I error probability bounded above by α, under the complete null hypothesis. Proof of strong FWER control is given in Section 4.2.6.

4.2.3 Non-Centrality Parameters

Let $\beta \in [0, 1]$ be given, and suppose we wish to power the study for H_0 at $1 - \beta$ for the effect $\delta_0 = \theta_0/\sigma$, where σ^2 is the pooled variance for the overall population. The non-centrality parameter $\lambda_0^{(0)} = \mathbb{E}\left(T_0^{(0)}|\delta_0\right) = \sqrt{\frac{n}{4}}\delta_0$ (and hence the required sample size) is obtained numerically by solving the following equation for $\lambda_0^{(0)}$:

$$\begin{aligned}
1 - \beta &= \mathbb{P}[\text{Reject } H_0 \mid \delta_0] \\
&= \mathbb{P}\left(T_{10} \geq C_{\gamma_0}, T_0^{(0)} \geq C_{\alpha_0} \mid \delta_0\right) \\
&= 1 - F_{T_{10}}\left(C_{\gamma_0} - w_{10}^{(0)}\lambda_0^{(0)}\right) - F_{T_0^{(0)}}\left(C_{\alpha_0} - \lambda_0^{(0)}\right) \\
&\qquad + F_{T_{10},T_0^{(0)}}\left(C_{\gamma_0} - w_{10}^{(0)}\lambda_0^{(0)}, C_{\alpha_0} - \lambda_0^{(0)}; w_{10}^{(0)}\right).
\end{aligned}$$

[1] CDFs in Equation (4.4) can for example be evaluated in Matlab using functions such as `mvncdf` or `qsimvnv`.

Let $\delta_j = \eta_j \delta_0$ denote the standardized treatment effect in Ω_j, where $\eta_j \geq 0$ for $j = 1, 2$. When no populations are dropped, non-centrality parameters for Ω_j are given as

$$\lambda_j^{(0)} = \mathbb{E}\left(T_j^{(0)} \mid \delta_j\right) = \sqrt{\frac{f_{0j} n_0}{4}} \cdot \delta_j = \eta_j \sqrt{f_{0j}} \cdot \lambda_0^{(0)}, \ j = 1, 2.$$

It is also easily seen that $\lambda_2^{(0)} = \frac{\eta_2}{\eta_1} \sqrt{f_{12}} \cdot \lambda_1^{(0)}$. If $\mathcal{P}_2 = \{1, 2\}$, we get the following non-centrality parameters:

$$\begin{aligned} \lambda_1^{(1)} = \mathbb{E}\left(T_1^{(1)} \mid \delta_1\right) &= \sqrt{\frac{t f_{01} n_0}{4} + \frac{(1-t) n_0}{4}} \cdot \delta_1 \\ &= \eta_1 \sqrt{t f_{01} + 1 - t} \cdot \lambda_0^{(0)}, \end{aligned}$$

and

$$\lambda_2^{(1)} = \mathbb{E}\left(T_2^{(1)} \mid \delta_2\right) = \sqrt{\frac{t f_{02} n_0}{4} + \frac{(1-t) f_{12} n_0}{4}} \cdot \delta_2 = \eta_2 \sqrt{t f_{02} + (1-t) f_{12}} \cdot \lambda_0^{(0)}.$$

Finally, if $\mathcal{P}_2 = \{2\}$, then

$$\lambda_2^{(2)} = \mathbb{E}\left(T_2^{(2)} \mid \delta_2\right) = \eta_2 \sqrt{t f_{02} + 1 - t} \cdot \lambda_0^{(0)}.$$

Note that $f_{01} \leq 1$ implies that $\lambda_1^{(1)} \geq \lambda_1^{(0)}$, and $f_{02} \leq f_{12} \leq 1$ implies $\lambda_2^{(k)} \leq \lambda_2^{(k')}$ for $k < k'$, signifying the potential increase in power when stage two is carried out on smaller populations only.

When subgroups are not nested (e.g. disjoint), quantities such as f_{12} may equal zero. In such cases, it may be more useful to think of f_{12} as the percentage of observations allocated to Ω_2, compared to that of Ω_1. As hypotheses have been ordered by importance (proving treatment efficacy in Ω_1 is more important than proving efficacy in Ω_2), it should still be the case that $f_{02} \leq f_{12} \leq 1$.

4.2.4 FE for $\ell \geq 2$ Subgroups

We now extend the procedure to the case of an arbitrary number of populations, $\Omega_0, \Omega_1, \ldots, \Omega_\ell$ where $\Omega_0 \supsetneq \Omega_j$ for $j \in \mathcal{P}_1 := \{0, 1, \ldots, \ell\}$. These populations admit a natural ordering, where establishing efficacy for Ω_i is of greater importance than for Ω_j, for $i < j$. We have the usual null hypotheses $H_j : \delta_j \leq 0$ for $j \in \mathcal{P}_1$. Local significance levels $\alpha_0, \alpha_1, \ldots, \alpha_\ell \in [0, 1]$ are specified such that $\sum_{j \in \mathcal{P}_1} \alpha_j = \alpha$, the desired FWER. Equivalently, values $c_j \in [0, 1]$ are specified such that $\sum_{j \in \mathcal{P}_1} c_j = 1$ and $\alpha_j = c_j \alpha$. Finally, before any data is unblinded we must specify enrichment parameters $\gamma_j \in [0, 1]$ for $j \in \mathcal{P}_1$. In Section 4.2.5 we describe how to obtain adjusted significance levels $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell$, but for now we take these as given. Standardized statistics (p-values) at stage i for Ω_j are

$$T_{ij}\ (p_{ij}), \text{ for } i = 1, 2;\ j \in \mathcal{P}_i.$$

As before, $f_{\mathcal{S},j}$ denotes the prevalence of Ω_j in $\Omega_{\mathcal{S}}$, $j \in \mathcal{P}_1$ and $\mathcal{S} \subseteq \mathcal{P}_1$. The allotted sample size for the whole trial is n, and $n_1 = tn$ and $n_2 = (1-t)n$ for stage one and two respectively. The sample size for Ω_j in stage i is then

$$n_{ij} = f_{\mathcal{P}_i, j} n_i, \text{ for } i = 1, 2;\ j \in \mathcal{P}_i.$$

If the trial is not stopped for futility, let $k = \min_j \mathcal{P}_2$ be the index of the first population (in our ordering) that is carried on to stage two. If $k = 0$, then combination weights are given as $w_{ij}^{(0)} = (n_{ij}/f_{0j}n)^{1/2} = (n_i/n)^{1/2}$ for $i = 1, 2$ and $j \in \mathcal{P}_1$. For $k > 0$, we use

$$\begin{aligned} w_{1j}^{(k)} &= \left(\frac{t f_{\mathcal{P}_1, j}}{t f_{\mathcal{P}_1, j} + (1-t) f_{\mathcal{P}_2, j}} \right)^{1/2} \\ w_{2j}^{(k)} &= \left(\frac{(1-t) f_{\mathcal{P}_2, j}}{t f_{\mathcal{P}_1, j} + (1-t) f_{\mathcal{P}_2, j}} \right)^{1/2}, \end{aligned}$$

for $j = k, \ldots, \ell$. Final statistics after stage two are given as

$$T_j^{(k)} = w_{1j}^{(k)} T_{1j} + w_{2j}^{(k)} T_{2j}^{(k)},\ j = k, \ldots, \ell.$$

Following the work in Section 2.1.2, we can give correlation identities of interest for this method:

$$
\begin{aligned}
\text{Corr}\left(T_{1j}, T_j^{(k)}\right) &= w_{1j}^{(k)},\ 0 \le k \le j \le \ell, \\
\text{Corr}\left(T_i^{(k)}, T_j^{(k)}\right) &= \sqrt{r_{ij} r_{ji}},\ 0 \le k \le i, j \le \ell, \\
\text{Corr}\left(T_{1i}, T_j^{(k)}\right) &= w_{1j}^{(k)} \sqrt{r_{ij} r_{ji}},\ 0 \le k, i, j \le \ell,\ j \ge k,
\end{aligned}
$$

where r_{ij} is given in Equation (2.5). Finally, the non-centrality parameter for $T_j^{(k)}$ is given as

$$\lambda_j^{(k)} = \eta_j \sqrt{t f_{\mathcal{P}_1,j} + (1-t) f_{\mathcal{P}_2,j}} \cdot \lambda_0^{(0)},\ j \in \mathcal{P}_2,$$

where $\delta_j = \eta_j \delta_0$ and $\lambda_0^{(0)}$ is obtained as described in Section 4.2.3. The testing algorithm for this procedure is given in Algorithm 2. Note that when we eliminate populations $\Omega_0, \dots, \Omega_{k-1}$ after stage one, hypotheses $H_0, \dots, H_{k-1}$ are regarded as having been accepted without further investigation.

```
Input: $\alpha_j$ for $j \in \mathcal{P}_1$, $\tilde{\alpha}_j$ for $j = 1, \dots, \ell$ and $T_{1j}$ for $j \in \mathcal{P}_1$
$k = 0$;
while $T_{1k} < C_{\gamma_k}$ do
    $k = k + 1$;
end
if $k < \ell + 1$ then
    if $k = 0$ then
        Use fallback procedure on $\Omega_0, \dots, \Omega_\ell$ with $\alpha_j$ and $T_j^{(0)}$, $j \in \mathcal{P}_1$;
    else
        Use fallback procedure on $\Omega_k, \dots, \Omega_\ell$ with $\tilde{\alpha}_j$ and $T_j^{(k)}$, $j = k, \dots, \ell$;
    end
else
    Stop the trial for futility;
end
```

Algorithm 2: Testing algorithm for FE.

4.2.5 Obtaining $\tilde{\alpha}_j$

In this section, we describe how to obtain the adjusted significance levels $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell$. First, we define some notation that is useful to describe decision paths taken through the procedure. In what follows, let $\bar{E}$ denote the complement of the set E. For stage one, let $S_k = [p_{1k} \leq \gamma_k]$, $k = 0, \ldots, \ell$. Hence if we observe $\bar{S}_0 \cap \cdots \cap \bar{S}_{k-1} \cap S_k$, then $\mathcal{P}_2 = \{k, \ldots, \ell\}$ for stage two. Let

$$s_{ji}^{(k)} = \sum_{m=j}^{i} \alpha_m^{(k)}, \text{ where } \alpha_m^{(0)} = \alpha_m \text{ and } \alpha_m^{(k)} = \tilde{\alpha}_m \text{ for } k = 1, \ldots, \ell.$$

For the second stage we define rejection events $R_{ij}^{(k)}$ as follows. If no enrichment is carried out (i.e. the event S_0 is observed), then

$$\begin{aligned}
R_{00}^{(0)} &= \left[T_0^{(0)} \geq C_{\alpha_0}\right] \\
R_{10}^{(0)} &= \left[T_1^{(0)} \geq C_{\alpha_0+\alpha_1}\right] \\
R_{11}^{(0)} &= \left[T_1^{(0)} \geq C_{\alpha_1}\right] \\
&\vdots \\
R_{ij}^{(0)} &= \left[T_i^{(0)} \geq C_{s_{ji}^{(0)}}\right], \ i = 0, 1, \ldots, \ell \text{ and } j = 0, \ldots, i.
\end{aligned}$$

Following stage one events $\bar{S}_0 \cap \cdots \cap \bar{S}_{k-1} \cap S_k$ for $k = 1, \ldots, \ell$, let

$$R_{ij}^{(k)} = \left[T_i^{(k)} \geq C_{s_{ji}^{(k)}}\right], \ i = k, \ldots, \ell \text{ and } j = k, \ldots, i.$$

The notation can be explained as follows: $R_{ij}^{(k)}$ is the rejection of H_i with $\mathcal{P}_2 = \{k, \ldots, \ell\}$ at level $s_{ji}^{(k)} = \sum\limits_{m=j}^{i} \alpha_m^{(k)}$, where H_{j-1} is the last hypothesis that was not rejected. If $j - 1 < k$, then all hypotheses prior to H_i have been rejected (minus those eliminated due to enrichment). We point out the special case where no enrichment took place, and all hypotheses prior to H_i were rejected. In this case (as in the original fallback procedure), H_i is simply tested at level $\sum\limits_{m=0}^{i} \alpha_m$.

Example 4.4. *Suppose we have three populations and* Ω_0 *is eliminated at the end of stage one. Then* $\ell = 2$*,* $k = 1$*, and*

$$R_{11}^{(1)} = \left[T_1^{(1)} \geq C_{\tilde{\alpha}_1}\right];\ R_{21}^{(1)} = \left[T_2^{(1)} \geq C_{\tilde{\alpha}_1+\tilde{\alpha}_2}\right];\ R_{22}^{(1)} = \left[T_2^{(1)} \geq C_{\tilde{\alpha}_2}\right].$$

Now, for stage one events $\bar{S}_0 \cap \cdots \cap \bar{S}_{k-1} \cap S_k$, $k \in \mathcal{P}_1$, which defines the index set $\mathcal{P}_2 = \{k, \ldots, \ell\}$ of populations to be tested after stage two, and for the index set $B^{(k)} \subseteq \mathcal{P}_2$, define the event

$$\mathcal{R}\left(B^{(k)}\right) := \left[\text{Reject } H_j,\ j \in B^{(k)} \text{ and accept } H_j,\ j \in \mathcal{P}_2 \setminus B^{(k)}\right].$$

We illustrate this notation in an example:

Example 4.5. *Suppose we have four populations and* Ω_0 *is eliminated after stage one. Then,* $\ell = 3$*,* $k = 1$ *and* $\mathcal{P}_2 = \{1, 2, 3\}$*. Suppose that* $B^{(1)} = \{1, 3\}$*. Then*

$$\begin{aligned}\mathcal{R}\left(B^{(1)}\right) &= \bar{S}_0 \cap S_1 \cap \left[R_{11}^{(1)} \cap \bar{R}_{21}^{(1)} \cap R_{33}^{(1)}\right] \\ &= \left[T_{10} < C_{\gamma_0}, T_{11} \geq C_{\gamma_1}, T_1^{(1)} \geq C_{\tilde{\alpha}_1}, T_2^{(1)} < C_{\tilde{\alpha}_1+\tilde{\alpha}_2}, T_3^{(1)} \geq C_{\tilde{\alpha}_3}\right].\end{aligned}$$

Suppose that h hypotheses are true, indexed as $A_h := \{m_1, m_2, \ldots, m_h\} \subseteq \mathcal{P}_1$. Then, the Type I error probability is given as

$$\mathbb{P}\left\{\bigcup_{k=0}^{\ell} \bigcup_{\substack{B^{(k)} \subseteq \{k,\ldots,\ell\}: \\ B^{(k)} \cap A_h \neq \varnothing}} \mathcal{R}\left(B^{(k)}\right) \;\middle|\; H_{m_1}, H_{m_2}, \ldots, H_{m_h}\right\}.$$

We now outline the steps needed to obtain adjusted significance levels $\tilde{\alpha}_j$, $j = 1, \ldots, \ell$:

1. Following the notation defined above, and under the global null hypothesis $\bigcap_{j=0}^{\ell} H_j$, the Type I error probability is

$$\mathbb{P}\left\{\bigcup_{k=0}^{\ell} \bigcup_{B^{(k)} \in 2^{\{k,\ldots,\ell\}} \setminus \varnothing} \mathcal{R}\left(B^{(k)}\right) \;\middle|\; H_0, H_1, \ldots, H_\ell\right\}. \tag{4.5}$$

As we did in Section 4.2.2 when $\ell = 2$, replace $\tilde{\alpha}_j$ in the expression given in Equation (4.5) with $r_j\xi_0$, where $r_j = c_j/(c_1 + \cdots + c_\ell)$, and solve numerically for ξ_0 with (4.5) set equal to α.

2. For $i = 1, \ldots, \ell - 1$, do

   - Consider the auxiliary system consisting of all populations except Ω_i.
   - Assign stage-one decision parameters $\gamma'_j = \gamma_j$ for $j = 0, \ldots, i-1$, and $\gamma'_{i+1} = 1$. (Subsequent γ_j do not matter.)
   - Assign local significance levels $\alpha'_j = \alpha_j$ for $j \neq i$, and $\alpha'_{i+1} = \alpha_i + \alpha_{i+1}$. Note that this implies $c'_j = c_j$ and $c'_{i+1} = c_i + c_{i+1}$.
   - Under the global null hypothesis, $H_0 \cap \cdots \cap H_{i-1} \cap H_{i+1} \cap \cdots \cap H_\ell$, in the auxiliary system, obtain adjusted significance levels $\tilde{\alpha}'_j = r'_j\xi_i$, where

     $$r'_j = c'_j \Big/ \sum_{k=1}^{\ell} c_k,\ j = 1, \ldots, i-1, i+1, \ldots, \ell.$$

     and ξ_i is obtained numerically to ensure that Type I error probability is equal to α under the global hypothesis for the auxiliary system.

3. Set $\xi^- = \bigwedge_{j=0}^{\ell-1} \xi_j$ and $\tilde{\alpha}_j = r_j\xi^-$ for $j = 1, \ldots, \ell$.

We see that in order to obtain the adjusted significance levels $\tilde{\alpha}_j$, we need to solve numerically for ξ_j in a total of ℓ equations. Namely, we must evaluate the probability of a Type I error under the global null hypothesis in the system of populations $\Omega_0, \Omega_1, \ldots, \Omega_\ell$, and in each of the auxiliary systems $\Omega_0, \ldots, \Omega_{i-1}, \Omega_{i+1}, \ldots, \Omega_\ell$ for $i = 1, \ldots, \ell - 1$. Many multiple comparison procedures that rely on the closure principle, require explicit specification of intersection hypotheses. The number of required intersection hypotheses increases exponentially with new elementary hypotheses, which can make the underlying testing strategy difficult to communicate.

This is not the case here, and the preparatory computational burden (computing adjusted significance values) does not increase exponentially with the number of hypotheses considered.

4.2.6 Control of the FWER

Proof for $\ell = 2$

We first provide a proof for the case $\ell = 2$. In the following, H_i^a denotes the alternative hypothesis to H_i. We start with the global null hypothesis $H_0 \cap H_1 \cap H_2$. In the notation of Section 4.2.5, a Type I error occurs with $\mathcal{P}_2 = \mathcal{P}_1$ if any of the following events occur:

$$\begin{aligned} S_0 \cap \Big[&\left(R_{00}^{(0)} \cap R_{10}^{(0)} \cap R_{20}^{(0)}\right) \cup \left(R_{00}^{(0)} \cap R_{10}^{(0)} \cap \bar{R}_{20}^{(0)}\right) \cup \left(R_{00}^{(0)} \cap \bar{R}_{10}^{(0)} \cap \bar{R}_{22}^{(0)}\right) \\ &\cup \left(R_{00} \cap \bar{R}_{10}^{(0)} \cap R_{22}^{(0)}\right) \cup \left(\bar{R}_{00}^{(0)} \cap R_{11}^{(0)} \cap R_{21}^{(0)}\right) \cup \left(\bar{R}_{00}^{(0)} \cap R_{11}^{(0)} \cap \bar{R}_{21}^{(0)}\right) \\ &\cup \left(\bar{R}_{00}^{(0)} \cap \bar{R}_{11}^{(0)} \cap R_{22}^{(0)}\right) \Big]. \end{aligned}$$

If $\mathcal{P}_2 = \{1, 2\}$, or if $\mathcal{P}_2 = \{2\}$, the following events lead to a Type I error:

$$\begin{aligned} &\left\{ \bar{S}_0 \cap S_1 \cap \left[\left(R_{11}^{(1)} \cap R_{21}^{(1)}\right) \cup \left(\bar{R}_{11}^{(1)} \cap R_{22}^{(1)}\right) \cup \left(R_{11}^{(1)} \cap \bar{R}_{21}^{(1)}\right) \right] \right\} \\ &\qquad \cup \left\{ \bar{S}_0 \cap \bar{S}_1 \cap S_2 \cap R_{22}^{(2)} \right\}. \end{aligned}$$

Taking the union of these two expressions and simplifying yields

$$\begin{aligned}
\mathbb{P}[\text{Type I error}] = \mathbb{P}\Bigg\{ &\left(S_0 \cap \left[R_{00}^{(0)} \cup \left(\bar{R}_{00}^{(0)} \cap R_{11}^{(0)}\right) \cup \left(\bar{R}_{00}^{(0)} \cap \bar{R}_{11}^{(0)} \cap R_{22}^{(0)}\right)\right]\right) \\
&\cup \left(\bar{S}_0 \cap S_1 \cap \left[R_{11}^{(1)} \cup \left(\bar{R}_{11}^{(1)} \cap R_{22}^{(1)}\right)\right]\right) \cup \left(\bar{S}_0 \cap \bar{S}_1 \cap S_2 \cap R_{22}^{(2)}\right)\Bigg\} \\
= \mathbb{P}&\left[T_{10} \geq C_{\gamma_0}, T_0^{(0)} \geq C_{\alpha_0}\right] + \mathbb{P}\left[T_{10} \geq C_{\gamma_0}, T_0^{(0)} < C_{\alpha_0}, T_1^{(0)} \geq C_{\alpha_1}\right] \\
&+ \mathbb{P}\left[T_{10} \geq C_{\gamma_0}, T_0^{(0)} < C_{\alpha_0}, T_1^{(0)} < C_{\alpha_1}, T_2^{(0)} \geq C_{\alpha_2}\right] \\
&+ \mathbb{P}\left[T_{10} < C_{\gamma_0}, T_{11} \geq C_{\gamma_1}, T_1^{(1)} \geq C_{\tilde{\alpha}_1}\right] \\
&+ \mathbb{P}\left[T_{10} < C_{\gamma_0}, T_{11} \geq C_{\gamma_1}, T_1^{(1)} < C_{\tilde{\alpha}_1}, T_2^{(1)} \geq C_{\tilde{\alpha}_2}\right] \\
&+ \mathbb{P}\left[T_{10} < C_{\gamma_0}, T_{11} < C_{\gamma_1}, T_{12} \geq C_{\gamma_2}, T_2^{(2)} \geq C_{\tilde{\alpha}_2}\right].
\end{aligned}$$

Expressing this in terms of the respective CDFs and simplifying leads to equation (4.4) in Section 4.2.2. Hence, by construction, we know that this probability is bounded above by α, which is the desired Type I probability.

For notational convenience we adopt the following convention. For a given set A, let B_A denote a set B which contains at least one element of A. Then, for $H_0^a \cap H_1 \cap H_2$, a Type I error can be expressed as

$$\begin{aligned}
\bigcup_{i=0}^{2} \bigcup_{B^{(i)}_{\{1,2\}} \subseteq \{i,\dots,2\}} \mathcal{R}\left(B^{(i)}\right) &\subseteq \left\{S_0 \cap \left[R_{10}^{(0)} \cup \left(\bar{R}_{11}^{(0)} \cap R_{22}^{(0)}\right)\right]\right\} \\
\cup \quad &\left\{\bar{S}_0 \cap S_1 \cap \left[R_{11}^{(1)} \cup \left(\bar{R}_{11}^{(1)} \cap R_{22}^{(1)}\right)\right]\right\} \cup \left\{\bar{S}_0 \cap \bar{S}_1 \cap S_2 \cap R_{22}^{(2)}\right\}.
\end{aligned}$$

Now, for sufficiently large n, the event paths involving $\bar{S}_0 \cap S_1$ and $\bar{S}_0 \cap \bar{S}_1 \cap S_2$ will be dominated by the first path so

$$\begin{aligned}
\mathbb{P}[\text{Type I error}] &\leq \mathbb{P}\left\{S_0 \cap \left[R_{10}^{(0)} \cup \left(\bar{R}_{11}^{(0)} \cap R_{22}^{(0)}\right)\right] \big| H_1, H_2\right\} \\
&\leq \mathbb{P}\left\{R_{10}^{(0)} \cup \left(\bar{R}_{11}^{(0)} \cap R_{22}^{(0)}\right) \big| H_1, H_2\right\}.
\end{aligned}$$

The last line is the probability of a Type I error event in a three population fallback procedure under $H_0^a \cap H_1 \cap H_2$, and hence it is bounded by α.

Next, for $H_0 \cap H_1^a \cap H_2$, a Type I error is expressed as

$$\begin{aligned}\bigcup_{i=0}^{2} \bigcup_{B_{\{0,2\}}^{(i)} \subseteq \{i,\ldots,2\}} \mathcal{R}\left(B^{(i)}\right) \;\subseteq\; & \left\{S_0 \cap \left[R_{00}^{(0)} \cup \left(\bar{R}_{00}^{(0)} \cap R_{21}^{(0)}\right)\right]\right\} \\ & \cup \left\{\bar{S}_0 \cap S_1 \cap R_{21}^{(1)}\right\} \cup \left\{\bar{S}_0 \cap \bar{S}_1 \cap S_2 \cap R_{20}^{(2)}\right\}.\end{aligned}$$

For large n, the last event path (corresponding to step II.2 in Section 4.2.1) is dominated by the first two event paths. Furthermore, $\bar{S}_0 \cap S_1 \cap R_{21}^{(1)} \subset \bar{S}_0 \cap R_{21}^{(1)}$, so

$$\begin{aligned}\mathbb{P}[\text{Type I error}] \;\le\; & \mathbb{P}\left\{S_0 \cap \left[R_{00}^{(0)} \cup \left(\bar{R}_{00}^{(0)} \cap R_{21}^{(0)}\right)\right] \big| H_0, H_2\right\} \\ & + \mathbb{P}\left\{\bar{S}_0 \cap R_{21}^{(1)} \big| H_0, H_2\right\}\end{aligned}$$

which is simply a Type I error probability for the same procedure in the reduced two population system consisting of Ω_0 and Ω_2. Here we are using critical values for the three population system, which are at least as large as those of the reduced system. Hence FWER is bounded above by α.

For $H_0 \cap H_1 \cap H_2^a$, a Type I error is given by

$$\bigcup_{i=0}^{2} \bigcup_{B_{\{0,1\}}^{(i)} \subseteq \{i,\ldots,2\}} \mathcal{R}\left(B^{(i)}\right) \;\subseteq\; \left\{S_0 \cap \left[R_{00}^{(0)} \cup \left(\bar{R}_{00}^{(0)} \cap R_{11}^{(0)}\right)\right]\right\} \cup \left\{\bar{S}_0 \cap S_1 \cap R_{11}^{(1)}\right\}.$$

This corresponds to a Type I error event for the same procedure, but for a system consisting only of Ω_0 and Ω_1. However, the critical values correspond to those of the three population system, and are hence larger. Therefore the above event occurs with probability less than or equal to α.

For $H_0 \cap H_1^a \cap H_2^a$, the Type I error probability is easily expressed as

$$\mathbb{P}[S_0 \cap R_{00} \mid H_0] \le \mathbb{P}[R_{00} \mid H_0] = \alpha_1 \le \alpha.$$

For $H_0^a \cap H_1 \cap H_2^a$, a Type I error is contained in

$$\left(S_0 \cap R_{10}^{(0)}\right) \cup \left(\bar{S}_0 \cap S_1 \cap R_{11}^{(1)}\right)$$

and, for sufficiently large n, this event is dominated by $S_0 \cap R_{10}^{(0)}$, which clearly has probability $\leq \alpha$. Similarly, it is easy to see that under $H_0^a \cap H_1^a \cap H_2$, a Type I error event is contained in

$$\left(S_0 \cap R_{20}^{(0)}\right) \cup \left(\bar{S}_0 \cap S_1 \cap R_{21}^{(0)}\right) \cup \left(\bar{S}_0 \cap \bar{S}_1 \cap R_{22}^{(1)}\right)$$

which is again dominated by $S_0 \cap R_{20}^{(0)}$, which has probability less than α. Thus, we have shown that when $\ell = 2$, FWER is strongly protected at level α.

Proof for General ℓ

Suppose $h < \ell + 1$ hypotheses are true. (The case where $h = \ell + 1$ is taken care of by construction.) Define the index set

$$A_h := \{m_1, m_2, \ldots, m_h\} \subsetneq \mathcal{P}_0$$

corresponding to those true hypotheses. We need to show that

$$\mathbb{P}\left\{\bigcup_{i=0}^{\ell} \bigcup_{B_{A_h}^{(i)} \subseteq \{i,\ldots,\ell\}} \mathcal{R}\left(B_{A_h}^{(i)}\right) \middle| H_{m_1}, \ldots, H_{m_h}\right\} \leq \alpha.$$

Let $\ell^* = \min\{I \backslash A\}$ correspond to the first hypothesis (in our ordering) which is not true. We then consider the following two cases.

Suppose $\ell^* = 0$. Then H_0 is not true and hence, for sufficiently large n, the Type I error event is dominated by

$$\begin{aligned} \bigcup_{B_{A_h}^{(0)} \subseteq \{0,\ldots,\ell\}} \mathcal{R}\left(B_{A_h}^{(0)}\right) &= S_0 \cap \left(\bigcup_{i \in A_h} [\text{Reject } H_i \text{ with fallback procedure}]\right) \\ &\subseteq \bigcup_{i \in A_h} [\text{Reject } H_i \text{ with fallback procedure}] \end{aligned}$$

which is simply a Type I error event for the fallback procedure. This is known to have probability less than or equal to α.

Now, suppose that $\ell^* > 1$. Then, as $\Pr[\bar{S}_{\ell^*}] \approx 0$ for sufficiently large n, the Type I error event is dominated by

$$\bigcup_{i=0}^{\ell^*} \bigcup_{B_{A_h}^{(i)} \subseteq \{i,\ldots,\ell\}} \mathcal{R}\left(B_{A_h}^{(i)}\right) \subseteq \left\{ \bigcup_{i=0}^{\ell^*-1} \bigcup_{B_{A_h}^{(i)} \subseteq \{i,\ldots,\ell\}} \mathcal{R}\left(B_{A_h}^{(i)}\right) \right\}$$
$$\cup \left\{ \bar{S}_1 \cap \cdots \cap \bar{S}_{\ell^*-1} \cap \left(\bigcup_{i \in A_h, i > \ell^*} [\text{Reject } H_i \text{ with adjusted } \alpha\text{-values}] \right) \right\}$$

This is a Type I error event for the reduced system in which Ω_{ℓ^*} is eliminated, and $\gamma_{\ell^*+1} = 1$. The critical values being used are for the $\ell+1$ population system, and are hence at least as large as the critical values needed to guarantee $\mathbb{P}[\text{Type I error}] \leq \alpha$ for the reduced system. Therefore $\mathbb{P}[\text{Type I error}] \leq \alpha$ for the ℓ population system as well, which proves strong control of FWER.

4.3 Hybrid Bayesian Adaptive Design with Utility-Based Interim Analysis (Two Stages)

As a generalization of Section 3.2.4 (the HUT design), we consider the use of adaptive seamless designs (ASD, see Section 2.2.3) for subgroup selection. The design proposed here is a two-stage design, representing the combination of Phases II and III in a clinical trial. For the interim analysis, we use a hierarchical Bayes setup, and specify a gain (loss) function to model the costs of making incorrect decisions before stage two. A key assumption is the exchangeability among treatment-by-subgroup interactions. That is, there is no *a priori* distinction to be made of the subgroups of interest with respect to treatment effect. (As stated in Chapter 3, this is often *not* a reasonable assumption.) An important feature of the hierarchical Bayes setup, is that it allows researchers to specify upfront the strength of their belief that subgroup-specific effects are present.

4.3.1 Setup

As in the notation introduced in Section 2.1.1, Ω_0 denotes the overall population, and $\Omega_1, \ldots, \Omega_\ell$ are ℓ disjoint subgroups, identified prior to the start of the trial. The assumption of disjointness can be made without loss of generality, as we can always redefine given subgroups to be disjoint (there may just be more of them). We consider $H_j : \theta_j \leq 0$, $j = 0, 1, \ldots, \ell$ as our hypotheses of interest. Let $\mathcal{P}_1 = \{1, \ldots, \ell\}$ denote the index set of populations under consideration during stage one. At the interim analysis, we may eliminate any number of subgroups, and $\mathcal{P}_2 \subseteq \mathcal{P}_1$ denotes the index set of populations carried on to stage two. If $\mathcal{P}_2 = \varnothing$, the study is abandoned due to futility.

Conditioning on $\mathcal{P}_i$, Z_{ij} denotes the standardized statistic for population Ω_j in stage i, $i = 1, 2$ and $j \in \mathcal{P}_i$. We use vector notation $\boldsymbol{Z}_i \in \mathbb{R}^{|\mathcal{P}_i|}$ to denote the stage-wise statistics for stage i, and $\boldsymbol{Z} \in \mathbb{R}^{|\mathcal{P}_i|}$ to denote the vector of combined test statistics for both stages. It is assumed that

$$\boldsymbol{Z}_i \sim \mathcal{N}\left(\boldsymbol{\delta}, I\right), \text{ where } \delta_j = \theta_j \sqrt{\mathcal{I}_{ij}}, \; i = 1, 2 \text{ and } j \in \mathcal{P}_i.$$

Here, I is the appropriately sized identity matrix. Conditioning on $\mathcal{P}_i$, the information $\mathcal{I}_{ij}$ is known, so the unknown component of δ_j is θ_j.

We shall specify a hierarchical Bayesian model for the parameters $\theta_1, \ldots, \theta_\ell$, which are assumed exchangeable. That is, the prior distribution $p(\theta_1, \ldots, \theta_\ell)$ is assumed to be invariant to permutations of the indices $1, \ldots, \ell$. For the treatment effect parameters, and for the hyperparameter ν, we use the priors

$$\theta_1, \ldots, \theta_\ell \overset{iid}{\sim} \mathcal{N}(\nu, \tau^2), \text{ and } \nu \sim \mathcal{N}(\phi, \omega^2).$$

Since $\theta_0 = \sum_{j \in \mathcal{P}_1} f_{0j}\theta_j$, a prior on θ_0 has been implicitly specified. We do not place priors on the dispersion parameters τ and ω, nor on the parameter ϕ.

Note the important role of the dispersion parameter τ. By varying the value of τ, researchers are able to specify the prior belief about presence of heterogeneity among subgroups. Namely, if τ is small, the θ_j are likely all concentrated narrowly around their mean, ν. On the other hand, if τ is large, it is implied that the parameters θ_j can vary considerably.

We now compute the posterior distribution of θ_j and ν, given first stage data $\boldsymbol{Z}_1 = \boldsymbol{z}_1$. For shorthand, write $b_j = \sqrt{\mathcal{I}_{1j}}$ and $\delta_j = b_j\theta_j$. Then

$$\begin{aligned} p(\theta_j \mid z_{1j}, \nu, \tau^2) &\propto p(z_{1j} \mid \theta_j)p(\theta_j \mid \nu, \tau^2) \\ &\propto \exp\left\{-\frac{1}{2}(z_{1j} - b_j\theta_j)^2 - \frac{1}{2\tau^2}(\theta_j - \nu)^2\right\} \\ &\propto \exp\left\{-\left(\frac{1}{2\sigma_j^2} + \frac{1}{2\tau^2}\right)\left(\theta_j^2 - 2\left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{1}{\sigma_j^2}\hat{\theta}_j + \frac{1}{\tau^2}\nu\right)\theta_j\right)\right\} \end{aligned}$$

where $\sigma_j^2 = 1/b_j^2$ and $\hat{\theta}_j = z_{1j}/b_j$. Note that

$$\left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{1}{\sigma_j^2}\hat{\theta}_j + \frac{1}{\tau^2}\nu\right) = \hat{\theta}_j - g(\sigma_j^2, \tau^2)(\hat{\theta}_j - \nu)$$

where $g(\sigma_j^2, \tau^2) = \frac{\sigma_j^2}{\sigma_j^2 + \tau^2}$, so

$$\begin{aligned} p(\theta_j \mid z_{1j}, \nu, \tau^2) &\propto \exp\left\{-\frac{1}{2g(\sigma_j^2, \tau^2)\tau^2}\left[\theta_j - \left(\hat{\theta}_j - g(\sigma_j^2, \tau^2)(\hat{\theta}_j - \nu)\right)\right]^2\right\} \\ &= \mathcal{N}\left\{\hat{\theta}_j - \frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\left(\hat{\theta}_j - \nu\right), \frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\tau^2\right\}. \end{aligned}$$

To obtain the posterior distribution of ν given $\boldsymbol{Z}_1 = \boldsymbol{z}_1$, we compute

$$\begin{aligned} p(\nu \mid \boldsymbol{z}_1, \tau^2, \phi, \omega^2) &\propto m(\boldsymbol{z}_1 \mid \nu, \tau^2) p(\nu \mid \phi, \omega^2) \\ &\propto \prod_{j=1}^{\ell} m(z_{1j} \mid \nu, \tau^2) p(\nu \mid \phi, \omega^2), \end{aligned}$$

where

$$\begin{aligned} m(z_{1j} \mid \nu, \tau^2) &\propto \int_{\Theta} p(z_{1j} \mid \theta_j) p(\theta_j \mid \nu, \tau^2) d\theta_j \\ &\propto \exp\left\{-\frac{1}{2} z_{1j}^2\right\} \exp\left\{\frac{1}{2}\left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{1}{\sigma_j^2}\hat{\theta}_j + \frac{1}{\tau^2}\nu\right)^2\right\} \\ &\quad \times \underbrace{\int_{\Theta} p(\theta_j \mid z_{1j}, \nu, \tau^2) d\theta_j}_{=1} \\ &\propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_j^2 + \tau^2}\left(\frac{z_{1j}}{b_j} - \nu\right)^2\right\} \\ &= \mathcal{N}\Big\{ b_j \nu, b_j^2 \left(\sigma_j^2 + \tau^2\right) \Big\}, \end{aligned}$$

and Θ denotes the parameter space for θ_j. Then,

$$\begin{aligned} p(\nu \mid \boldsymbol{z}_1, \tau^2, \phi, \omega^2) &\propto m(\boldsymbol{z}_1 \mid \nu, \tau^2) p(\nu \mid \phi, \omega^2) \\ &\propto \exp\left\{\sum_{i=1}^{\ell} -\frac{1}{2}\frac{1}{b_j^2(\sigma_j^2 + \tau^2)}(z_{1j} - b_j\nu)^2 - \frac{1}{2\omega^2}(\nu - \phi)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\sigma_{\nu|z}^{-2}\left[\nu - \sigma_{\nu|z}^2\left(\sum_{i=1}^{\ell}\frac{\hat{\theta}_j}{\sigma_j^2 + \tau^2} + \frac{\phi}{\omega^2}\right)\right]\right\}, \end{aligned}$$

where

$$\sigma_{\nu|z}^2 = \left(\sum_{j=1}^{\ell} \frac{1}{\mathcal{I}_{1j}^{-1} + \tau^2} + \frac{1}{\omega^2}\right)^{-1}.$$

We have hence shown that

$$\begin{aligned} \theta_j \mid z_{1j}, \nu, \tau^2 &\sim \mathcal{N}\left\{\hat{\theta}_{1j} - \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}\left(\hat{\theta}_{1j} - \nu\right), \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}\tau^2\right\}, \ j \in \mathcal{P}_1 \\ \nu \mid z_j, \tau^2, \phi, \omega^2 &\sim \mathcal{N}\left\{\sigma_{\nu|z}^2\left(\sum_{j=1}^{\ell}\frac{\hat{\theta}_{1j}}{\mathcal{I}_{1j}^{-1} + \tau^2} + \frac{\phi}{\omega^2}\right), \sigma_{\nu|z}^2\right\}, \end{aligned}$$

In the posterior mean of θ_j, we see the role that τ plays in strengthening/weakening evidence for subgroup specific effects:

$$\begin{aligned}
\hat{\theta}_{1j} - \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}\left(\hat{\theta}_{1j} - \nu\right) \quad &\underset{\tau \to \infty}{\to} \quad \hat{\theta}_{1j} \\
\hat{\theta}_{1j} - \frac{\mathcal{I}_{1j}^{-1}}{\mathcal{I}_{1j}^{-1} + \tau^2}\left(\hat{\theta}_{1j} - \nu\right) \quad &\underset{\tau \to 0}{\to} \quad \nu.
\end{aligned}$$

That is, if τ is large, the posterior mean for θ_j will tend towards $\hat{\theta}_j$, strongly implying heterogeneity. If τ is small, the posterior mean for every θ_j will tend towards ν.

4.3.2 Interim Analysis

We specify a loss (gain) function to quantify the cost (profit) of the available decisions, depending on true values of the parameters θ_j. We say that a population Ω_j is chosen if $j \in \mathcal{P}_2$, where $\mathcal{P}_2$ is the set of populations carried on to stage two. As in Chapter 3, let $\theta^-, \theta^+ \in \mathbb{R}$ such that $\theta^- < \theta^+$ and impose the following costs:

1. If $\theta_j \leq \theta^-$, we say that population Ω_j is not responsive, and a cost of c_1 is incurred if Ω_j is chosen;
2. If $\theta_j \geq \theta^+$, we say that population Ω_j is responsive, and a cost of c_2 is incurred if Ω_j is not chosen;
3. If $\theta_j \in (\theta^-, \theta^+)$, no costs are incurred either way.

A natural value for θ^- is zero or the null value of θ, while θ^+ could be set as proportional to the clinically significant treatment effect. Note that the terms "reject H_j" and "choose Ω_j" are not exchangeable. In the current setup, we can only reject hypotheses at the end of the trial, and all such decisions are carried out using frequentist analysis. It is at the interim analysis that we choose a population, and this decision is based purely on posterior expected loss, given first stage data.

Define the set of nonresponsive populations as $\mathcal{B} = \{j \in \mathcal{P}_1 \ : \ \theta_j \leq \theta^-\}$, and the set of responsive populations as $\mathcal{G} = \{j \in \mathcal{P}_1 \ : \ \theta_j \geq \theta^+\}$. Let $k = c_2/c_1$, representing the relative cost of false negatives compared to false positives. Thus, when $k < 1$ we will eliminate populations early unless results are good, while if $k > 1$ the cost for false negatives is higher and early elimination is less likely. Now, for a decision $\mathcal{P}_2 \subseteq \mathcal{P}_1$, we define our gain function as

$$G(\boldsymbol{\theta}, d) = -|\mathcal{P}_2 \setminus \mathcal{B}^c| - k|\mathcal{G} \setminus \mathcal{P}_2|.$$

The set $\mathcal{P}_2 \setminus \mathcal{B}^c$ contains all false positive decisions, i.e. selected populations Ω_j with $\theta_j \leq \theta^-$. Similarly, the set $\mathcal{G} \setminus \mathcal{P}_2$ contains all false negative decisions, i.e. responsive populations (with $\theta_j \geq \theta^+$) that were not selected.

At the interim analysis, we will determine the decision $\mathcal{P}_2$ such that the posterior expected gain (utility) is maximized. That is, we take

$$\mathcal{P}_2 = \arg\max_{\mathcal{S} \subseteq \mathcal{P}_1} \mathbb{E}\left[G(\boldsymbol{\theta}, \mathcal{S}) \mid \boldsymbol{Z}_1 = \boldsymbol{z}_1\right] =: \arg\max_{\mathcal{S} \subseteq \mathcal{P}_1} \mathcal{U}(\mathcal{S}, \boldsymbol{z}_1),$$

where $\boldsymbol{z}_1 \in \mathbb{R}^\ell$ is the observed data from stage one. Note that, for given $\mathcal{P}_2 \subseteq \mathcal{P}_1$,

$$\begin{aligned}
\mathcal{U}(\mathcal{P}_2, \boldsymbol{z}_1) &= \mathbb{E}\left[G(\boldsymbol{\theta}, \mathcal{S}) \mid \boldsymbol{Z}_1 = \boldsymbol{z}_1\right] \\
&= \mathbb{E}\left[-\sum_{j \in \mathcal{P}_2} 1\{\theta_j \leq \theta^-\} - k \sum_{j \notin \mathcal{P}_2} 1\{\theta_j \geq \theta^+\} \;\middle|\; \boldsymbol{Z}_1 = \boldsymbol{z}_1 \right] \\
&= -\sum_{j \in \mathcal{P}_2} \mathbb{P}[\theta_j \leq \theta^- \mid \boldsymbol{Z}_1 = \boldsymbol{z}_1] - k \sum_{j \notin \mathcal{P}_2} \mathbb{P}[\theta_j \geq \theta^+ \mid \boldsymbol{Z}_1 = \boldsymbol{z}_1]
\end{aligned}$$

As the parameters θ_j have a normal posterior distribution, computing the quantity $\mathcal{U}(\mathcal{P}_2, \boldsymbol{z}_1)$ is straightforward. Obtaining the posterior mean requires an estimate of ν, as

$$\hat{\mu}_{\theta_j|z} = \mathbb{E}[\theta_j | z_{1j}, \nu, \tau] = \hat{\theta}_j - \frac{\sigma_j^2}{\sigma_j^2 + \tau^2}(\hat{\theta}_j - \hat{\nu}), \ j \in \mathcal{P}_1.$$

The estimate $\hat{\nu}$ can be obtained in various ways, for example by using its posterior expected value, i.e.

$$\hat{\nu} = \left(\sum_{j=1}^{\ell} \frac{1}{\sigma_j^2 + \tau^2} + \frac{1}{\omega^2} \right)^{-1} \left(\sum_{j=1}^{\ell} \frac{\hat{\theta}_j}{\sigma_j^2 + \tau^2} + \frac{\phi}{\omega^2} \right).$$

4.3.3 Final Analysis

The final analysis can make use of any procedure that controls FWER strongly at level α. For example we could obtain adjusted p-values via the Hochberg or the Hommel method, see Chapter 2 for more detail. Here, we show two alternate methods that are somewhat less standard. The first strategy relies on conditional error rates, while the second approach involves modifications to the hypotheses of interest, and may entail a simpler final analysis.

Before proceeding, we introduce the following notation. For $\mathcal{S} \subseteq \mathcal{P}$, define $H_{\mathcal{S}}^{\cap}$ as the intersection hypothesis containing all hypotheses H_j for which $j \in \mathcal{S}$. That is, $H_{\mathcal{S}}^{\cap} = \bigcap\limits_{j \in \mathcal{S}} H_j$. Recall that $\theta_{\mathcal{S}}$ is the weighted average of all θ_j with $j \in \mathcal{S}$, and define $H_{\mathcal{S}}^{\cup}$ as the hypothesis of no overall treatment effect in $\Omega_{\mathcal{S}}$, that is $H_{\mathcal{S}}^{\cup} : \theta_{\mathcal{S}} \leq 0$.

Conditional Error Rates

We follow the approach used by Müller and Schäfer (2001), discussed in Section 2.2.4, whereby the conditional error function is defined in terms of a predefined test φ. Let $\varphi_{\mathcal{S}}$ denote a predefined levels α test for the hypothesis $H_{\mathcal{S}}^{\cap}$, $\mathcal{S} \subseteq \mathcal{P}_1$. If $H_{\mathcal{S}}^{\cap}$ is rejected (accepted), then $\varphi_{\mathcal{S}} = 1$ ($\varphi_{\mathcal{S}} = 0$).

Definition 4.3. *Let $\mathcal{S} \subseteq \mathcal{P}_1$. The conditional error rate (CER) for $H_{\mathcal{S}}^{\cap}$ is defined* *as*

$$A_{\mathcal{S}}(\boldsymbol{Z}_1) = \mathbb{E}_{H_{\mathcal{S}}^{\cap}}\left[\varphi_{\mathcal{S}} \mid Z_{1j}, j \in \mathcal{S}\right] = \mathbb{P}_{H_{\mathcal{S}}^{\cap}}\left[\textit{Reject } H_{\mathcal{S}}^{\cap} \mid Z_{1j}, j \in \mathcal{S}\right].$$

$A_{\mathcal{S}}(\boldsymbol{Z}_1)$ is a measurable function of the first stage results[2], taking values in [0,1], and satisfying $\mathbb{E}_{H_{\mathcal{S}}^{\cap}}[A(\boldsymbol{Z}_1)] \leq \alpha$. After the first stage is completed, we can compute $A_{\mathcal{S}}(\boldsymbol{Z}_1)$ for all $\mathcal{S} \subseteq \mathcal{P}_1$. We choose $\mathcal{P}_2$ based on the utility function discussed in the previous section, and following the second stage we obtain p-values $p_{2,\mathcal{S}}$, based only on second stage results. Hence $p_{2,\mathcal{S}}$ is p-clud (see Section 2.2.4) and testing $H_{\mathcal{S}}^{\cap}$ at level $A_{\mathcal{S}}(\boldsymbol{z}_1)$ ensures that Type I error is controlled at level α for $H_{\mathcal{S}}^{\cap}$. Applying the closure principle of Marcus et al. (1976) then implies that FWER is strongly protected at level α.

Following the above discussion, the CER combination procedure is carried out as follows:

1. Specify tests $\varphi_{\mathcal{S}}$ of $H_{\mathcal{S}}^{\cap}$ for all $\mathcal{S} \subseteq \mathcal{P}_1$.
2. After stage one, determine $\mathcal{P}_2 = \arg\max_{\mathcal{S} \subseteq \mathcal{P}_1} \mathcal{U}(\mathcal{S}, \boldsymbol{z}_1)$.
3. After stage two, for subsets $\mathcal{S}$ of $\mathcal{P}_1$:
   (a) If $\mathcal{S} \subseteq \mathcal{P}_2$, then test $H_{\mathcal{S}}^{\cap}$ with the preplanned test $\varphi_{\mathcal{S}}$.
   (b) If $\mathcal{S} \not\subseteq \mathcal{P}_2$, then compute $A_{\mathcal{S}}(\boldsymbol{Z}_1)$ and the p-value $p_{2,\mathcal{S}} := p_{2,\mathcal{S} \cap \mathcal{P}_2}$ and reject $H_{\mathcal{S}}^{\cap}$ if and only if $p_{2,\mathcal{S}} \leq A_{\mathcal{S}}(\boldsymbol{Z}_1)$. Per convention, $p_{2,\varnothing} = 1$.

There is a great deal of flexibility regarding the types of tests φ that can be used. Here, we give one example, similar to the approach taken by Koenig et al. (2008) in which many competing treatments are compared to a common control.

[2] Measurable with respect to the σ-algebra generated by the first stage results, see (Liu et al., 2002).

Specify weights r_j, $j \in \mathcal{P}_1$ such that $\sum_{j \in \mathcal{P}_1} r_j = 1$ and $\alpha_j = r_j \alpha$. Let $\mathcal{S} \subseteq \mathcal{P}_1$, and define adjusted weights $r_j^{\mathcal{S}}$, computed as

$$r_j^{\mathcal{S}} = r_j / r_{\cdot}^{\mathcal{S}}, \quad \text{where} \quad r_{\cdot}^{\mathcal{S}} := \sum_{j \in \mathcal{S}} r_j.$$

Now define critical values $C_j^{\mathcal{S}} = \Phi^{-1}\left(1 - r_j^{\mathcal{S}} \xi\right)$, where Φ^{-1} is the inverse normal CDF, and $\xi \in [0, 1]$ is determined numerically such that

$$1 - \alpha = \mathbb{P}_{H_{\mathcal{S}}^{\cap}}\left[Z_j < C_j^{\mathcal{S}}, \forall j \in \mathcal{S}\right].$$

The test $\varphi_{\mathcal{S}}$ rejects $H_{\mathcal{S}}^{\cap}$ if $Z_j \geq C_j^{\mathcal{S}}$ for some $j \in \mathcal{S}$. By construction it is a level α test. Based on this test, we can compute the CER for $H_{\mathcal{S}}^{\cap}$:

$$A_{\mathcal{S}}(\boldsymbol{z}_1) = \mathbb{P}_{H_{\mathcal{S}}^{\cap}}(\varphi_{\mathcal{S}} = 1 | \boldsymbol{Z}_1 = \boldsymbol{z}_1) = \mathbb{P}_{H_{\mathcal{S}}^{\cap}}\left[\max_{j \in \mathcal{S}}\left(Z_j - C_j^{\mathcal{S}}\right) \geq 0 \;\middle|\; z_{1,j}, j \in \mathcal{S}\right].$$

A simpler approach is to compute $1 - \alpha$ equicoordinate boundaries $D^{\mathcal{S}}$ of the $|\mathcal{S}|$-variate normal distribution with the appropriate correlation structure. In the above notation this is equivalent to setting $r_j = \alpha / |\mathcal{P}_1|$ for all $j \in \mathcal{P}_1$. Note that since we allow for early stopping due to futility, the procedure is conservative and the true FWER is strictly less than α. A common remedy is to simply adjust α via simulation ("buy back alpha" or "reclaim alpha") to make the procedure tight.

The final analysis relies on combining test statistics from different stages. We use the inverse normal combination method, see Equation (2.4) of Section 2.1.1. Combination weights are as defined for FE in Section 4.2.4, i.e.

$$\begin{aligned} w_{1j} &= \left(\frac{t f_{\mathcal{P}_1,j}}{t f_{\mathcal{P}_1,j} + (1-t) f_{\mathcal{P}_2,j}}\right)^{1/2} \\ w_{2j} &= \left(\frac{(1-t) f_{\mathcal{P}_2,j}}{t f_{\mathcal{P}_1,j} + (1-t) f_{\mathcal{P}_2,j}}\right)^{1/2}, \end{aligned}$$

for $j \in \mathcal{P}_2$. Finally, we specify how to obtain p-values $p_{2,\mathcal{S}}$ after the conclusion of the second stage. If $\mathcal{S} \subseteq \mathcal{P}_2$, then $H_{\mathcal{S}}^{\cap}$ is tested with the preplanned test $\varphi_{\mathcal{S}}$. If, however, $\mathcal{S} \nsubseteq \mathcal{P}_2$, then we use conditional second stage tests (Koenig et al., 2008), which reject $H_{\mathcal{S}}^{\cap}$ if

$$p_{2,\mathcal{S}} := \mathbb{P}_{H_{\mathcal{S}}^{\cap}} \left[\max_{j \in \mathcal{S} \cap \mathcal{P}_2} Z_j \geq \max_{j \in \mathcal{S} \cap \mathcal{P}_2} z_j \;\middle|\; z_{1j}, j \in \mathcal{S} \cap \mathcal{P}_2 \right] \leq A_{\mathcal{S}}(\boldsymbol{z}_1).$$

Equivalently, we can use a numerical search to find critical values $\tilde{C}_j^{\mathcal{S}} = \Phi^{-1}\left(1 - \tilde{r}_j^{\mathcal{S}} \xi\right)$, where $\tilde{r}_j^{\mathcal{S}} = r_j / \sum\limits_{j \in \mathcal{S} \cap \mathcal{P}_2} r_j$, such that

$$\mathbb{P}_{H_{\mathcal{S}}^{\cap}} \left[\max_{j \in \mathcal{S} \cap \mathcal{P}_2} \left(Z_j - \tilde{C}_j^{\mathcal{S}} \right) \geq 0 \;\middle|\; z_{1j}, j \in \mathcal{S} \cap \mathcal{P}_2 \right] = A_{\mathcal{S}}(\boldsymbol{z}_1),$$

and reject $H_{\mathcal{S}}^{\cap}$ if $\max\limits_{j \in \mathcal{S} \cap \mathcal{P}_2} \left(Z_j - \tilde{C}_j^{\mathcal{S}} \right) \geq 0$. We can see that using the preplanned test $\varphi_{\mathcal{S}}$ or $A_{\mathcal{S}}(\boldsymbol{Z}_1)$ yields the same rejection region when $\mathcal{S} \subseteq \mathcal{P}_2$.

Example 4.6. *Suppose that $\mathcal{P}_1 = \{1, 2\}$ so Ω_0 is partitioned into two smaller populations. If, at the interim stage, it is decided to continue only with Ω_1, then $\mathcal{P}_2 = \{1\}$. Hence,*

$$\begin{aligned} p_{2,\mathcal{S}} &= \mathbb{P}_{H_1}(Z_1 \geq z_1 \mid z_{11}) \\ &= \mathbb{P}_{H_1}\left(w_{11} z_{11} + w_{21} Z_{21} \geq w_{11} z_{11} + w_{21} z_{21}\right) = 1 - \Phi(z_{21}) =: p_2, \end{aligned}$$

for $\mathcal{S} = \{1\}$ and $\mathcal{S} = \{1, 2\}$. Now, $H_{12} = H_1 \cap H_2$ is rejected if $p_2 \leq A_{\{1,2\}}(\boldsymbol{z}_1)$. Then, H_1 is rejected if $p_2 \leq A_{\{1\}}(\boldsymbol{z}_1)$ **and** *if H_{12} was rejected. Note that since $\{1\} \subseteq \mathcal{P}_2$, we could have just tested H_1 using the originally planned test $\varphi_{\{1\}}$. This is seen by observing that for $\mathcal{S} \subseteq \mathcal{P}_2$, $\tilde{C}_j^{\mathcal{S}} = C_j^{\mathcal{S}}$.*

Population Combination

We consider an alternate methodology for selecting subgroups that is somewhat simpler than the CER principle discussed above. For $\mathcal{S} \subseteq \mathcal{P}_1$, recall that $\Omega_{\mathcal{S}} = \bigcup\limits_{j \in \mathcal{S}} \Omega_j$. Then $\Omega_{\mathcal{P}_2}$ is the overall population carried on to stage two after the interim analysis. Recall that $\theta_{\mathcal{S}}$ is the overall treatment effect in $\Omega_{\mathcal{S}}$, specified as the weighted average of θ_j for which $j \in \mathcal{S}$, see Equation 2.2 of Section 2.1.1. Now, the hypothesis of interest is

$$H^{\cup}_{\mathcal{P}_2} : \theta_{\mathcal{P}_2} \leq 0, \text{ no treatment effect in } \Omega_{\mathcal{P}_2}.$$

which is tested using the standardized statistic $Z_{\mathcal{P}_2}$ for $\Omega_{\mathcal{P}_2}$. Applying the closure principle, we must now test all intersection hypotheses $H^{\cap}_{\mathcal{S}}$ such that $\mathcal{P}_2 \subsetneq \mathcal{S} \subseteq \mathcal{P}_1$. Let $Z_{i,\mathcal{S}}$ denote the pooled standardized statistic for $\Omega_{\mathcal{S}}$ over stage i. Then $H^{\cap}_{\mathcal{S}}$ can be tested using the statistic

$$Z_{\mathcal{S}} = w_{1,\mathcal{S}} Z_{1,\mathcal{S}} + w_{2,\mathcal{S}\cap\mathcal{P}_2} Z_{2,\mathcal{S}\cap\mathcal{P}_2}, \tag{4.6}$$

where empirical data weights are given as

$$w_{1,\mathcal{S}} = \sqrt{\frac{tf_{0,\mathcal{S}}}{tf_{0,\mathcal{S}} + (1-t)}}, \quad \text{and} \quad w_{2,\mathcal{S}\cap\mathcal{P}_2} = \sqrt{\frac{1-t}{tf_{0,\mathcal{S}} + (1-t)}}.$$

If, instead of using the empirical weights, we use prespecified weights, then the procedure controls FWER strongly at level α. As previously noted in the thesis, and in (Wang et al., 2009), using empirical data weights may result in some inflation of error rates. With limited options for adaptation however, inflation is often negligible. If early stopping is enforced, the procedure is likely conservative, even if empirical data weights are used.

Note that this procedure simplifies the final analysis as we only consider one elementary hypothesis, $H_{\mathcal{P}_2}$. This can be both a virtue and a weakness. If we fail to reject any $H_{\mathcal{S}}$ where $\mathcal{P}_2 \subseteq \mathcal{S}$ then there is no room to consider smaller populations, $\mathcal{P}_2' \subsetneq \mathcal{P}_2$, without violating the closure principle. We see this by noting that the test statistics for $H^{\cup}_{\mathcal{P}_2}$ and $H^{\cap}_{\mathcal{P}_2}$, defined in Equation (4.6), are exactly the same. Hence, if $H^{\cup}_{\mathcal{P}_2}$ cannot be rejected, then neither can $H^{\cap}_{\mathcal{P}_2}$ in this setting. Note however, that rejection of $H^{\cup}_{\mathcal{P}_2}$ *does imply* rejection of $H^{\cap}_{\mathcal{P}_2}$. Hence, if

$H^{\cup}_{\mathcal{P}_2}$ is indeed rejected, we may proceed to investigate intersection hypotheses $H^{\cap}_{\mathcal{P}'_2}$ for $j \in \mathcal{P}_2$ in a step-down manner without inflating Type I error probability.

Chapter 5

K-Stage Group-Sequential Design with Subgroup Selection

In this chapter, we propose a K-stage sequential confirmatory design that incorporates the data-driven option to redefine the target population at the first interim analysis (the procedure name is abbreviated as GSDS). In practice, this design can be thought of as a seamless Phase II/III design, where the first stage represents Phase II, after which populations appearing to be non-responsive are dropped. Subsequent stages represent interim analyses in a large-scale confirmatory Phase III trial which is intended to prove treatment efficacy in the overall remaining population following Phase II.

To demonstrate the procedure in practice, we include two worked examples in Section 5.8. The first is an application to development of antidepressants, and the second is for the testing of a promising agent in oncology. In Section 5.9, two numerical examples are conducted to analyze basic operating characteristics of the GSDS design. Also, we correct for the selection bias that results from stage-wise testing. We conclude the chapter with a comparison of GSDS with the FE and HUT procedures, developed in chapters 3 and 4.

5.1 Setup

We have an experimental treatment that is to be tested on a population of interest Ω_0. Suppose that from earlier phases of development, or from previous clinical trials, there is evidence suggesting that Ω_0 can be partitioned into disjoint subsets

$\Omega_1, \ldots, \Omega_\ell$, and that treatment efficacy may differ across these smaller populations (subgroups). As before, define $\mathcal{P} = \{1, \ldots, \ell\}$ as the index set of subgroups under consideration. For $\mathcal{S} \subseteq \mathcal{P}$, define

$$\Omega_{\mathcal{S}} = \bigcup_{j \in \mathcal{S}} \Omega_j.$$

Let treatment efficacy in Ω_j be parameterized by θ_j, and define $\theta_{\mathcal{S}}$ as the weighted average treatment effect over populations Ω_j where $j \in \mathcal{S}$, and $\mathcal{S} \subseteq \mathcal{P}$. We plan to allow for a total of $K-1$ interim analyses.

We construct our tests in terms of the efficient score and observed Fisher's information. Referring to Chapter 2, for $k = 1, \ldots, K$, and $j \in \mathcal{P}$, define Y_{kj} and $\mathcal{I}_{kj}$ as the efficient score and the cumulative observed information for θ_j at stage k, respectively. Define the respective stage-wise increments as $X_{kj} = Y_{kj} - Y_{k-1,j}$ and $\Delta_{kj} = \mathcal{I}_{kj} - \mathcal{I}_{k-1,j}$. It can be shown that, asymptotically for small θ_j we have

$$X_{kj} \sim \mathcal{N}\left(\theta_j \Delta_{kj}, \Delta_{kj}\right), \ k = 1, \ldots, K \text{ and } j \in \mathcal{P}. \tag{5.1}$$

For now, we shall assume that the observation variance, σ^2, is known. Hence the relation in Equation (5.1) is exact. For composite populations $\Omega_{\mathcal{S}}$ with $\mathcal{S} \subseteq \mathcal{P}$, we define the efficient score, $Y_{k,\mathcal{S}}$, and the observed information $\mathcal{I}_{k,\mathcal{S}}$ in a similar manner.

5.2 Population Selection

At the first interim analysis, we allow non-responsive populations to be dropped from the trial. We propose two decision rules for use during the first interim analysis. The first rule applies when we can assume no *a priori* ordering on treatment effects. In other words, other than believing that populations may respond differently to the proposed treatment, we have no prior belief that any one subgroup is more responsive than another. The second rule applies when subgroups are ordered according to their expected treatment response. That is, the prior belief is that $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_\ell$ (this might be applicable for nested populations). For example, this could apply when testing an antidepressant believed to work best on severely depressed individuals. Both decision rules rely on the use of a stopping rule boundary l_1, which for now we will assume is given (see Section 5.4 for computational details.)

Decision Rules:

I For each $j = 1, \ldots, \ell$, if $X_{1j} \leq l_1\sqrt{\mathcal{I}_{1j}}$, then population Ω_j is eliminated from the trial. If $\mathcal{S}$ denotes the index set of retained populations, we know that $Y_{1,\mathcal{S}} > \sum_{j\in\mathcal{S}} l_1\sqrt{\mathcal{I}_{1j}} =: \tilde{l}_{1,\mathcal{S}}$. Also note that $\sum_{j\in\mathcal{S}}\sqrt{\mathcal{I}_{1j}} \geq \sqrt{\sum_{j\in\mathcal{S}}\mathcal{I}_{1j}} = \sqrt{\mathcal{I}_{1,\mathcal{S}}}$.

II Starting at Ω_ℓ and searching backwards (through the indices), find the first Ω_j for which $X_{1j} \geq l_1\sqrt{\mathcal{I}_{1j}}$. Call this population Ω_r, eliminate all Ω_j for $j > r$ and retain all populations for which $j \leq r$.

We refer to these decision rules as DR-I and DR-II. In each case, the remaining populations, indexed by $\mathcal{P}^* \subseteq \mathcal{P}$ say, are pooled together and the hypothesis of interest becomes $H_{\mathcal{P}^*} : \theta_{\mathcal{P}^*} = 0$[1]. We also enrich the chosen subgroups, so all planned observations for remaining stages may be allocated to $\Omega_{\mathcal{P}^*}$. Note that enrichment is not required, but intentions should be prospectively specified. We shall see that computation of stopping boundaries requires knowledge of planned information accumulation at each stage.

[1] Note that this is the hypothesis $H^{\cup}_{\mathcal{P}^*}$ which was defined in Section 4.3. However, as there is no ambiguity in the notation in this chapter, we simplify our notation by removing the "∪" from the superscript.

If populations do not admit a natural ordering, then DR-I may be applied. In this case however, it is possible that subgroups are selected in such a way as to make the resulting composite population appear biologically implausible. To guard against this, the decision rules could be combined to guide population selection in a "medically reasonable" way. This may be done by partitioning $\mathcal{P}$ into $q < \ell$ disjoint subsets $\mathcal{P}_1, \ldots, \mathcal{P}_q$, and imposing an ordering within each of these subsets. Then within each subset $\mathcal{P}_i$, for $i = 1, \ldots, q$, DR-II is used to determine the population $\Omega_{\mathcal{S}_i}$, $\mathcal{S}_i \subseteq \mathcal{P}_i$, that will be carried on to stage two. The overall population taken to stage two is thus $\Omega_{\mathcal{S}_1} \cup \cdots \cup \Omega_{\mathcal{S}_q}$.

5.3 Control of the FWER

We aim to control Type I error rates at level α. Note that since the treatment effect in $\Omega_{\mathcal{S}}$ is a weighted average, $\theta_{\mathcal{S}} = 0$ only if $\theta_j = 0$ for all $j \in \mathcal{S}$. Hence, false positives can only occur when we have selected a composite population in which all $\theta_j = 0$. Note that null hypotheses are simple rather than composite, i.e. $H : \theta = 0$ rather than $H : \theta \leq 0$. This is done so as to eliminate the possible presence of opposite direction treatment effects, which are believed to be very unlikely (see discussion in Section 1.2).

To control FWER strongly at level α, we must ensure that no combination of ineffective populations is chosen, and the resulting null hypothesis is rejected, with probability greater than α. It is easy to see that the probability of selecting a population $\mathcal{S}$ for which $\theta_{\mathcal{S}} = 0$ at the first interim analysis is maximized when all populations are nonresponsive, i.e. $\theta_j = 0$ for all $j \in \mathcal{P}$. (Else we would have a non-zero probability of selecting a composite population with positive treatment effect.) Hence, in order to protect FWER strongly at level α, it suffices to obtain decision boundaries such that Type I error probability is bounded above by α when the global null hypothesis $H_0 : \theta_1 = \cdots = \theta_\ell = 0$ is true.

The decision rules employed at the first interim analysis mean that information levels for subsequent stages depend on stage one observations $Y_{1j},\ j \in \mathcal{P}$. In (Jennison and Turnbull, 2000, Ch. 7.4), an example is given where such dependence can result in inflated Type I error. However, this inflation occurs as a result of skipping an interim analysis if early results are unimpressive, and the critical boundaries that are used in the example are not computed to account for the exact decision rules. Hence Type I error probability is not guaranteed to be less than α. In our procedure however, we specifically evaluate the probability of rejection for any combination of populations. These probabilities are then summed up to obtain the marginal probability of rejection, and a numerical search obtains stopping boundaries that ensure that probability of false rejection is no larger than α for the whole trial.

As currently stated, the decision rules introduced in Section 5.2 only allow populations to be discarded early for futility. In particular, it is not permitted to reject the null hypothesis for an individual population Ω_j and then to proceed with the remaining populations to later stages. There are two reasons for imposing this restriction. First, as we see in the next section, the fact that we need only be concerned with the global null hypothesis greatly simplifies computation of stopping boundaries. If we allowed early rejection of individual populations, then protecting FWER strongly would be significantly more complicated. Second, Wang et al. (2007b) caution against early rejection of small subgroups even if they appear promising; early findings are often based on sample sizes that are too small to assert with confidence that a treatment is effective in such populations. On the other hand, if early results appear very negative (or potentially harmful), we have an ethical obligation not to unnecessarily expose patients to a treatment that is unlikely to be of any benefit. Therefore it is desirable to eliminate populations that appear to be nonresponsive at a reasonably early point in the trial.

5.4 Construction of Sequential Stopping Rules

Let $\mathcal{P}^*$ be the random index set of populations chosen after the first stage. Suppose we have planned interim analysis times $\{t_k\}_{k=0}^K$, such that $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$. A stopping rule is constructed via the use of spending functions (Lan and DeMets, 1983), defined in terms of calendar times (Lan and DeMets, 1989). We compute upper and lower boundaries $(l_k, u_k)_{k=1}^K$ with $l_k \leq u_k$ for all k, where crossing the lower boundary results in termination of the trial and acceptance of all hypotheses, while crossing the upper boundary results in termination with rejection of $H_{\mathcal{P}^*}$. To ensure trial termination at stage K, we require that $l_K = u_K$.

As we use the efficient score as our test statistic, actual rejection and acceptance boundaries will depend on information already observed. To this end, define adjusted boundaries

$$\tilde{l}_{k,\mathcal{S}} := l_k \sqrt{\mathcal{I}_{k,\mathcal{S}}} \text{ and } \tilde{u}_{k,\mathcal{S}} := u_k \sqrt{\mathcal{I}_{k,\mathcal{S}}} \text{ for } k = 1, \ldots, K, \ \mathcal{S} \subseteq \mathcal{P}.$$

If $Y_{k,\mathcal{S}} \geq \tilde{u}_{k,\mathcal{S}}$ ($Y_{k,\mathcal{S}} \leq \tilde{l}_{k,\mathcal{S}}$) for some k, then $H_{\mathcal{S}}$ is rejected (accepted) and the trial stops. Else, if $Y_{k,\mathcal{S}} \in \left(\tilde{l}_{k,\mathcal{S}}, \tilde{u}_{k,\mathcal{S}}\right)$, the trial proceeds to stage $k+1$. We derive boundaries using the approach of Stallard and Facey (1996), in which two spending functions are defined under the assumption that the global null hypothesis is true. Let $\alpha_U^* : [0,1] \to [0,\alpha]$ and $\alpha_L^* : [0,1] \to [0, 1-\alpha]$ be non-decreasing functions with

$\alpha_U^*(0) = \alpha_L^*(0) = 0$, $\alpha_U^*(1) = \alpha$, and $\alpha_L^*(1) = 1 - \alpha$. Let $\boldsymbol{\theta} \in \mathbb{R}^\ell$ denote a particular treatment efficacy configuration, and define (for $\mathcal{S} \subseteq \mathcal{P}$)

$$\begin{aligned}
&\psi_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right) \\
&\quad := \mathbb{P}_{\boldsymbol{\theta}}[\text{Select } \mathcal{S} \text{ and reject } H_{\mathcal{S}} \text{ exactly at stage } k] \\
&\quad = \mathbb{P}_{\boldsymbol{\theta}}\left[\mathcal{P}^* = \mathcal{S}, Y_{1,\mathcal{S}} \in \left(\tilde{l}_{1,\mathcal{S}}, \tilde{u}_{1,\mathcal{S}}\right), \ldots, Y_{k-1,\mathcal{S}} \in \left(\tilde{l}_{k-1,\mathcal{S}}, \tilde{u}_{k-1,\mathcal{S}}\right), Y_{k,\mathcal{S}} \geq \tilde{u}_{k,\mathcal{S}}\right],
\end{aligned} \tag{5.2}$$

and

$$\begin{aligned}
&\xi_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right) \\
&\quad := \mathbb{P}_{\boldsymbol{\theta}}[\text{Select } \mathcal{S} \text{ and accept } H_{\mathcal{S}} \text{ exactly at stage } k] \\
&\quad = \mathbb{P}_{\boldsymbol{\theta}}\left[\mathcal{P}^* = \mathcal{S}, Y_{1,\mathcal{S}} \in \left(\tilde{l}_{1,\mathcal{S}}, \tilde{u}_{1,\mathcal{S}}\right), \ldots, Y_{k-1,\mathcal{S}} \in \left(\tilde{l}_{k-1,\mathcal{S}}, \tilde{u}_{k-1,\mathcal{S}}\right), Y_{k,\mathcal{S}} \leq \tilde{l}_{k,\mathcal{S}}\right],
\end{aligned} \tag{5.3}$$

where $\mathbb{P}_{\boldsymbol{\theta}}$ indicates that the probability is computed under the configuration $\boldsymbol{\theta}$. By summing over subsets $\mathcal{S} \subseteq \mathcal{P}$, we get marginal stopping probabilities for a given stage k, *viz.*

$$\begin{aligned}
\psi_k\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right) &:= \mathbb{P}_{\boldsymbol{\theta}}[\text{Stop trial exactly at stage } k \text{ with rejection}] \\
&= \sum_{\mathcal{S} \subseteq \mathcal{P}} \psi_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right),
\end{aligned}$$

and

$$\begin{aligned}
\xi_k\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right) &:= \mathbb{P}_{\boldsymbol{\theta}}[\text{Stop trial exactly at stage } k \text{ with no rejection}] \\
&= \sum_{\mathcal{S} \subseteq \mathcal{P}} \xi_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right).
\end{aligned}$$

Upper and lower boundaries are now obtained by recursively solving the expressions

$$\psi_k\left(l_1, u_1, \ldots, l_k, u_k; \mathbf{0}\right) = \alpha_U^*(t_k) - \alpha_U^*(t_{k-1}) \tag{5.4}$$

and

$$\xi_k\left(l_1, u_1, \ldots, l_k, u_k; \mathbf{0}\right) = \alpha_L^*(t_k) - \alpha_L^*(t_{k-1}), \tag{5.5}$$

where $\mathbf{0}$ is an ℓ-vector of zeroes. Hence, both upper and lower boundaries are obtained by evaluating path probabilities under the complete null hypothesis. The definition of functions $\alpha_U^*(t)$ and $\alpha_L^*(t)$ ensures that the boundaries meet at stage K, forcing the trial to end. To ensure a desired power, a numerical search for required observed information is required. This is discussed in more detail in Section 5.5.

We can specify probabilities $\psi_{k,\mathcal{S}}$ and $\xi_{k,\mathcal{S}}$ in terms of iterated integrals. First, given $\mathcal{S} \subseteq \mathcal{P}$, we write

$$\psi_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right) = \mathbb{P}_{\boldsymbol{\theta}}\left[\mathcal{P}^* = \mathcal{S}\right] \times \tilde{\psi}_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right), \tag{5.6}$$

where

$$\begin{aligned}&\tilde{\psi}_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right)\\ &= \mathbb{P}_{\boldsymbol{\theta}}\left[Y_{1,\mathcal{S}} \in \left(\tilde{l}_{1,\mathcal{S}}, \tilde{u}_{1,\mathcal{S}}\right), \ldots, Y_{k-1,\mathcal{S}} \in \left(\tilde{l}_{k-1,\mathcal{S}}, \tilde{u}_{k-1,\mathcal{S}}\right), Y_{k,\mathcal{S}} \geq \tilde{u}_{k,\mathcal{S}} \mid \mathcal{P}^* = \mathcal{S}\right]\end{aligned}$$

A key insight is that the sequence of random variables $Y_{k,\mathcal{S}}$ is Markovian, so the conditional distribution of $Y_{k,\mathcal{S}}$ given $Y_{1,\mathcal{S}} = y_{1,\mathcal{S}}, \ldots, Y_{k-1,\mathcal{S}} = y_{k-1,\mathcal{S}}$ depends only on y_{k-1}. Hence, for each $k = 2, \ldots, K$,

$$\begin{aligned}&\tilde{\psi}_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right)\\ &= \mathbb{P}_{\boldsymbol{\theta}}\left[Y_{1,\mathcal{S}} \in \left(\tilde{l}_{1,\mathcal{S}}, \tilde{u}_{1,\mathcal{S}}\right), \ldots, Y_{k-1,\mathcal{S}} \in \left(\tilde{l}_{k-1,\mathcal{S}}, \tilde{u}_{k-1,\mathcal{S}}\right), Y_{k,\mathcal{S}} \geq \tilde{u}_{k,\mathcal{S}} \mid \mathcal{P}^* = \mathcal{S}\right]\\ &= \int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} \cdots \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} \int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta}) f_2(y_{2,\mathcal{S}}|y_{1,\mathcal{S}}; \boldsymbol{\theta}) \cdots f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}}; \boldsymbol{\theta}) \prod_{i=1}^{k} dy_{i,\mathcal{S}},\end{aligned} \tag{5.7}$$

where $f_i(y_{i,\mathcal{S}}|y_{i-1,\mathcal{S}}; \boldsymbol{\theta})$ is the conditional density of $Y_{i,\mathcal{S}}$ given $\mathcal{P}^* = \mathcal{S}$ and $Y_{i-1,\mathcal{S}} = y_{i-1,\mathcal{S}}$ for $2 \leq i \leq k$, and $f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta})$ is the conditional density of $Y_{1,\mathcal{S}}$ given $\mathcal{P}^* = \mathcal{S}$.

Similarly,

$$\begin{aligned}
&\tilde{\xi}_{k,\mathcal{S}}\left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}}\Big[Y_{1,\mathcal{S}} \in \big(\tilde{l}_{1,\mathcal{S}}, \tilde{u}_{1,\mathcal{S}}\big), \ldots, Y_{k-1,\mathcal{S}} \in \big(\tilde{l}_{k-1,\mathcal{S}}, \tilde{u}_{k-1,\mathcal{S}}\big), Y_{k,\mathcal{S}} \leq \tilde{l}_{k,\mathcal{S}} \mid \mathcal{P}^* = \mathcal{S}\Big] \\
&= \int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} \cdots \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} \int_{-\infty}^{\tilde{l}_{k,\mathcal{S}}} f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta}) f_2(y_{2,\mathcal{S}}|y_{1,\mathcal{S}};\boldsymbol{\theta}) \cdots f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta}) \prod_{i=1}^{k} dy_{i,\mathcal{S}}.
\end{aligned} \tag{5.8}$$

For $i = 2, \ldots, k$, the conditional densities $f_i(y_{i,\mathcal{S}}|y_{i-1,\mathcal{S}};\boldsymbol{\theta})$ are normal with mean $y_{i-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{i,\mathcal{S}}$ and variance $\Delta_{i,\mathcal{S}}$. That is,

$$\begin{aligned}
f_i(y_{i,\mathcal{S}}|y_{i-1,\mathcal{S}};\boldsymbol{\theta}) &\equiv f_{Y_{i,\mathcal{P}^*}|\mathcal{P}^*=\mathcal{S}}(y_{i,\mathcal{P}^*}|y_{i-1,\mathcal{P}^*}, \mathcal{P}^* = \mathcal{S};\boldsymbol{\theta}) \\
&= \frac{1}{\sqrt{\Delta_{i,\mathcal{S}}}}\varphi\left(\frac{y_{i,\mathcal{S}} - (y_{i-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{i,\mathcal{S}})}{\sqrt{\Delta_{i,\mathcal{S}}}}\right),
\end{aligned} \tag{5.9}$$

where $\varphi(\cdot)$ is the standard normal density. What remains then is to derive the density

$$f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta}) \equiv f_{Y_{1,\mathcal{P}^*}|\mathcal{P}^*=\mathcal{S}}(y_{1,\mathcal{P}^*}|\mathcal{P}^* = \mathcal{S};\boldsymbol{\theta}),$$

the form of which depends on the decision rule that is followed at the first interim analysis.

5.4.1 Derivation of f_1 with respect to DR-I

For $j \in \mathcal{P}$, we have $X_{1j} \sim \mathcal{N}\left(\nu_j, \varsigma_j^2\right)$, where $\nu_j = \theta_j\Delta_{1j}$, $\varsigma_j^2 = \Delta_{1j}$. Write $l_{1j} = l_1\sqrt{\Delta_{1j}}$. Then, define the density of $X_{1j}|X_{1j} > l_{1j}$ as

$$\begin{aligned}
h_j(x_j) &:= \left[\mathbb{P}(X_{1j} > l_{1j})\right]^{-1} 1\{x_j > l_{1j}\}\frac{1}{\varsigma_j}\varphi\left(\frac{x_j - \nu_j}{\varsigma_j}\right) \\
&= \left[\Phi\left(\frac{\nu_j - l_{1j}}{\varsigma_j}\right)\right]^{-1} 1\{x_j > l_{1j}\}\frac{1}{\varsigma_j}\varphi\left(\frac{x_j - \nu_j}{\varsigma_j}\right).
\end{aligned}$$

Without loss of generality, suppose that the first r populations were chosen, and the rest were eliminated. (If this were not the case we can easily relabel the populations to make it so.) Now define the partial sum

$$\tilde{Y}_i = \sum_{j=1}^{i} X_{1j},$$

and let $g_i(y_i)$ denote the density of $\tilde{Y}_i$. Using convolution, we can write the densities g_i for $i = 2, \ldots, r$:

$$\begin{aligned} g_2(y_2) &= \int_{-\infty}^{\infty} h_1(x_1)h_2(y_2 - x_1)dx_1 \\ &= \int_{l_{11}}^{y_2 - l_{12}} h_1(x_1)h_2(y_2 - x_1)dx_1, \end{aligned}$$

where the integration limits arise from the fact that $h_1(x_1) > 0$ if and only if $x_1 > l_{11}$, and $h_2(y_2 - x_1) > 0$ if and only if $y_2 - x_1 > l_{12}$. Similarly,

$$\begin{aligned} g_3(y_3) &= \int_{l_{11}+l_{12}}^{y_3 - l_{13}} g_2(y_2)h_3(y_3 - y_2)dy_2, \\ &\vdots \\ g_r(y_r) &= \int_{\sum_{j=1}^{r-1} l_{1j}}^{y_r - l_{1r}} g_{r-1}(y_{r-1})h_r(y_r - y_{r-1})dy_{r-1}. \end{aligned}$$

Hence, the density of $Y_{1,\mathcal{S}}$, given $\mathcal{P}^* = \mathcal{S}$, can be written as a $(r-1)$-fold multiple integral:

$$f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta}) = \int\limits_{l_{r-1,\mathcal{S}}}^{y_{1,\mathcal{S}} - l_{1r}} \int\limits_{l_{r-2,\mathcal{S}}}^{y_{r-1} - l_{1,r-1}} \cdots \int\limits_{l_{1,\mathcal{S}}}^{y_2 - l_{12}} m_r(y_{1,\mathcal{S}}, y_{r-1}, \ldots, y_1) \prod_{i=1}^{r-1} dy_i, \qquad (5.10)$$

where

$$m_r(y_{1,\mathcal{S}}, y_{r-1}, \ldots, y_1) = h_r(y_{1,\mathcal{S}} - y_{r-1})h_{r-1}(y_{r-1} - y_{r-2}) \cdots h_2(y_2 - y_1)h_1(y_1),$$

and $l_{i,\mathcal{S}} := \sum_{j=1}^{i} l_{1j} = l_1 \sum_{j=1}^{i} \sqrt{\Delta_{1j}}$. If we want non-zero probability of stopping for rejection at the first interim analysis, then u_1 will be finite, and this probability can be written as

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}\left(Y_{1,\mathcal{S}} \geq \tilde{u}_{1,\mathcal{S}} | \mathcal{P}^* = \mathcal{S}\right) &= \int_{\tilde{u}_{1,\mathcal{S}}}^{\infty} \int_{l_{r-1,\mathcal{S}}}^{y_r - l_{1r}} g_{r-1}(y_{r-1}) h_r(y_r - y_{r-1}) dy_{r-1} dy_r \\ = & \int_{l_{r-1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}} - l_{1r}} g_{r-1}(y_{r-1}) \left(\int_{\tilde{u}_{1,\mathcal{S}}}^{\infty} h_r(y_r - y_{r-1}) dy_r \right) dy_{r-1} \\ & + \int_{\tilde{u}_{1,\mathcal{S}} - l_{1r}}^{\infty} g_{r-1}(y_{r-1}) \left(\int_{y_{r-1}+l_{1r}}^{\infty} h_r(y_r - y_{r-1}) dy_r \right) dy_{r-1}.\end{aligned}$$

Define

$$\begin{aligned}e_{r-1}\left(y_{r-1}, \tilde{u}_{1,\mathcal{S}}; \boldsymbol{\theta}\right) &= \int_{\tilde{u}_{1,\mathcal{S}}}^{\infty} h_r(y_r - y_{r-1}) dy_r = \int_{\tilde{u}_{1,\mathcal{S}} - y_{r-1}}^{\infty} h_r(z) dz \\ &= \mathbb{P}_{\boldsymbol{\theta}}\left(X_{1r} \geq \tilde{u}_{1,\mathcal{S}} - y_{r-1} \mid X_{1r} > l_{1r}\right).\end{aligned}$$

Since $y_{r-1} < \tilde{u}_{1,\mathcal{S}} - l_{1r}$, $X_{1r} \geq \tilde{u}_{1,\mathcal{S}} - y_{r-1}$ implies that $X_{1r} > l_{1r}$, and hence

$$e_{r-1}\left(y_{r-1}, \tilde{u}_{1,\mathcal{S}}; \boldsymbol{\theta}\right) = \left[\Phi\left(\frac{\nu_r - l_{1r}}{\varsigma_r}\right)\right]^{-1} \Phi\left(\frac{\nu_r - (\tilde{u}_{1,\mathcal{S}} - y_{r-1})}{\varsigma_r}\right).$$

Also note that by setting $z = y_r - y_{r-1}$, we get

$$\begin{aligned}\int_{y_{r-1}+l_{1r}}^{\infty} h_r(y_r - y_{r-1}) dy_r &= \int_{l_{1r}}^{\infty} h_r(z) dz \\ &= \mathbb{P}_{\boldsymbol{\theta}}(X_{1r} > l_{1r} | X_{1r} > l_{1r}) = 1.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}\left(Y_{1,\mathcal{S}} \geq \tilde{u}_{1,\mathcal{S}} | \mathcal{P}^* = \mathcal{S}\right) &= \int_{l_{r-1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}} - l_{1r}} g_{r-1}(y_{r-1}) e_{r-1}\left(y_{r-1}, \tilde{u}_{1,\mathcal{S}}; \boldsymbol{\theta}\right) dy_{r-1} \\ &\quad + \int_{\tilde{u}_{1,\mathcal{S}} - l_{1r}}^{\infty} g_{r-1}(y_{r-1}) dy_{r-1}, \qquad (5.11)\end{aligned}$$

which can be computed recursively using numerical integration.

5.4.2 Derivation of f_1 with respect to DR-II

Deriving the density $f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta})$ is quite a bit simpler when DR-II is used, as the ordering assumption only restricts the values that X_{1r} can take on. Suppose we are given $\mathcal{P}^* = \mathcal{S} \subseteq \mathcal{P}$ with a total of r populations chosen. Then we know that $X_{1r} > l_1\sqrt{\Delta_{1r}} = l_{1r}$, and its density is $h_r(x_r)$ as defined above. As there are no restrictions on the values of $X_{11}, \dots, X_{1,r-1}$, the partial sum $\tilde{Y}_{r-1}$ is normally distributed with mean $\tilde{\nu}_{\mathcal{S}} = \theta_{\mathcal{S}\setminus\{r\}}\Delta_{1,\mathcal{S}\setminus\{r\}}$ and variance $\tilde{\varsigma}^2_{\mathcal{S}} = \Delta_{1,\mathcal{S}\setminus\{r\}}$. Denote its density by $\tilde{h}_{r-1}(\cdot)$. In the notation introduced in the derivation for DR-I, the density of $Y_{1,\mathcal{S}}$ is then

$$f_1(y_{1,\mathcal{S}}|\boldsymbol{\theta}) = \int_{l_{1r}}^{\infty} h_r(x_r)\tilde{h}_{r-1}(y_{1,\mathcal{S}} - x_r)dx_r, \; y_{1,\mathcal{S}} \in \mathbb{R}. \tag{5.12}$$

Hence, if we allow non-zero probability of termination with rejection after stage one, this can be expressed as

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}}\left(Y_{1,\mathcal{S}} > \tilde{u}_{1,\mathcal{S}}|\mathcal{P}^* = \mathcal{S}\right) &= \int_{\tilde{u}_{1,\mathcal{S}}}^{\infty}\int_{l_{1r}}^{\infty} h_r(x_r)\tilde{h}_{r-1}(y_r - x_r)dx_r dy_r \\
&= \int_{l_{1r}}^{\infty} h_r(x_r)\int_{\tilde{u}_{1,\mathcal{S}}-x_r}^{\infty}\tilde{h}_{r-1}(z)dz dx_r \\
&= \int_{l_{1r}}^{\infty} h_r(x_r)\Phi\left(\frac{\tilde{\nu}_{\mathcal{S}} - (\tilde{u}_{1,\mathcal{S}} - x_r)}{\tilde{\varsigma}_{\mathcal{S}}}\right)dx_r,
\end{aligned}$$

where, after changing the order of integration in the second line, the substitution $z = y_r - x_r$ was performed to yield the final result.

5.5 Power and Maximum Information

Our specification of stopping boundaries ensures that the Type I error probability in this design is bounded above by α. To achieve desired probability of a positive result, we must conduct a numerical search for the required information. As there are many populations under consideration, the term "power" is not well defined. We therefore list a number of treatment efficacy configurations under which one might wish to achieve a certain rejection probability. Denote a clinical treatment difference as θ^*, and let $\beta \in (0,1)$ be given. Let

$$\mathcal{I}_{\max} = \sum_{k=1}^{K} (\mathcal{I}_k - \mathcal{I}_{k-1}) = \sum_{k=1}^{K} \Delta_k$$

denote the maximum cumulative observed information, where $\mathcal{I}_k$ and Δ_k are the cumulative and stage-wise observed information levels, respectively. (If σ is known, we can think of this as proportional to the maximum sample size that the trial might require before a conclusion is reached.)

1. Find $\mathcal{I}_{\max}$ to ensure that

   $$1 - \mathbb{P}[\text{Accept all } H_j \mid \theta_j = \theta^*, \ \forall j \in \mathcal{P}] = 1 - \beta,$$

   that is

   $$\sum_{\mathcal{S} \subseteq \mathcal{P}} \sum_{k=1}^{K} \psi_{k,\mathcal{S}} \left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}^* \right) = 1 - \beta. \tag{5.13}$$

   This guarantees a "positive result" with probability $1 - \beta$ assuming all subgroups enjoy a clinically significant benefit, but might be underpowered to yield stronger results involving many subgroups.

2. Find $\mathcal{I}_{\max}$ to ensure that

   $$\mathbb{P}[\text{Reject } H_0 \text{ (all subgroups)} \mid \theta_j = \theta^*, \ \forall j \in \mathcal{P}] = 1 - \beta,$$

   $$\iff \sum_{k=1}^{K} \psi_{k,\mathcal{P}} \left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}^* \right) = 1 - \beta \tag{5.14}$$

   This gives a "complete result" with probability $1 - \beta$, and makes positive results involving some but not all subgroups quite likely. However, this power requirement might be too expensive in terms of required sample size.

3. Fix $\mathcal{P}^* \subseteq \mathcal{P}$ and suppose $\theta_j = \theta^*$ for $j \in \mathcal{P}^*$ and $\theta_j = 0$ elsewhere. We can find $\mathcal{I}_{\max}$ to ensure that

   $$\mathbb{P}[\text{Reject } H_{\mathcal{S}}, \text{ some } \mathcal{P}^* \subseteq \mathcal{S} \subseteq \mathcal{P} \mid \theta_j = \theta^*, j \in \mathcal{P}^* \text{ and } \theta_j = 0 \text{ else}] = 1 - \beta,$$

which is equivalent to

$$\sum_{\mathcal{P}^* \subseteq \mathcal{S} \subseteq \mathcal{P}} \sum_{k=1}^{K} \psi_{k,\mathcal{S}} \left(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}^*\right) = 1 - \beta \tag{5.15}$$

This might be appropriate if we have a subset $\mathcal{P}^* \subseteq \mathcal{P}$ of populations of particular interest, and guarantees a positive result with probability $1 - \beta$ when only a portion of the overall population is responsive.

An additional option is to search for $\mathcal{I}_{\max}$ such that these power requirements are satisfied with an average probability of $1 - \beta$.

5.6 Sample Size Distribution

Let $\mathcal{I}_{\text{term}}$ denote the observed cumulative information upon termination of a clinical trial. Then $\mathcal{I}_{\text{term}}$ is a random variable, whose expectation can be computed. If we assume that planned stage-wise information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are fixed before the trial starts, then the expected information on termination is

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\mathcal{I}_{\text{term}}\right] = \sum_{\mathcal{S} \subseteq \mathcal{P}} \left\{ \sum_{k=1}^{K} \left[\psi_{k,\mathcal{S}}(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta}) + \xi_{k,\mathcal{S}}(l_1, u_1, \ldots, l_k, u_k; \boldsymbol{\theta})\right] \mathcal{I}_k \right\}, \tag{5.16}$$

where quantities $\psi_{k,\mathcal{S}}$ and $\xi_{k,\mathcal{S}}$ are defined in Equations (5.2) and (5.3).

Note that this definition of expected information includes the information that is “discarded” after nonresponsive populations are eliminated during the first stage. If we prefer that expected information refers only to $\Omega_{\mathcal{S}}$, we can replace $\mathcal{I}_k$ with $\mathcal{I}_{k,\mathcal{S}}$ in Equation (5.16).

5.7 Point Estimation

In this section we consider point estimation of the parameter $\theta_{\mathcal{S}}$ where $\mathcal{P}^* = \mathcal{S} \subseteq \mathcal{P}$ is the index set of retained populations after stage one. Let $\mathcal{C}_{k,\mathcal{S}} = (\tilde{l}_{k,\mathcal{S}}, \tilde{u}_{k,\mathcal{S}})$ be the continuation region at stage k. If we assume that $\mathcal{S}$ and k fix $\mathcal{I}_{k,\mathcal{S}}$, we can express the result of a trial with the statistic $(\mathcal{S}, T, Y_{T,\mathcal{S}})$, where

$$T = \min\{k : Y_{k,\mathcal{S}} \notin \mathcal{C}_{k,\mathcal{S}}\},$$

i.e. T is the stage at which a stopping boundary was crossed. The form of the density for $Y_{1,\mathcal{P}^*}|\mathcal{P}^* = \mathcal{S}$ does not easily imply a sufficient statistic for $\theta_{\mathcal{S}}$. However, we can estimate $\theta_{\mathcal{S}}$ (conditioning on $\mathcal{P}^* = \mathcal{S}$) using the MLE from group sequential designs, $\hat{\theta}_{\mathcal{S}} = Y_{T,\mathcal{S}}/\mathcal{I}_{T,\mathcal{S}}$, see for example (Jennison and Turnbull, 2000, Ch. 8.2). The conditional bias can be estimated by computing $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_{\mathcal{P}^*} - \theta_{\mathcal{P}^*}|\mathcal{P}^* = \mathcal{S}\right]$, as follows. Let $p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta})$ denote the density of $Y_{k,\mathcal{P}^*}$ given that $\mathcal{P}^* = \mathcal{S}$. That is, let $p_1(y_{k,\mathcal{S}}|\boldsymbol{\theta}) = f_1(y_{k,\mathcal{S}}|\boldsymbol{\theta})$, where f_1 is given in Equation (5.10) or (5.12) depending on which decision rule is used. Then, for $k = 2, \dots, K$, let

$$p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta}) = \int_{\mathcal{C}_{k-1}} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta}) f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta}) dy_{k-1,\mathcal{S}},$$

where $f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta})$ is given in Equation (5.9). Then,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_{\mathcal{P}^*}|\mathcal{P}^* = \mathcal{S}\right] &= \sum_{k=1}^{K} \int_{y_{k,\mathcal{S}} \notin \mathcal{C}_{k,\mathcal{S}}} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} \\ &= \sum_{k=1}^{K} \left\{ \int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} + \int_{-\infty}^{\tilde{l}_{k,\mathcal{S}}} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} \right\}.\end{aligned} \tag{5.17}$$

It is straightforward to show that

$$\begin{aligned}
&\int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} \\
= \; & \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} \int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta}) f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} \; dy_{k,\mathcal{S}} \; dy_{k-1,\mathcal{S}} \\
= \; & \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta}) U_{\mathcal{S}}^{(1)} \left(\tilde{u}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right) dy_{k-1,\mathcal{S}},
\end{aligned}$$

where

$$\begin{aligned}
U_{\mathcal{S}}^{(1)} \left(\tilde{u}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right) &:= \int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} \\
= \; & \frac{1}{\mathcal{I}_{k,\mathcal{S}}} \Bigg\{ \varphi\left(\frac{\tilde{u}_{k,\mathcal{S}} - (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})}{\sqrt{\Delta_{k,\mathcal{S}}}} \right) \sqrt{\Delta_{k,\mathcal{S}}} \\
& + \Phi\left(\frac{y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}} - \tilde{u}_{k,\mathcal{S}}}{\sqrt{\Delta_{k,\mathcal{S}}}} \right) (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}}) \Bigg\}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\int_{-\infty}^{\tilde{l}_{k,\mathcal{S}}} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} \\
= \; & \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta}) L_{\mathcal{S}}^{(1)} \left(\tilde{l}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right) dy_{k-1,\mathcal{S}},
\end{aligned}$$

where

$$\begin{aligned}
L_{\mathcal{S}}^{(1)} \left(\tilde{l}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right) &:= \int_{\infty}^{\tilde{l}_{k,\mathcal{S}}} f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta}) \frac{y_{k,\mathcal{S}}}{\mathcal{I}_{k,\mathcal{S}}} dy_{k,\mathcal{S}} \\
= \; & \frac{1}{\mathcal{I}_{k,\mathcal{S}}} \Bigg\{ -\varphi\left(\frac{\tilde{l}_{k,\mathcal{S}} - (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})}{\sqrt{\Delta_{k,\mathcal{S}}}} \right) \sqrt{\Delta_{k,\mathcal{S}}} \\
& + \Phi\left(\frac{\tilde{l}_{k,\mathcal{S}} - (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})}{\sqrt{\Delta_{k,\mathcal{S}}}} \right) (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}}) \Bigg\}.
\end{aligned}$$

We can hence evaluate the conditional bias of $\hat{\theta}_{\mathcal{P}^*}$ given $\mathcal{P}^* = \mathcal{S}$, using a given value of $\theta_{\mathcal{S}}$. Computations are straightforward, involving only a succession of univariate integrals (as was the case when computing stopping boundaries).

It may also be of interest to evaluate the conditional variance of our estimates, i.e. $\mathrm{Var}\Big(\hat{\theta}_{\mathcal{P}^*}|\mathcal{P}^* = \mathcal{S}\Big)$. For this, we need to compute

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}^2_{\mathcal{P}^*}|\mathcal{P}^* = \mathcal{S}\right] = \sum_{k=1}^{K}\left\{\int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta})\frac{y^2_{k,\mathcal{S}}}{\mathcal{I}^2_{k,\mathcal{S}}}dy_{k,\mathcal{S}} + \int_{-\infty}^{\tilde{l}_{k,\mathcal{S}}} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta})\frac{y^2_{k,\mathcal{S}}}{\mathcal{I}^2_{k,\mathcal{S}}}dy_{k,\mathcal{S}}\right\}. \tag{5.18}$$

In similar fashion as above, we can show that

$$\begin{aligned}
&\int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta})\frac{y^2_{k,\mathcal{S}}}{\mathcal{I}^2_{k,\mathcal{S}}}dy_{k,\mathcal{S}} \\
=\;& \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}}\int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta})f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta})\frac{y^2_{k,\mathcal{S}}}{\mathcal{I}^2_{k,\mathcal{S}}}dy_{k,\mathcal{S}}dy_{k-1,\mathcal{S}} \\
=\;& \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta})U^{(2)}_{\mathcal{S}}\left(\tilde{u}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right)dy_{k-1,\mathcal{S}},
\end{aligned}$$

where

$$\begin{aligned}
U^{(2)}_{\mathcal{S}}\left(\tilde{u}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right) &:= \int_{\tilde{u}_{k,\mathcal{S}}}^{\infty} f_k(y_{k,\mathcal{S}}|y_{k-1,\mathcal{S}};\boldsymbol{\theta})\frac{y^2_{k,\mathcal{S}}}{\mathcal{I}^2_{k,\mathcal{S}}}dy_{k,\mathcal{S}} \\
&= \frac{1}{\mathcal{I}^2_{k,\mathcal{S}}}\Big\{\big[\Delta_{k,\mathcal{S}}z_u + 2(y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})\big]\varphi(z_u) \\
&\quad + \big[\Delta_{k,\mathcal{S}} + (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})^2\big]\Phi(-z_u)\Big\},
\end{aligned}$$

and $z_u = \big[\tilde{u}_{k,\mathcal{S}} - (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})\big]/\sqrt{\Delta_{k,\mathcal{S}}}$. Similarly,

$$\begin{aligned}
&\int_{-\infty}^{\tilde{l}_{k,\mathcal{S}}} p_k(y_{k,\mathcal{S}}|\boldsymbol{\theta})\frac{y^2_{k,\mathcal{S}}}{\mathcal{I}^2_{k,\mathcal{S}}}dy_{k,\mathcal{S}} \\
=\;& \int_{\tilde{l}_{k-1,\mathcal{S}}}^{\tilde{u}_{k-1,\mathcal{S}}} p_{k-1}(y_{k-1,\mathcal{S}}|\boldsymbol{\theta})L^{(2)}_{\mathcal{S}}\left(\tilde{l}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right)dy_{k-1,\mathcal{S}},
\end{aligned}$$

where

$$\begin{aligned}
L^{(2)}_{\mathcal{S}}\left(\tilde{u}_{k,\mathcal{S}}, y_{k-1,\mathcal{S}}, \Delta_{k,\mathcal{S}}, \mathcal{I}_{k,\mathcal{S}};\boldsymbol{\theta}\right) &:= \frac{1}{\mathcal{I}^2_{k,\mathcal{S}}}\Big\{-\big[\Delta_{k,\mathcal{S}}z_l + 2(y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})\big]\varphi(z_l) \\
&\quad + \big[\Delta_{k,\mathcal{S}} + (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})^2\big]\Phi(z_l)\Big\},
\end{aligned}$$

and $z_l = \big[\tilde{l}_{k,\mathcal{S}} - (y_{k-1,\mathcal{S}} + \theta_{\mathcal{S}}\Delta_{k,\mathcal{S}})\big]/\sqrt{\Delta_{k,\mathcal{S}}}$.

We can get unconditional bias $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_{\mathcal{S}} - \theta_{\mathcal{S}}|\mathcal{S}\right]\right]$ using Equation (5.17), and by summing over stage one decision probabilities. In the case that $\mathcal{P}^* = \{j\}$, $j \in \mathcal{P}$, we can compute the conditional bias of $\hat{\theta}_j$ using the above formulae. It might also be of interest to obtain the unconditional bias of $\hat{\theta}_j$. In that case, we could evaluate the quantities $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_j|j \notin \mathcal{P}^*\right]$ and $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_j|j \in \mathcal{P}^*\right]$. The first of these is straightforward: evaluate $\mathbb{E}_{\boldsymbol{\theta}}\left[Y_{1j}/\mathcal{I}_{1j}|Y_{1j} < l_1\sqrt{\mathcal{I}_{1j}}\right]$ using the inverse Mills ratio. In the analysis of Section 5.4 we do not keep track of individual populations past the first stage, so $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_j|j \in \mathcal{P}^*\right]$ might be easier to obtain with simulation. More work is required to express the unconditional expectation $\mathbb{E}_{\boldsymbol{\theta}}(\hat{\theta}_j)$ analytically.

It is known that group sequential designs yield biased estimators. Wang and Leung (1997) proposed a bootstrap algorithm for bias-reduction in a single-population group sequential design, and we employ a simple extension of this algorithm that accounts for multiple populations. At the termination of a clinical trial, we have maximum likelihood estimates $\hat{\theta}_j$ for all populations Ω_j. For populations eliminated at the first interim analysis, this is simply $\hat{\theta}_{1j}$, and for remaining populations use $\hat{\theta}_{T,j}$. The bootstrap algorithm is now outlined, taking input $\hat{\boldsymbol{\theta}}$.

1. Generate B bootstrap samples via simulation, using $\hat{\boldsymbol{\theta}}$ as the treatment effect parameter. That is, simulate B clinical trials using original information levels and stopping boundaries. Each run will yield a bootstrap estimate $\hat{\theta}_j^{*b}$ for $j \in \mathcal{P}$ and $b = 1, \ldots, B$. The mean bootstrap MLE for θ_j is then defined as
$$\bar{\theta}_{1,j}^{*B} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_j^{*b}.$$
2. Set $\hat{\theta}_{1,j}^* = \hat{\theta}_j - \left(\bar{\theta}_{1,j}^{*B} - \hat{\theta}_j\right)$, where $\left(\bar{\theta}_{1,j}^{*B} - \hat{\theta}_j\right)$ is the simulated bootstrap bias estimate.

As suggested by Wang and Leung (1997), the algorithm can be repeated to reduce higher order bias, using the most recent bias-adjusted estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}^*$, as the input. In general, suppose we have applied the adjustment algorithm $r-1$ times, and currently have adjusted estimates $\hat{\theta}^*_{r-1,j}$ for $j \in \mathcal{P}^*$. Now, redo step 1 above to get $\bar{\theta}^{*B}_{r,j}$. Then our new bias-adjusted estimate of θ_j is given as

$$\hat{\theta}^*_{r,j} = \hat{\theta}_j - \left(\bar{\theta}^{*B}_{r,j} - \hat{\theta}^*_{r-1,j}\right). \tag{5.19}$$

This re-application of the bootstrap might be considered as a simple version of the "double bootstrap" introduced by Beran (1988). In subsequent numerical examples, we also construct bootstrap confidence intervals. A $1-2\alpha$ interval is obtained in a non-parametric fashion by evaluating α and $1-\alpha$ percentiles of the B bootstrap simulations that are generated via the algorithm discussed above.

5.8 Worked Examples

In this section, the GSDS procedure is illustrated by showing how it could be employed in two types of clinical trials where the target populations may exhibit heterogeneity with respect to treatment response.

5.8.1 Clinical Trial for Depression Treatment

We consider the example of testing a new treatment intended to treat patients suffering from depression disorder. The efficacy of the new treatment is compared to that of an active control, and the primary endpoint is change in the Hamilton Depression Index (HAMD) after 6 weeks of therapy. Let μ^E and μ^C be the respective mean 6-week declines (negative change from baseline, so $\mu > 0$ implies improvement) in HAMD for the experimental treatment and control, and let $\theta = \mu^E - \mu^C$

be the HAMD improvement resulting from the new treatment. As has been seen with past experience in similar trials (Mehta and Patel, 2006), we assume that the observation standard deviation σ is known and equal to 10. Assuming that patient responses are normally distributed, the efficient score and observed information are given as

$$Y = \mathcal{I} \cdot \left(\bar{X}^E - \bar{X}^C \right), \quad \mathcal{I} = \frac{n}{4\sigma^2},$$

where $\bar{X}^E$ and $\bar{X}^C$ are the respective sample mean treatment responses from experimental and control treatments, and n is the total sample size. Note that the numbers of patients assigned to the new and control treatments are equal to $\frac{n}{2}$.

The HAMD takes integer values 0–52, where scores from 8–13 indicate mild depression, 14–18 moderate depression, 19–22 severe depression and 23 or greater very severe depression. Meta-analyses conducted by Kirsch et al. (2008) and Fournier et al. (2010) have found evidence suggesting that benefits for treatment versus control are mostly confined to patients suffering from the most severe form of depression. This issue has also received widespread mainstream media attention, see for example Begley (2010). In our example, we therefore assume that there are two populations of interest. Namely, let Ω_1 consist of patients whose initial HAMD score is $\geq$ 23, and let Ω_2 be the remaining patients suffering from less severe depression (i.e. HAMD $\leq$ 22). Fournier et al. (2010) report that roughly 70% of treatment-seeking outpatients have HAMD scores $\leq$ 22, so we assume that $f_{01} = 0.3 = 1 - f_{02}$. We let θ_1 and θ_2 denote the respective mean HAMD improvements of the new treatment over control for Ω_1 and Ω_2. Then, $\theta_0 = f_{01}\theta_1 + f_{02}\theta_2$ denotes the mean HAMD improvement for the general population, $\Omega_0 = \Omega_1 \cup \Omega_2$.

As it is believed that patients with severe depression are likely more responsive, we consider the population structure to be a nested one, and so DR-II is applicable (see Section 5.2). A three stage trial is planned, where each of the three stages uses an equal amount (33%) of the planned sample size.[2] If the trial continues but Ω_2 is eliminated, then all planned observations are allocated to Ω_1 for stages two and three. Hence there are three possible conclusions to this trial:

- Fail to reject any null hypothesis, hence concluding that the new treatment does not constitute an improvement, i.e. overall futility.
- Reject the null hypothesis $H_1 : \theta_1 = 0$, while accepting $H_2 : \theta_2 = 0$. The treatment only improves on the control for the most severely depressed patients (HAMD ≥ 23).
- Reject the null hypothesis $H_0 : \theta_0 = 0$, where $\theta_0 = f_{01}\theta_1 + f_{02}\theta_2$. The treatment constitutes an improvement for all patients.

Spending functions are defined as

$$\alpha_U^*(0.33) = 0.0083;\ \alpha_U^*(0.67) = 0.0167;\ \alpha_U^*(1) = 0.025,$$

and

$$\alpha_L^*(0.33) = 0.3250;\ \alpha_L^*(0.67) = 0.6500;\ \alpha_L^*(1) = 0.975.$$

Recursively solving Equations (5.4) and (5.5) as described in Section 5.4 results in the following standardized boundaries:

$$(l_1, u_1) = (0.1766, 2.5551);\ (l_2, u_2) = (0.4580, 2.4649);\ (l_3, u_3) = (2.3365, 2.3365).$$

Following the work of Mehta and Patel (2006), we consider the value $\theta^* = 4$ to represent a clinically meaningful improvement over the active control. We set $\alpha = 0.025$, and determine the required sample size so as to guarantee a positive

[2] It is not necessary to make all stages be of equal length, as the procedure outlined is able to handle unequal increments in information (as long as planned information accumulation is explicitly stated at the outset).

result (reject H_0 or H_1) with probability $1 - \beta = 0.9$, when $\theta_1 = 4$ and $\theta_2 = 2$ (i.e. treatment effect in Ω_2 is positive but not practically significant). Solving for $\mathcal{I}_{\max}$ in Equation (5.13) of Section 5.5, this power requirement results in $\mathcal{I}_{\max} = 1.6494$ which, given the fact that $\sigma = 10$, is approximately equivalent to a sample size of 660 patients. For comparison, requiring that the power specification be satisfied when $(\theta_1, \theta_2) = (4, 0)$ results in a sample size of 1,665, while $(\theta_1, \theta_2) = (4, 4)$ requires only 313 patients. With a total sample size of 660, each stage will sample 220 patients, with 66 coming from Ω_1 and 154 from Ω_2.

After the first stage, observed information for Ω_1 and Ω_2 is $\mathcal{I}_{11} = 0.1649$ and $\mathcal{I}_{12} = 0.3849$, respectively. Hence the population-specific lower boundaries are $l_{11} = l_1\sqrt{\mathcal{I}_{11}} = 0.0717$ and $l_{12} = l_1\sqrt{\mathcal{I}_{12}} = 0.1096$. Suppose that we observe efficient scores $Y_{11} = 0.1803$ and $Y_{12} = -0.0119$ (giving point estimates $\hat{\theta}_{11} = 1.09$ and $\hat{\theta}_{12} = -0.03$). Then, as $Y_{12} < l_{12}$, patients from Ω_2 will no longer be recruited for the trial ($H_2 : \theta_2 = 0$ is accepted at this point). The upper boundary for Ω_1 is $u_{11} = u_1\sqrt{I_{11}} = 1.0377$, so $Y_{11} \in (l_{11}, u_{11})$ and sampling continues with all patients coming from Ω_1 (i.e. we enrich the subpopulation Ω_1). The Ω_1 increment in information for stage two will now be $\Delta_{21} + \Delta_{22} = 0.5498$ (220 patients). Stage two boundaries are $\tilde{l}_{2,\{1\}} = l_2\sqrt{\mathcal{I}_{2,\{1\}}} = 0.3872$ and $\tilde{u}_{2,\{1\}} = u_2\sqrt{\mathcal{I}_{2,\{1\}}} = 2.0839$. Suppose $Y_{2,\{1\}} = 2.3408$. Then $Y_{2,\{1\}} > \tilde{u}_{2,\{1\}}$ and the trial stops with rejection of $H_1 : \theta_1 = 0$. Stage three is not conducted, and the trial reached a positive conclusion based on observed information equal to $\mathcal{I}_{2,\{1\}} = 0.7147$ which corresponds to approximately 286 patients (from Ω_1). The 220 observations planned for stage three are not needed.

The resulting point estimates, $\hat{\theta}_1 = Y_{2,\{1\}}/\mathcal{I}_{2,\{1\}} = 3.2752$ and $\hat{\theta}_2 = Y_{12}/\mathcal{I}_{12} = -0.0309$, can be adjusted according to the bias-correction procedure outlined in

Section 5.7. Applying the bootstrap algorithm, we obtain bias adjusted estimates $\hat{\theta}_1^{*\text{new}} = 2.8199$ and $\hat{\theta}_2^{*\text{new}} = 0.5736$. The respective standard errors for these adjustments (using B =10,000 bootstrap replications) are 1.59 and 1.16. We can also obtain $1 - 2\alpha$ bootstrap percentile intervals in a non-parametric fashion by using the α and $1 - \alpha$ bootstrap sample percentiles. Doing so for $\alpha = 0.025$, we obtain the intervals

$$\text{For } \theta_1 : (1.64, 7.08); \quad \text{For } \theta_2 : (-2.73, 1.11).$$

Bootstrapped bias estimates can exhibit a fair amount of variability, so the standard errors and intervals we observe here are not surprising. We do note that $\theta^* = 4$ is excluded from the interval for θ_2, but is included in the interval for θ_1. Also of note is that 0 is included in the interval for θ_2 but excluded from the interval for θ_1. Hence the intervals are concordant with the trial conclusions.

We end this example by investigating what the trial conclusion would have been, had a flexible adaptive seamless Phase II/III design been employed, where arbitrary trial modifications are allowed at interim analysis points (Bretz et al., 2006). We still assume that $\alpha = 0.025$, and that Type I error spending is equally allocated over the three stages, so $\alpha^* = 0.0083$ will be spent at each stage. Stage-wise p-values are computed as

$$p_{1j} = 1 - \Phi\left[Y_{1j}/\sqrt{\mathcal{I}_{1j}}\right], \text{ and } p_{kj} = 1 - \Phi\left[\frac{Y_{kj} - Y_{k-1,j}}{\sqrt{\mathcal{I}_{kj} - \mathcal{I}_{k-1,j}}}\right],$$

for $k = 2, 3$ and $j = 0, 1, 2$. For stage k, the p-values are combined using the weighted inverse normal combination method:

$$C(p_1, \ldots, p_k) = \left(\sum_{s=1}^{k} w_s^2\right)^{-1/2} \sum_{i=1}^{k} w_i \Phi^{-1}(1 - p_i),$$

and rejection occurs if $1 - \Phi[C(p_1, \ldots, p_k)] < \alpha^*$. As each stage uses an equal amount of the planned sample size, combination weights w_k, for $k = 1, 2, 3$, are all equal to $\sqrt{1/3}$.

The hypotheses of interest are $H_0 : \theta_0 = 0$ and $H_1 : \theta_1 = 0$, so we must account for multiple testing by computing multiplicity-adjusted p-values.[3] We do this according to Simes' procedure (Simes, 1986), whereby the k^{th} stage p-value for the intersection hypothesis $H_{01} = H_0 \cap H_1$ is

$$p^*_{k0} = \min\left\{2\min\{p_{k1}, p_{k0}\}, \max\{p_{k1}, p_{k0}\}\right\}, \; k = 1, 2, 3.$$

Now suppose that at the end of stage one, a decision is (arbitrarily) made to discontinue sampling from Ω_2, and to continue only with patients from Ω_1. Using the numbers given above, stage one results in p-values $p_{11} = 0.3285$, $p_{12} = 0.5077$ so $p_{10} = 0.4102 = p^*_{10}$. As Ω_2 is not sampled during stage two, H_{01} must be tested by computing $C(p^*_{10}, p_{21})$. Using the observed value of Y_{21}, we get that $p_{21} = 0.0018$, and hence

$$C(p_{11}, p_{21}) = 2.3743, \text{ and } C(p^*_{10}, p_{21}) = 2.2209.$$

Noting that $1 - \Phi[2.3743] = 0.0088 > \alpha^* = 0.0083$, we see that H_1 cannot be rejected at this stage (neither can H_{01}). Note that $C(p_{11}, p_{21})$ and $C(p^*_{10}, p_{21})$ were computed using the pre-specified weights, so $w_k = \sqrt{1/3}$ for all k, which are inefficient as the population Ω_1 was enriched for the second stage.

Using empirical data weights, i.e. using weights that are adjusted to reflect observed increments in information (accounting for enrichment), we get

$$\widetilde{C}(p_{11}, p_{21}) = 2.7689, \text{ and } \widetilde{C}(p^*_{10}, p_{21}) = 2.6647.$$

which leads to rejection of both H_{01} and H_1. However, as noted in (earlier section in thesis), data-dependent adjustment of weights is known to inflate Type I error as tests are not necessarily based on standard normal statistics.

[3] As the trial allows for any modification, we could feasibly be interested in testing $H_2 : \theta_2 = 0$ as well. However, to follow the population structure, we assume that restricting to Ω_2 was not part of the trial protocol.

5.8.2 I-SPY 2 – Neo-Adjuvant Treatment of Breast Cancer

The I-SPY 2 trial ("Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis 2") is an ongoing highly adaptive Phase II clinical trial, intended to foster efficient clinical development of oncologic therapies and biomarkers (Barker et al., 2009). The trial tests numerous promising agents in the neo-adjuvant[4] setting for women with breast cancer ($>$ 3.0 cm) using a Bayesian design for adaptive randomization. Treatments that show a high Bayesian predictive probability of being more effective than standard therapy will "graduate" from the Phase II trial, and pass on to a concentrated Phase III trial. The overall trial design features two arms of standard neo-adjuvant chemotherapy treatment. Patients are administered "weekly paclitaxel (plus trastuzumab (Herceptin) for HER2 positive patients) followed by doxorubicin (Adriamycin) and cyclophosphamide (Cytoxan)" (Barker et al., 2009). Other arms will be assigned to five new drugs, to be tested simultaneously in addition to the standard therapy. The primary endpoint is the measurement of pathologic complete response[5] (pCR), and secondary endpoints include up to 10 years monitoring for overall and disease-free survival.

Part of the unique approach of I-SPY 2 is the inclusion of biomarker identification for each candidate drug. Hence each drug that is successfully passed to a Phase III trial comes with a number of recommended subgroups, as obtained by use of a hierarchical model that enables borrowing of information across all possible subgroups. Though research over the last decade has generally established

[4] In oncology, neo-adjuvant therapy refers to the administration of chemotherapy prior to surgery.

[5] Pathologic complete response refers to the disappearance of all clinical evidence of disease. This does not necessarily mean cure, as microscopic metastases may remain undetected, and can regrow and become resistant to treatment.

that cancer is a number of heterogeneous diseases, the development and validation of biomarkers that can accurately classify patients according to prognosis and treatment has proven to be immensely challenging. The hierarchical model used in I-SPY 2, in combination with the ongoing testing, analytic validation and qualification of biomarkers, is a revolutionary approach to breaking the "biomarker barrier."

Biomarkers consist of three classes: standard, qualifying and exploratory. Standard biomarkers are approved by the FDA and are used to determine patient eligibility for the trial, while qualifying and exploratory biomarkers are not yet approved, and some are defined on the basis of promising preliminary data in I-SPY 2. Tissue and blood samples are collected prospectively in I-SPY 2, which contributes to the validation of exploratory biomarkers both during the trial, and with retrospective analyses following the trial. The standard biomarkers are hormone receptor (HR) status (+/−), human epidermal growth factor receptor 2 (HER2, discussed in Chapter 1) status (+/−), and MammaPrint status (high risk MP2, low risk MP1). Hormone receptor status is determined by assessing estrogen receptor status (ER) and progesterone receptor status (PR). If ER and PR are both negative, then HR is said to be negative as well, and if either of ER or PR is positive (or both), then HR is positive. MammaPrint (Mook et al., 2007) is a molecular diagnostic test that assesses the risk for recurrence of a cancer. To be eligible for enrollment in the trial, a patient's cancer must be one of the following:

1. MammaPrint High Risk score;
2. MammaPrint Low Risk score and ER negative;
3. MammaPrint Low Risk score and ER positive and HER2 positive.

Patients with low MammaPrint score, as well as ER positive and HER2 negative, are excluded from the trial as they would not be considered ideal candidates for chemotherapy. Combinations and unions of combinations of the subgroups defined by the standard biomarkers are narrowed down from 256 possibilities to 14 distinct signatures (Barker et al., 2009). This is done based on clinical relevance, as most signatures are rare and biologically uninteresting.

I-SPY 2 involves exploratory testing of numerous novel drugs in comparison with the efficacy of standard chemotherapy alone. Each drug is added to standard therapy, and is tested on a minimum of 20 patients, and a maximum of 120 patients. Based on biomarker signature and eligibility, patients are randomized to the novel drugs, and Bayesian methods of adaptive randomization ensure that randomization probabilities are relatively high for drugs that do well with a particular biomarker signature. Namely, if $\pi(z,t)$ is the probability of pCR for a patient with biomarker signature z and for treatment t, then allocation probabilities are proportional to

$$\mathbb{P}\left[\pi(z,t) > \pi(z,t'),\ t' \neq t \mid \text{data}\right],$$

the posterior probability that treatment t is most likely to benefit this particular patient (Berry et al., 2011). Not only does this allow promising drugs to progress through the trial more rapidly and efficiently, but adaptively randomizing in this fashion also addresses ethical concerns as patients are more likely to be assigned to drugs genuinely believed to be helpful.

As the I-SPY 2 trial progresses, a Bayesian posterior predictive probability of success in a future Phase III trial is computed for each drug, and for each possible biomarker signature. The future Phase III trial is intended to compare the drug in question against standard therapy with a prespecified sample size. Hence, drugs that have a high Bayesian posterior predictive probability of being more effective than standard therapy, are “graduated” along with their corresponding subgroups and passed on to confirmatory testing in a separate Phase III trial. When the posterior predictive probability of success becomes sufficiently low for all signatures, the drug is dropped from consideration. New drugs can be added at any time during the I-SPY 2 trial as other drugs are either dropped or graduated.

For the purposes of demonstrating an application of the design developed in this chapter, we carry out an illustrative example of how a Phase III trial might be conducted for a novel drug that has graduated from the I-SPY 2 trial. Assumptions and conclusions are not necessarily realistic in a clinical sense; rather they are presented here for the sake of demonstrating our procedure. Suppose then that a promising agent has graduated from a Phase II study as described above, and passed on to confirmatory testing in a Phase III trial. We assume that there were three biomarker signatures positively associated with the new drug, and for ease of exposition these are taken to be the populations defined by the inclusion criteria for I-SPY 2. Hence we have three disjoint patient populations:

- Ω_1 = Patients with MammaPrint High Risk score;
- Ω_2 = Patients with MammaPrint Low Risk score and ER negative;
- Ω_3 = Patients with MammaPrint Low Risk score and ER positive and HER2 positive.

As the three biomarkers were passed on with the drug, but no obvious preference is present, we can apply DR-I of Section 5.2. The best case conclusion for this procedure is that all three populations respond well enough during the first stage, and the trial is consequently concluded having determined a positive effect for all patients involved. In the case that one (or more) of the three populations appears to be nonresponsive, we can drop said population(s) and proceed to later stages while pooling the remaining populations. Hence there is some flexibility to salvage the trial should this become necessary. This is desirable, as the Phase II testing conducted in I-SPY 2 did not necessarily involve very large sample sizes.

To proceed, we need to specify patient stratification among the signatures passed to this Phase III trial. For efficient screening, we will stratify patients according to the estimated prevalence of each of the three populations defined above. Barker et al. (2009) provide a supplemental document containing expected prevalence (obtained in an earlier program, I-SPY 1) for patient populations defined by HR (+/−), HER2 (+/−) and MammaPrint (MP1/MP2) status, from which we obtain signature prevalence (estimates are somewhat crudely rounded for simplicity) as $f_{01} = 0.6$, $f_{02} = 0.18$ and $f_{03} = 0.22$. We note that patient inclusion criteria are based on ER (and not PR), while prevalence information is only available for HR (depends both on ER and PR). However, as this example is only for illustrative purposes, we proceed under the simplifying assumption that ER prevalence is the same as that of HR.

Let p^E and p^C denote the respective pCR probabilities for experimental and control treatments, and let $\theta = p^E - p^C$. Alternatively, we could define θ as the log-odds ratio

$$\theta = \log\left\{\frac{p^E\left(1-p^C\right)}{p^C\left(1-p^E\right)}\right\} = \log\left\{\frac{p^E}{1-p^E}\right\} - \log\left\{\frac{p^C}{1-p^C}\right\}.$$

Let θ_j denote the mean improvement in pCR probability for Ω_j, so $\theta_0 = f_{01}\theta_1 + f_{02}\theta_2 + f_{03}\theta_3$ is the mean improvement for the general population. Using the parametrization of θ defined above, and assuming that the numbers of patients assigned to the new and control treatments are equal to $\frac{n}{2}$ (recall that n is the total sample size), the efficient score and observed information are given as

$$Y = \hat{\mathcal{I}} \cdot \left(\hat{p}^E - \hat{p}^C\right), \quad \hat{\mathcal{I}} = \frac{n}{2\tilde{p}(1-\tilde{p})}. \tag{5.20}$$

Here, $\hat{p}^E$ and $\hat{p}^C$ are the respective estimates of success probabilities for experimental and control treatments. Furthermore, $\bar{p} = \left(p^E + p^C\right)/2$ denotes the common response probability, estimated by $\tilde{p}$ under the null hypothesis by pooling observations from both treatment arms. We note that information depends on the unknown $\bar{p}$, so when designing the trial we shall use the conservative value of $\bar{p} = \frac{1}{2}$. Hence, observed information will be larger than expected, and actual power will be greater than desired.

A two-stage trial is envisaged, and planned sample size is divided evenly over the two stages. If the trial continues past the first stage, but one or more populations are eliminated, then all planned observations are allocated to the remaining populations for the second stage.[6] Given the population structure, there are eight possible conclusions to the trial (futility, or reject $H_{\mathcal{S}}$ for some $\mathcal{S} \subseteq \mathcal{P} = \{1, 2, 3\}$). Spending functions are defined as

$$\alpha_U^*(0.5) = 0.0125;\ \alpha_U^*(1) = 0.025, \text{ and } \alpha_L^*(0.5) = 0.4875;\ \alpha_L^*(1) = 0.975.$$

Recursively solving Equations (5.4) and (5.5) as described in Section 5.4 results in the following standardized boundaries:

$$(l_1, u_1) = (0.7962, 2.7625);\ (l_2, u_2) = (2.5204, 2.5204).$$

Pooled pCR response for control (standard chemotherapy) from I-Spy 1 was roughly 40%, and we arbitrarily consider $\theta^* = 0.2$ to be a clinically significant effect. We set $\alpha = 0.025$ and determine the required sample size to guarantee that

[6] Depending on subgroup prevalence and type of disease, using enrichment in this setting may not be feasible, particularly if the screening process is expensive and/or time-consuming.

Table 5.1: Results of the Phase III trial following I-SPY 2. Displayed are efficient scores, observed Fisher's information levels, sample sizes, Z-scores and p-values.

Stage 1:	X_{1j}	Δ_{1j}	n_{1j}	Z_{1j}	P_{1j}
Ω_1	-10.71	480	240	-0.49	0.6875
Ω_2	12.84	144	72	1.07	0.1423
Ω_3	19.06	176	88	1.44	0.0754
Stage 2:	X_{2j}	Δ_{2j}	n_{2j}	Z_{2j}	P_{2j}
Ω_2	34.07	360	180	1.80	0.0363
Ω_3	69.60	440	220	3.32	0.0005
$\Omega_{\{2,3\}}$	103.67	800	400	3.67	0.0001
Total:	Y_{2j}	$\mathcal{I}_{2j}$	n_j	Z_j	P_j
Ω_2	46.91	504	252	2.09	0.0183
Ω_3	88.66	616	308	3.57	0.0002
$\Omega_{\{2,3\}}$	135.57	1120	560	4.05	0.0000

H_0 is rejected (use a numerical search for $\mathcal{I}_{\max}$ in Equation (5.14) of Section 5.5) with probability $1 - \beta = 0.9$ when $\theta^* = 0.2$ for all populations. This results in $\mathcal{I}_{\max} = 1495.5$, which from Equation (5.20) is equivalent to approximately 748 patients, which we round up to 800 to allow for drop-outs. Hence, each stage recruits 240 patients from Ω_1, 72 patients from Ω_2, and 88 patients from Ω_3 (assuming all three populations are carried on to the second stage).

After the first stage, observed information for Ω_1, Ω_2 and Ω_3 will be $\mathcal{I}_{11} = 480, \mathcal{I}_{12} = 144$ and $\mathcal{I}_{13} = 176$, respectively. Hence, population-specific lower boundaries are $l_{11} = l_1\sqrt{\mathcal{I}_{11}} = 17.44, l_{12} = l_1\sqrt{\mathcal{I}_{12}} = 9.55$ and $l_{13} = l_1\sqrt{\mathcal{I}_{13}} = 10.56$. Suppose we observe the efficient scores $Y_{11} = -10.71, Y_{12} = 12.84$, and $Y_{13} = 19.06$ (giving point estimates $\hat{\theta}_{11} = -0.0223, \hat{\theta}_{12} = 0.0892$ and $\hat{\theta}_{13} = 0.1083$). We see that $Y_{11} < l_{11}$, so patients from population Ω_1 will not be recruited for stage two. Furthermore, $Y_{12} + Y_{13} = 31.90 < u_{1,\{2,3\}} = u_1\sqrt{\Delta_{12} + \Delta_{13}} = 49.42$, so the trial continues to stage two. Increments in information for stage two will be $\tilde{\Delta}_{22} = 360$ and $\tilde{\Delta}_{23} = 440$ (180 and 220 patients) for Ω_2 and Ω_3 respectively. As

Table 5.2: Effect estimates, first and second bootstrap adjustments, and the resulting percentile confidence intervals.

Estimates	Raw $\hat{\theta}_j$	$\hat{\theta}^{*\text{new}}_{1,j}$	$\hat{\theta}^{*\text{new}}_{2,j}$	CI-Low	CI-High
Ω_1	-0.0223	-0.0302	-0.0410	-0.0778	0.0049
Ω_2	0.0931	0.0527	0.0479	0.0134	0.2157
Ω_3	0.1439	0.1179	0.1169	0.0665	0.2507
$\Omega_{\{2,3\}}$	0.1210	0.0886	0.0858	0.0590	0.2064

the two populations are now pooled together, the critical values after stage two will be $\tilde{l}_{2,\{2,3\}} = \tilde{u}_{2,\{2,3\}} = u_2\sqrt{\Delta_{12} + \Delta_{13} + \tilde{\Delta}_{22} + \tilde{\Delta}_{23}} = 84.35$. Suppose respective second stage efficient scores are $Y_{22} = 46.91$ and $Y_{23} = 88.66$, yielding the combined score $Y_{2,\{2,3\}} = 135.79$. Then, as $Y_{2,\{2,3\}} > \tilde{u}_{2,\{2,3\}}$, the null hypothesis $H_{\{2,3\}} : \theta_{\{2,3\}} = 0$ is rejected and the treatment is approved for patients in the population $\Omega_2 \cup \Omega_3$.

The resulting point estimates are $\hat{\theta}_2 = 0.0931$ and $\hat{\theta}_3 = 0.1439$, and $\hat{\theta}_{\{2,3\}} = 0.1210$. After twice applying the bootstrap algorithm of Section 5.7 to remove first and second order bias, we get adjusted estimates $\hat{\theta}^{*\text{new}}_2 = 0.0479$, $\hat{\theta}^{*\text{new}}_3 = 0.1169$ and $\hat{\theta}^{*\text{new}}_{\{2,3\}} = 0.0858$. The bias-adjusted estimate of θ_1 is $\hat{\theta}^{*\text{new}}_1 = -0.0410$. The $1 - 2\alpha$ bootstrap percentile intervals for $\theta_2, \theta_3, \theta_{\{2,3\}}$ and θ_1 are:

$$\begin{aligned}
\theta_2 &: (0.0134, 0.2157);\\
\theta_3 &: (0.0665, 0.2507);\\
\theta_{\{2,3\}} &: (0.0590, 0.2064);\\
\theta_1 &: (-0.0778, 0.0049).
\end{aligned}$$

We note that $\theta^* = 0.2$ is included in the first three intervals, but excluded from the interval for θ_1. Furthermore, the interval for θ_1 is the only one that contains the value zero. Table 5.2 contains the raw effect estimates, as well as bootstrap adjustments and confidence intervals.

Table 5.3: Tests of intersection hypotheses using the weighted inverse normal combination rule and Simes' p-value adjustment method. Weights are prespecified and set equal to $\sqrt{1/2}$.

Hypothesis	P-values		Combined	
	Stage 1	Stage 2	Z	P
H_1	0.6875	–	–	–
H_2	0.1423	0.0363	2.026	0.0214
H_3	0.0754	0.0005	3.362	0.0004
H_{12}	0.2846	0.0363	1.672	0.0472
H_{13}	0.1508	0.0005	3.077	0.0010
H_{23}	0.1423	0.0009	2.962	0.0015
H_{123}	0.2135	0.0009	2.767	0.0028

As in Section 5.8.1, we consider an alternate approach to the Phase III trial. Using two stages and equal spending $\alpha^* = 0.0125$ for each stage, we assume that a more adaptive approach was taken, whereby arbitrary modifications were allowed at the first analysis (Bretz et al., 2006). P-values are combined using the weighted inverse normal combination rule, and Simes' method is used to adjust p-values for multiple testing. As equal sample size is used for both stages, combination weights are preset to equal $\sqrt{1/2}$. Statistics are observed exactly as given in Table 5.1. Table 5.3 shows p-values for each stage, as well as Z-scores and p-values combined over both stages using Simes' method.

We assume that Ω_1 was dropped after stage one, so after stage two it is only possible to reject H_2 and/or H_3. The adjusted p-value, $\widetilde{P}_j$, for testing H_j with strong protection of FWER is the maximum p-value for all intersection hypotheses containing j. Hence, for testing H_2, we get the multiplicity adjusted p-value

$$\widetilde{P}_2 = \max\{P_2, P_{12}, P_{23}, P_{123}\} = 0.0472,$$

and for testing H_3, we have

$$\widetilde{P}_3 = \max\{P_3, P_{13}, P_{23}, P_{123}\} = 0.0028.$$

Hence, as $\widetilde{P}_2 = 0.0472 > \alpha^* = 0.0125$, H_2, which concerns the subgroup of patients with low risk MammaPrint and ER negative, cannot be rejected. On the other hand, H_3, concerning the subgroup of patients with low risk MammaPrint, ER positive and HER2 positive, is rejected as $\widetilde{P}_3 = 0.0028 < \alpha^* = 0.0125$. In this example, the adaptive approach did not result in a negative trial, but the conclusion was not as strong as that obtained using the method developed in this chapter. The flexibility obtained by allowing arbitrary modifications at interim analysis points hence comes at a price. We also note that in contrast to our method, the prespecified weights are inefficient after stage one, and final test statistics are not functions of sufficient statistics.

5.9 Numerical Results

In this section, we present two simple numerical examples to demonstrate operating characteristics of the GSDS procedure. In the setting of two subgroups using DR-I, we show how stopping boundaries are computed, as well as power, point estimates and expected information. Example 1 is conducted without early stopping for rejection, while Example 2 does allow early rejection.

5.9.1 Example 1

Suppose we have two subgroups and plan for two stages of equal length. Then $\ell = 2$, $K = 2$, and $t_1 = 0.5$ and $t_2 = 1$. Set $\alpha = 0.05$, and define $\alpha^*_U(0.5) = 0$ and $\alpha^*_U(1) = \alpha$, so rejection can only occur after stage two. For lower spending function, define $\alpha^*_L(0.5) = (1 - \alpha)/2$ and $\alpha^*_L(1) = 1 - \alpha$. We can compute boundary values using an arbitrary value for total information (say 1), and then use a numerical search to find a value for $\mathcal{I}_{\max}$ that ensures sufficient power. Let $\theta^* = 1$ denote a clinically significant effect, and use Equation (5.13) to compute required maximum observed information. Note that the definition of $\alpha_L^*(t_1)$ ensures that under the global null hypothesis, the trial stops with early acceptance after the first stage with probability 0.475.

We compile results for $f_{01} = 1/4$ and $f_{01} = 1/2$, which can be seen in Tables (5.4) and (5.5). We first comment on the results, and then give equations, particular to this setup, for computing stopping boundaries, as well as the rejection probabilities, point estimates and expected sample size seen in the aforementioned tables.

First, we see in Tables (5.4) and (5.5) that conditional parameter estimates can be severely biased upwards, which is an expected consequence of the stage-one selection algorithm. To counter this, we apply the bootstrap algorithm discussed in Section 5.7. The algorithm is applied twice, using $B = 10{,}000$ iterations each time. The mean adjusted bias shown in Tables (5.4) and (5.5) is the result of the second adjustment, and we observe a substantial overall reduction in bias for the cases considered. Mean lower and upper endpoints of bootstrap 95% confidence intervals are also reported, and on average these are generally concordant with the desired conclusion. For example, when $\theta_1 = 0$, the resulting interval for θ_1 contains zero, and does not contain the clinically significant effect size $\theta^* = 1$. Coverage probabilities are also reported and can vary considerably. When the "correct" decision is made (e.g. selecting $\mathcal{P}^* = \{1, 2\}$ when $\theta_1 = \theta_2 = 1$, or selecting $\mathcal{P}^* = \{1\}$ when $\theta_1 = 1$ and $\theta_2 = 0$), coverage probabilities are approximately 95%. On the other hand, if an incorrect decision is made after the first stage, we see some decrease in coverage probability. For instance, if $\theta_1 = 1$ and $\theta_2 = 0$, and $\mathcal{P}^* = \{2\}$ is selected, coverage probability drops to 89%. It is well known that bootstrap samples are highly variable, which explains why the percentile-based intervals we use may appear as being too wide.

In Figures (5.1) and (5.2), we show results from running the bootstrap bias-reduction method using $B = 10{,}000$ replications for various values of θ_1 and θ_2. Plots depict raw bias, as obtained by evaluating Equation (5.17), and empirical adjusted bias as obtained by applying the bootstrap algorithm both once (single adj-bias) and twice (double adj-bias). We see that while raw bias decreases as θ gets large, adjusting twice all but eliminates any discernable bias for all values of θ_1 and θ_2 considered. As the algorithm is not computationally expensive (Matlab can run 10,000 simulations in fractions of a second), we conclude that adjusting twice rather than once is clearly beneficial. Further adjustments do not seem necessary.

The column $\mathbb{E}[\mathcal{I}_{\text{term}}]$ gives expected observed information at trial termination. The first figure is the expected observed information for the selected population(s), while the second is the total expected observed information. The upper halves of Tables (5.4) and (5.5) also include rejection probabilities and expected observed information for a group sequential design that uses the same spending functions, but does not allow any subgroup analysis. This design is discussed, for example, by Stallard and Facey (1996). As expected, when $\theta_1 = \theta_2 = 1$, the standard design reaches the correct conclusion (positive for the whole population) with greater probability than GSDS. However, when effect is limited to Ω_1, we see a clear advantage for our procedure. This is especially clear in Table 5.5, where $f_{01} = \frac{1}{4}$, and a strong effect in Ω_1 is hard to detect without specifically checking both subgroups. For example, when $\theta_1 = 2$ and $\theta_2 = 0$, our procedure rejects H_1

with probability 0.657, and either H_1 or H_0 with 0.858 probability. The standard procedure, on the other hand, only reaches a positive conclusion with probability 0.428.

In this setup, l_1 is found by solving

$$\alpha_L^*(t_1) = \frac{1-\alpha}{2} = \mathbb{P}_0\left[X_{11} \le l_1\sqrt{\mathcal{I}_{11}}, X_{12} \le l_1\sqrt{\mathcal{I}_{12}}\right],$$

which yields

$$l_1 = \Phi^{-1}\left(\sqrt{\frac{1-\alpha}{2}}\right) = 0.4936.$$

where Φ^{-1} denotes the inverse of the standard normal CDF. As we do not allow early stopping for rejection, $u_1 = \infty$. Using the values for l_1 and u_1, we find l_2 and u_2 with a two-dimensional numerical search, such that Equations (5.4) and (5.5) are satisfied. To this end, we compute

$$\begin{aligned}\psi_{2,\{1\}}(l_1,u_1,l_2,u_2;\boldsymbol{\theta}) &= \mathbb{P}_{\boldsymbol{\theta}}[\mathcal{P}^* = \{1\}] \cdot \tilde{\psi}_{2,\{1\}}(l_1,u_1,l_2,u_2;\boldsymbol{\theta}) \\ &= \Phi\left[\frac{l_{12} - \theta_2\mathcal{I}_{12}}{\sqrt{\mathcal{I}_{12}}}\right] \int_{l_{11}}^{\infty} \frac{1}{\sqrt{\mathcal{I}_{11}}}\varphi\left(\frac{y_{11} - \theta_1\mathcal{I}_{11}}{\sqrt{\mathcal{I}_{11}}}\right) \\ &\quad \times \Phi\left(\frac{y_{11} + \theta_1\Delta_{2,\{1\}} - u_2\sqrt{\mathcal{I}_{2,\{1\}}}}{\sqrt{\Delta_{2,\{1\}}}}\right) dy_{11}, \qquad (5.21)\end{aligned}$$

and

$$\begin{aligned}\xi_{2,\{1\}}(l_1,u_1,l_2,u_2;\boldsymbol{\theta}) &= \mathbb{P}_{\boldsymbol{\theta}}[\mathcal{P}^* = \{1\}] \cdot \tilde{\xi}_{2,\{1\}}(l_1,u_1,l_2,u_2;\boldsymbol{\theta}) \\ &= \Phi\left[\frac{l_{12} - \theta_2\mathcal{I}_{12}}{\sqrt{\mathcal{I}_{12}}}\right] \int_{l_{11}}^{\infty} \frac{1}{\sqrt{\mathcal{I}_{11}}}\varphi\left(\frac{y_{11} - \theta_1\mathcal{I}_{11}}{\sqrt{\mathcal{I}_{11}}}\right) \\ &\quad \times \Phi\left(\frac{u_2\sqrt{\mathcal{I}_{2,\{1\}}} - y_{11} - \theta_1\Delta_{2,\{1\}}}{\sqrt{\Delta_{2,\{1\}}}}\right) dy_{11}. \qquad (5.22)\end{aligned}$$

Here, $\Delta_{2,\{1\}}$ is the information increment for stage two, given that we are only sampling from Ω_1. In this example, it is assumed that all observations during stage two are taken from Ω_1 (including observations initially intended for Ω_2).

$\mathcal{I}_{2,\{1\}}$ denotes the cumulative observed information after stage two, given that stage two sampled only from Ω_1. We obtain $\psi_{2,\{2\}}$ and $\xi_{2,\{2\}}$ in a similar fashion.

For the choice $\mathcal{P}^* = \{1,2\}$, we compute (for brevity, we use 0 to denote the set $\{1,2\}$ below)

$$\psi_{2,0}(l_1, u_1, l_2, u_2; \boldsymbol{\theta}) = \int_{\tilde{l}_{1,0}}^{\infty} f_1(y_{10}|\boldsymbol{\theta}) \Phi\left(\frac{y_{10} + \theta_0 \Delta_{20} - u_2\sqrt{\mathcal{I}_{2,0}}}{\sqrt{\Delta_{20}}}\right) dy_{10}, \qquad (5.23)$$

$$\xi_{2,0}(l_1, u_1, l_2, u_2; \boldsymbol{\theta}) = \int_{\tilde{l}_{1,0}}^{\infty} f_1(y_{10}|\boldsymbol{\theta}) \Phi\left(\frac{u_2\sqrt{\mathcal{I}_{20}} - y_{10} - \theta_0 \Delta_{20}}{\sqrt{\Delta_{20}}}\right) dy_{10}, \qquad (5.24)$$

where

$$f_1(y_{10}|\boldsymbol{\theta}) = \int_{l_{11}}^{y_{10}-l_{12}} \frac{1}{\sqrt{\mathcal{I}_{11}\mathcal{I}_{12}}} \varphi\left(\frac{x_{11} - \theta_1 \mathcal{I}_{11}}{\sqrt{\mathcal{I}_{11}}}\right) \varphi\left(\frac{y_{10} - x_{11} - \theta_2 \mathcal{I}_{12}}{\sqrt{\mathcal{I}_{12}}}\right) dx_{11},$$

and Δ_{20} and $\mathcal{I}_{20}$ denote the information increment and total information for stage two, respectively.

Using identities (5.21-5.24), and the specified spending functions along with Equations (5.4) and (5.5), we get that

$$l_1 = 0.4936 \text{ and } u_2 = l_2 = 1.8937.$$

Note that these are on the scale of standardized statistics where, in the single-stage, single-population problem, using $\alpha = 0.05$ results in the one-sided boundary 1.645. For comparison, we can compute stopping boundaries for an equivalent design, only where no populations are eliminated after stage one (using the same error spending functions). In that case, $l_1 = -0.0627$ and $u_2 = l_2 = 1.6347$. As expected, the added flexibility of target population adaptation results in larger stopping boundaries.

Using Equation (5.17), we can write down $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_{\mathcal{S}}|\mathcal{P}^* = \mathcal{S}\right]$ for $\mathcal{S} = \{1\}, \{2\}$

and $\{1,2\}$. For $j=1,2$, we can show that

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_j|\mathcal{P}^*=\{j\}\right]=\theta_j+\frac{\sqrt{\Delta_{1j}}}{\mathcal{I}_{2,\{j\}}}\cdot\varphi\left(\frac{l_{1j}-\theta_j\Delta_{1j}}{\sqrt{\Delta_{1j}}}\right)\left[\Phi\left(\frac{\theta_j\Delta_{1j}-l_{1j}}{\sqrt{\Delta_{1j}}}\right)\right]^{-1}.$$

Note that, for large Δ_{1j}, if $\theta_j\neq 0$,

$$\varphi\left(\frac{l_{1j}-\theta_j\Delta_{1j}}{\sqrt{\Delta_{1j}}}\right)\approx 0 \text{ and } \left[\Phi\left(\frac{\theta_j\Delta_{1j}-l_{1j}}{\sqrt{\Delta_{1j}}}\right)\right]^{-1}\approx 1.$$

Hence, the conditional estimate of $\hat{\theta}_j$ given that $\mathcal{P}^*=\{j\}$ for $j=1,2$, is asymptotically unbiased as Δ_{1j} gets large. With a bit of algebra, we can also show that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_0|\mathcal{P}^*=\{1,2\}\right]\\ =\ &\frac{1}{\mathcal{I}_{20}}\left\{\theta_0\Delta_{20}+\int_{l_{11}}^{\infty}h_1(x_{11})\left[(\theta_2\Delta_{12}+x_{11})+\sqrt{\Delta_{12}}\frac{\varphi(w)}{\Phi(-w)}\right]dx_{11}\right\}\end{aligned}$$

where $w=\frac{l_{12}-\theta_2\Delta_{12}}{\sqrt{\Delta_{12}}}$. Hence we can again see that if Δ_{11} and Δ_{12} are large (and $\theta_1,\theta_2\neq 0$), then

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_0|\mathcal{P}^*=\{1,2\}\right]\approx\frac{1}{\mathcal{I}_{20}}(\theta_0\Delta_{10}+\theta_0\Delta_{20})=\theta_0.$$

5.9.2 Example 2

Again suppose we have two subgroups and plan for two stages. Spending functions are the same as in Example 1, but this time we set $\alpha_U^*(t_1)=\alpha/2$, which enables early stopping with rejection. We find l_1 in the same way as in Example 1, so $l_1=\Phi^{-1}\left(\sqrt{(1-\alpha)/2}\right)=0.4936$. Early stopping for rejection is now allowed, so we must also compute u_1. For this, we compute

$$\begin{aligned}\psi_{11}&=\mathbb{P}_{\boldsymbol{\theta}}\left[\mathcal{P}^*=\{1\}\right]\cdot\tilde{\psi}_{1,\{1\}}(l_1,u_1;\boldsymbol{\theta})\\&=\Phi\left(\frac{l_{12}-\theta_2\Delta_{12}}{\sqrt{\Delta_{12}}}\right)\cdot\Phi\left(\frac{\theta_1\Delta_{11}-u_1\sqrt{\Delta_{11}}}{\sqrt{\Delta_{11}}}\right),\\ \psi_{12}&=\Phi\left(\frac{l_{11}-\theta_1\Delta_{11}}{\sqrt{\Delta_{11}}}\right)\cdot\Phi\left(\frac{\theta_2\Delta_{12}-u_1\sqrt{\Delta_{12}}}{\sqrt{\Delta_{12}}}\right).\end{aligned}$$

For the decision $\mathcal{P}^* = \{1, 2\}$, we use Equation (5.11) and get

$$\psi_{10} = \int_{l_{11}}^{\tilde{u}_{10}-l_{12}} \varphi\left(\frac{x_{11} - \theta_1\Delta_{11}}{\sqrt{\Delta_{11}}}\right) \frac{1}{\sqrt{\Delta_{11}}} \Phi\left(\frac{\theta_2\Delta_{12} - (\tilde{u}_{10} - x_{11})}{\sqrt{\Delta_{12}}}\right) dx_{11}$$
$$+ \Phi\left(\frac{\theta_1\Delta_{11} - l_{11}}{\sqrt{\Delta_{11}}}\right) \Phi\left(\frac{\theta_1\Delta_{11} - (\tilde{u}_{10} - l_{12})}{\sqrt{\Delta_{11}}}\right).$$

Next, we compute $\psi_1 = \psi_{10} + \psi_{11} + \psi_{12}$, and perform a numerical search for u_1 such that $\psi_1 = \alpha_U^*(0.5) = \alpha/2$. Doing so for $f_{01} = \frac{1}{2}$ yields the first stage boundaries $(l_1, u_1) = (0.4936, 2.297)$. For stage two, l_2 and u_2 are found as in Example 1 where identities from Equations (5.21-5.24) are evaluated and a two-dimensional numerical search for l_2 and u_2 is conducted. Note that we now replace ∞ with the corresponding upper limits, e.g. use $\tilde{u}_{1,\{1\}}$ instead of ∞ in Equation (5.21). Doing this yields $l_2 = u_2 = 2.0980$.

To compute $\mathbb{E}\left[\hat{\theta}_{\mathcal{P}^*} | \mathcal{P}^* = \mathcal{S}\right]$, we again use Equation (5.17), but now we need to account for the possibility that the trial was terminated with rejection of $H_{\mathcal{P}^*}$ after the first stage. For $j = 1, 2$, we get

$$\begin{aligned}
&\int_{\tilde{u}_{1,\{j\}}}^{\infty} p_1(y_{1j} | \mathcal{P}^* = \{j\}; \boldsymbol{\theta}) \frac{y_{1j}}{\mathcal{I}_{1j}} dy_{11} \\
&= \frac{1}{\Delta_{1j}} \int_{\tilde{u}_{1,\{j\}}}^{\infty} y_{1j} h_j(y_{1j}) dy_{1j} \\
&= \frac{1}{\Delta_{1j}} \left[\Phi\left(\frac{\theta_j\Delta_{1j} - l_{1j}}{\sqrt{\Delta_{1j}}}\right)\right]^{-1} \\
&\quad \times \left\{\varphi\left(\frac{\tilde{u}_{1,\{j\}} - \theta_j\Delta_{1j}}{\sqrt{\Delta_{1j}}}\right)\sqrt{\Delta_{1j}} + \Phi\left(\frac{\theta_j\Delta_{1j} - \tilde{u}_{1,\{j\}}}{\sqrt{\Delta_{1j}}}\right)\theta_j\Delta_{1j}\right\},
\end{aligned}$$

and

$$\begin{aligned}
\int_{-\infty}^{\infty} p_2(y_{2j}; \boldsymbol{\theta}) \frac{y_{2j}}{\mathcal{I}_{2,\{j\}}} dy_{1j} &= \frac{1}{\mathcal{I}_{2,\{j\}}} \int_{l_{1j}}^{\tilde{u}_{1,\{j\}}} h_j(y_{1j})(y_{1j} + \theta_j\Delta_{2,\{j\}}) dy_{1j} \\
&= \frac{1}{\mathcal{I}_{2,\{j\}}} \Phi\left(-w_{lj}\right)^{-1} \Big\{ \left[\varphi(w_{lj}) - \varphi(w_{uj})\right] \sqrt{\Delta_{1j}} \\
&\quad + \left[\theta_j(\Delta_{1j} + \Delta_{2,\{j\}})\right] \left[\Phi(w_{uj}) - \Phi(w_{lj})\right] \Big\},
\end{aligned}$$

where $w_{uj} = \left(\tilde{u}_{1,\{j\}} - \theta_j \Delta_{1j}\right) / \sqrt{\Delta_{1j}}$ and $w_{lj} = (l_{1j} - \theta_j \Delta_{1j}) / \sqrt{\Delta_{1j}}$.

For $\mathcal{P}^* = \{1, 2\}$, we get (for brevity, 0 denotes the set $\{1, 2\}$)

$$\begin{aligned}
&\int_{\tilde{u}_{10}}^{\infty} p_1(y_{10}; \boldsymbol{\theta}) \frac{y_{10}}{\mathcal{I}_{10}} \\
&= \frac{1}{\mathcal{I}_{10}} \int_{\tilde{u}_{10}}^{\infty} y_{10} \int_{l_{11}}^{y_{10}-l_{12}} h_1(x_{11}) h_2(y_{10} - x_{12}) dx_{11} \; dy_{10} \\
&= \left[\mathcal{I}_{10} \Phi(-w_{l2})\right]^{-1} \\
&\quad \times \int_{l_{11}}^{\tilde{u}_{10}-l_{12}} h_1(x_{11}) \left[\varphi\left(\widetilde{w}(x_{11})\right) + (\theta_2 \Delta_{12} + x_{11}) \Phi\left(-\widetilde{w}(x_{11})\right)\right] dx_{11} \\
&\quad + \int_{\tilde{u}_{10}-l_{12}}^{\infty} h_1(x_{11}) \left[(\theta_2 \Delta_{12} + x_{11}) + \sqrt{\Delta_{12}} \frac{\varphi(w_{l2})}{\Phi(-w_{l2})}\right] dx_{11},
\end{aligned}$$

where w_{l2} is defined above, and $\widetilde{w}(x_{11}) = (\tilde{u}_{10} - x_{11} - \theta_2 \Delta_{12}) / \sqrt{\Delta_{12}}$. Finally,

$$\int_{-\infty}^{\infty} p_2(y_{20}; \boldsymbol{\theta}) \frac{y_{20}}{\mathcal{I}_{20}} dy_{20} = \frac{1}{\mathcal{I}_{20}} \left[I + \theta_0 \Delta_{20} \cdot II\right],$$

where

$$\begin{aligned}
I = &\left[\prod_{j=1}^{2} \Phi\left(-w_{lj}\right)\right]^{-1} \\
&\times \Bigg\{ \left[\Phi(\widetilde{w}_{u2}) - \Phi(w_{l2})\right] \cdot \left[\sqrt{\Delta_{11}} \left(\varphi(w_{l1}) - \varphi(\widetilde{w}_{u1})\right) + \theta_1 \Delta_{11} \left(\Phi(\widetilde{w}_{u1}) - \Phi(w_{l1})\right)\right] \\
&+ \left[\sqrt{\Delta_{12}} \left(\varphi(w_{l2}) - \varphi(\widetilde{w}_{u2})\right) + \theta_2 \Delta_{12} \left(\Phi(\widetilde{w}_{u2}) - \Phi(w_{l2})\right)\right] \cdot \left[\Phi(\widetilde{w}_{u1}) - \Phi(w_{l1})\right] \Bigg\},
\end{aligned}$$

and

$$II = \left[\prod_{j=1}^{2} \Phi\left(-w_{lj}\right)\right]^{-1} \Big(\Phi(\widetilde{w}_{u2}) - \Phi(w_{l2})\Big) \cdot \Big(\Phi(\widetilde{w}_{u1}) - \Phi(w_{l1})\Big).$$

Here, w_{lj} are as defined above, and

$$\widetilde{w}_{u1} = \frac{\tilde{u}_{10} - l_{12} - \theta_1 \Delta_{11}}{\sqrt{\Delta_{11}}}, \text{ and } \widetilde{w}_{u2} = \frac{\tilde{u}_{10} - \theta_2 \Delta_{12}}{\sqrt{\Delta_{12}}}.$$

We again use $\theta^* = 1$ as clinical significance, and Equation (5.13) to obtain the maximum observed information. Results are summarized in Tables 5.6 and 5.7.

Figures (5.3) and (5.4) show results from running the bootstrap bias-reduction method using $B = 10{,}000$ replications.

Many of the observations made in Example 1 hold true here as well. We do note that bias reduction is not always as successful as in Example 1, and for values of θ close to 2 there is sometimes lingering bias. For example, referring to Table 5.7 where $\theta_1 = 2$ and $\theta_2 = 0$, the mean (double) adjusted bias of $\hat{\theta}_1$ is 0.041. However, when compared to the raw bias of 0.307 this does not seem too alarming, and if we increase θ further (> 2) then lingering bias was observed to decrease towards zero again.

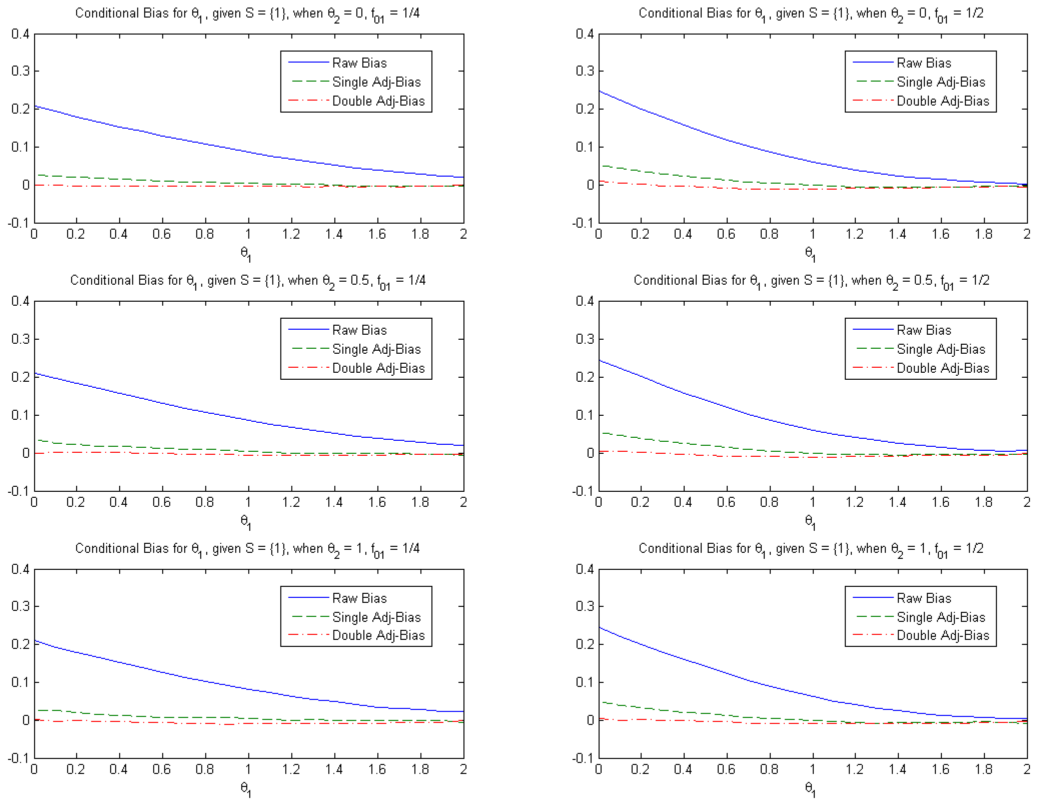
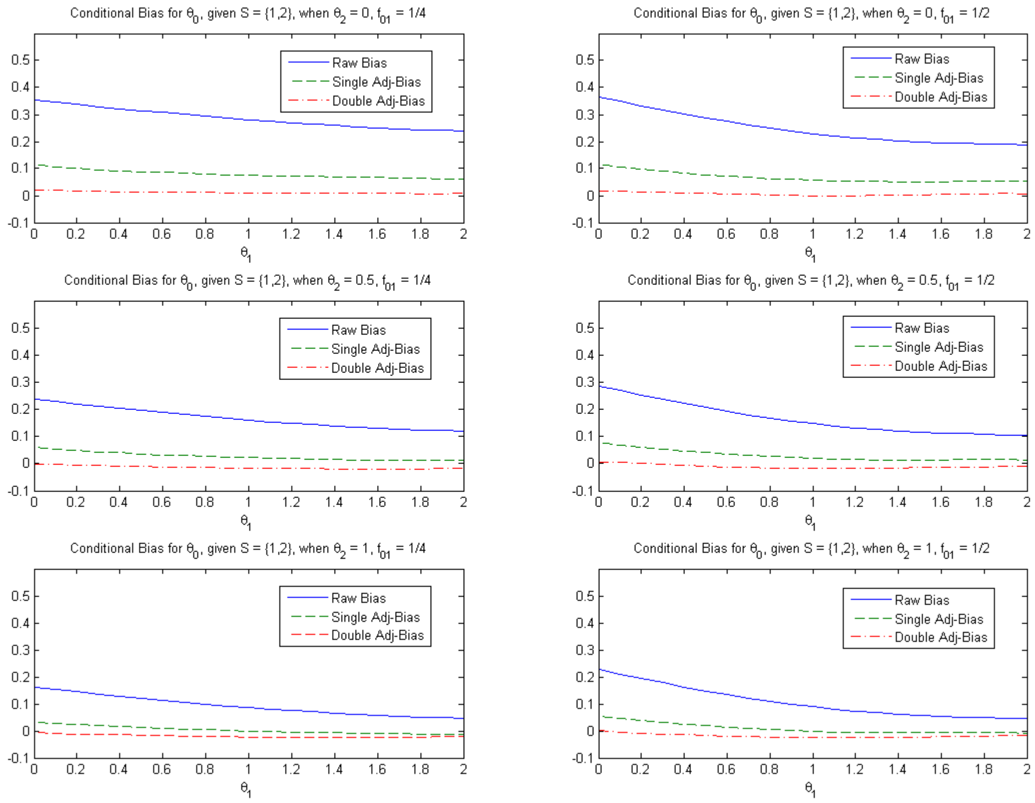


Figure 5.1: Results from bootstrap bias-reduction algorithm in example 1, in which $u_1 = \infty$. Plots depict conditional bias of θ_1 given $\mathcal{S} = \{1\}$ for various values of θ_2. In the right column, $f_{01} = \frac{1}{4}$, and in the left column, $f_{01} = \frac{1}{2}$.



Conditional Bias for θ_0, given S = {1,2}, when θ_2 = 0, f_{01} = 1/4

Conditional Bias for θ_0, given S = {1,2}, when θ_2 = 0, f_{01} = 1/2

Conditional Bias for θ_0, given S = {1,2}, when θ_2 = 0.5, f_{01} = 1/4

Conditional Bias for θ_0, given S = {1,2}, when θ_2 = 0.5, f_{01} = 1/2

Conditional Bias for θ_0, given S = {1,2}, when θ_2 = 1, f_{01} = 1/4

Conditional Bias for θ_0, given S = {1,2}, when θ_2 = 1, f_{01} = 1/2

Figure 5.2: Results from bootstrap bias-reduction algorithm in example 1, in which $u_1 = \infty$. Plots depict conditional bias of θ_0 given $\mathcal{S} = \{1, 2\}$, for various values of θ_2. In the right column, $f_{01} = \frac{1}{4}$, and in the left column, $f_{01} = \frac{1}{2}$.

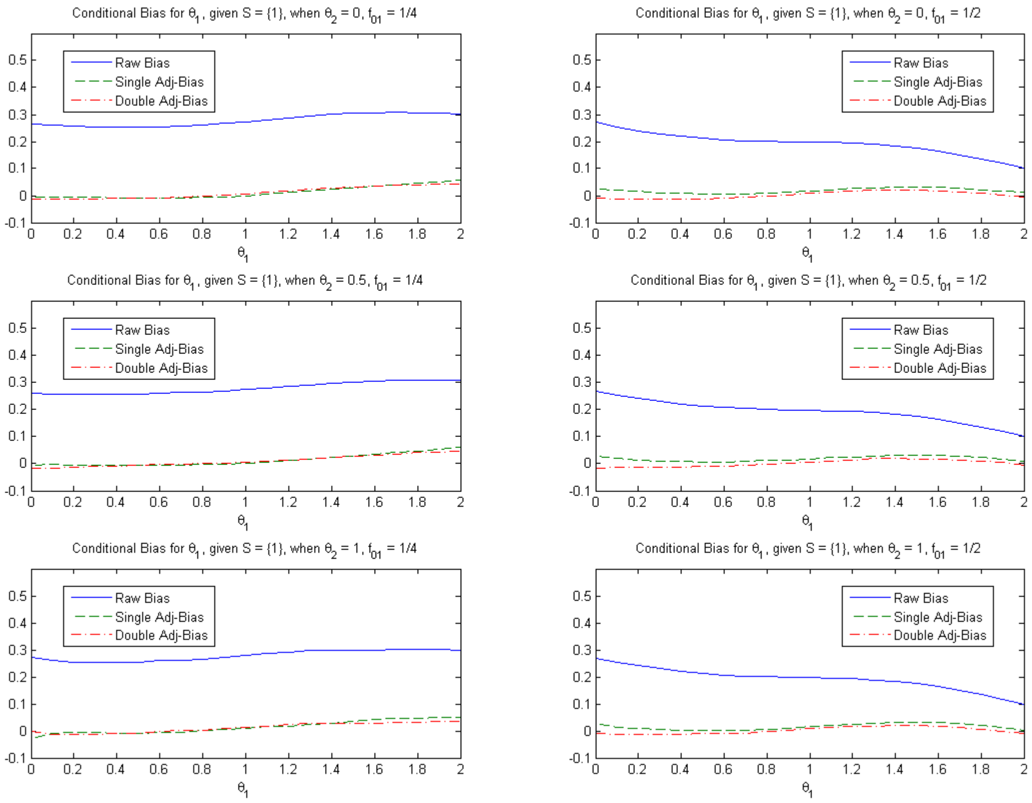


Figure 5.3: Results from bootstrap bias-reduction algorithm in example 2, in which $u_1 < \infty$. Plots depict conditional bias of θ_1 given $\mathcal{S} = \{1\}$ for various values of θ_2. In the right column, $f_{01} = \frac{1}{4}$, and in the left column, $f_{01} = \frac{1}{2}$.

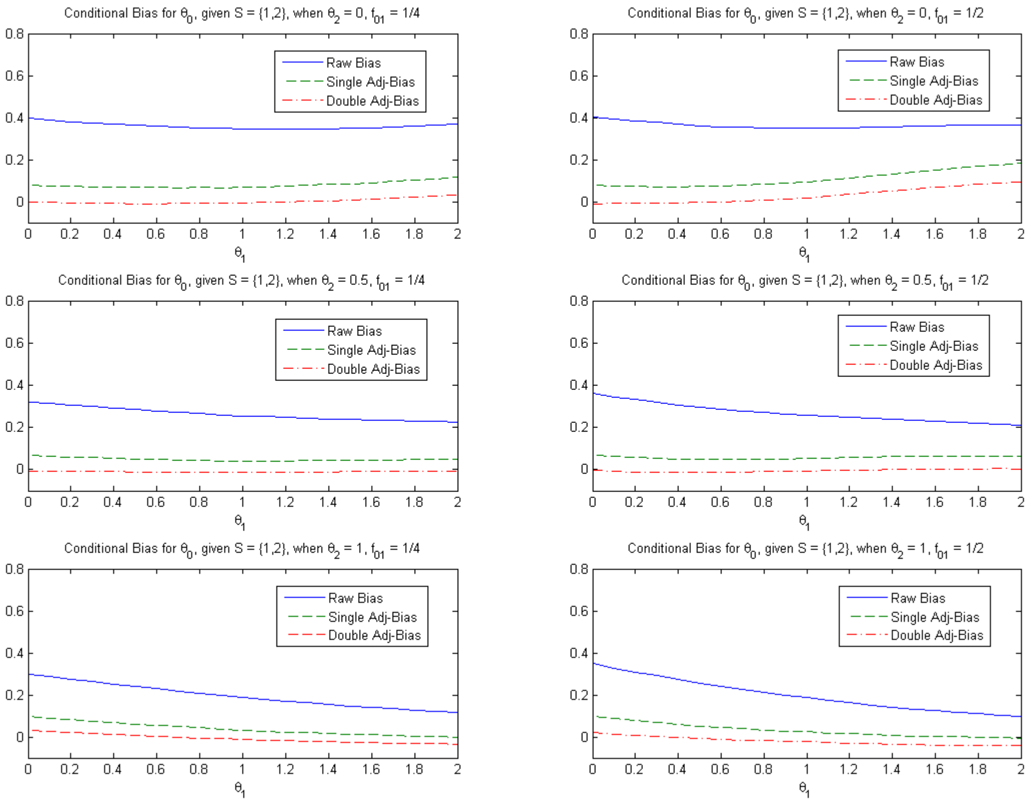


Figure 5.4: Results from bootstrap bias-reduction algorithm in example 2, in which $u_1 < \infty$. Plots depict conditional bias of θ_0 given $\mathcal{S} = \{1, 2\}$, for various values of θ_2. In the right column, $f_{01} = \frac{1}{4}$, and in the left column, $f_{01} = \frac{1}{2}$.

Table 5.4: GSDS operating characteristics for $\alpha = 0.05$, $f_{01} = f_{02} = 1/2$, $\alpha^*_U(t_1) = 0, \alpha^*_U(t_2) = \alpha, \alpha^*_L(t_1) = (1-\alpha)/2, \alpha^*_L(t_2) = 1-\alpha$. Boundaries are $l_1 = 0.4936$, $u_1 = \infty$, $l_2 = u_2 = 1.8937$. $\mathcal{I}_{\max} = 9.46$ which gives a positive result with probability 0.9 when $\theta_0 = \theta_1 = \theta_2 = 1$. Parenthesized values in the **upper table** refer to corresponding properties of a group sequential trial that only considers Ω_0 as a whole (Stallard and Facey, 1996). Spending functions are the same, critical values are $l_1 = -0.0627$ and $l_2 = u_2 = 1.6347$, and $\mathcal{I}_{\max} = 8.65$ which gives power 0.9 when $\theta_0 = 1$. Bold-faced entries indicate the "correct" decision given (θ_1, θ_2). Parenthesized values in the **lower table** refer to standard error of conditional bias, obtained numerically. Adjusted conditional bias estimates are a result of double-bootstrap adjustment, and confidence interval points are average upper and lower endpoints based on 10,000 simulations.

Effect Size	Final Decision				Stage 1 Decision				$\mathbb{E}[\hat{\theta}_{\mathcal{S}}\|\mathcal{S}]$			$\mathbb{E}[\mathcal{I}_{\text{term}}]$
$(\theta_0, \theta_1, \theta_2)$	None	H_0	H_1	H_2	Futile	Both	Ω_1	Ω_2	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	
(0,0,0)	**0.950** (0.950)	0.016 (0.050)	0.017	0.017	**0.475** (0.475)	0.097 (0.525)	0.214	0.214	0.370 (0.183)	0.246	0.246	6.2/7.2 (6.6)
(1,1,1)	0.100 (0.100)	**0.688** (0.900)	0.106	0.106	0.022 (0.017)	**0.726** (0.983)	0.126	0.126	1.088 (1.011)	1.059	1.059	8.8/9.4 (8.6)
(0.5,1,0)	0.331 (0.572)	0.173 (0.428)	**0.492**	0.004	0.102 (0.135)	0.265 (0.865)	**0.587**	0.046	0.729 (0.560)	1.059	0.246	7.5/9.0 (8.1)
(1,2,0)	0.012 (0.100)	0.302 (0.900)	**0.686**	0.000	0.003 (0.017)	0.309 (0.983)	**0.686**	0.002	1.187 (1.011)	2.003	0.246	7.8/9.4 (8.6)

Effect Size	Bias (Std.Err.)			Adj-Bias			CI: $\overline{\text{low}}/\overline{\text{upp}}$			CI: Cov. Prob.		
$(\theta_0, \theta_1, \theta_2)$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
(0,0,0)	0.370 (0.259)	0.246 (0.327)	0.246 (0.327)	0.018	0.010	0.010	-0.051 0.981	-0.341 0.948	-0.346 0.941	0.610	0.874	0.881
(1,1,1)	0.088 (0.295)	0.059 (0.352)	0.059 (0.352)	-0.024	-0.008	-0.007	0.547 1.698	0.391 1.771	0.392 1.772	0.949	0.951	0.949
(0.5,1,0)	0.229 (0.277)	0.059 (0.352)	0.246 (0.327)	-0.011	-0.009	0.003	0.247 1.341	0.391 1.770	-0.345 0.943	0.862	0.952	0.888
(1,2,0)	0.187 (0.292)	0.003 (0.373)	0.246 (0.327)	0.007	-0.005	-0.046	0.665 1.813	1.278 2.733	-0.383 0.897	0.896	0.948	0.945

Table 5.5: GSDS operating characteristics for $\alpha = 0.05$, $f_{01} = 1/4, f_{02} = 3/4$, $\alpha^*_U(t_1) = 0, \alpha^*_U(t_2) = \alpha$, $\alpha^*_L(t_1) = (1-\alpha)/2, \alpha^*_L(t_2) = 1-\alpha$. Boundaries are $l_1 = 0.4936$, $u_1 = \infty$, $l_2 = u_2 = 1.8707$. $\mathcal{I}_{\max} = 9.44$ which gives a positive result with probability 0.9 when $\theta_0 = \theta_1 = \theta_2 = 1$. Parenthesized values in the **upper table** refer to corresponding properties of a group sequential trial that only considers Ω_0 as a whole (Stallard and Facey, 1996). Spending functions are the same, critical values are $l_1 = -0.0627$ and $l_2 = u_2 = 1.6347$, and $\mathcal{I}_{\max} = 8.65$ which gives power 0.9 when $\theta_0 = 1$. Bold-faced entries indicate the "correct" decision given (θ_1, θ_2). Parenthesized values in the **lower table** refer to standard error of conditional bias, obtained numerically. Adjusted conditional bias estimates are a result of double-bootstrap adjustment, and confidence interval points are average upper and lower endpoints based on 10,000 simulations.

Effect Size	Final Decision				Stage 1 Decision				$\mathbb{E}[\hat{\theta}_{\mathcal{S}} \mid \mathcal{S}]$			$\mathbb{E}[\mathcal{I}_{\text{term}}]$
$(\theta_0, \theta_1, \theta_2)$	None	H_0	H_1	H_2	Futile	Both	Ω_1	Ω_2	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	
(0,0,0)	**0.950** (0.950)	0.016 (0.050)	0.015	0.019	**0.475** (0.475)	0.097 (0.525)	0.214	0.214	0.357 (0.183)	0.209	0.259	6.1/7.1 (6.6)
(1,1,1)	0.100 (0.100)	**0.629** (0.900)	0.047	0.224	0.024 (0.017)	**0.664** (0.983)	0.060	0.254	1.086 (1.011)	1.085	1.038	8.7/9.2 (8.6)
(0.25,1,0)	0.513 (0.819)	0.085 (0.181)	**0.394**	0.008	0.191 (0.280)	0.225 (0.720)	**0.499**	0.086	0.530 (0.363)	1.085	0.259	6.6/8.4 (7.4)
(.5,2,0)	0.142 (0.572)	0.201 (0.428)	**0.657**	0.001	0.032 (0.135)	0.296 (0.865)	**0.657**	0.015	0.738 (0.560)	2.019	0.259	6.7/9.2 (8.1)

Effect Size	Bias (Std.Err.)			Adj-Bias			CI: $\overline{\text{low}}/\overline{\text{upp}}$			CI: Cov. Prob.		
$(\theta_0, \theta_1, \theta_2)$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
(0,0,0)	0.357 (0.259)	0.209 (0.381)	0.259 (0.289)	0.024	-0.000	0.007	-0.062 0.978	-0.512 0.991	-0.249 0.902	0.619	0.912	0.849
(1,1,1)	0.086 (0.299)	0.085 (0.391)	0.038 (0.328)	-0.021	-0.001	-0.012	0.537 1.707	0.3322 1.8754	0.420 1.704	0.948	0.947	0.956
(0.5,1,0)	0.280 (0.265)	0.085 (0.391)	0.259 (0.289)	0.009	-0.004	0.011	0.080 1.143	0.330 1.872	-0.246 0.905	0.786	0.945	0.851
(1,2,0)	0.238 (0.273)	0.019 (0.404)	0.259 (0.289)	0.011	-0.008	0.007	0.268 1.355	1.227 2.817	-0.250 0.901	0.843	0.950	0.853

Table 5.6: GSDS operating characteristics for $\alpha = 0.05$, $f_{01} = 1/2, f_{02} = 1/2, \alpha^*_U(t_1) = \alpha/2, \alpha^*_U(t_2) = \alpha, \alpha^*_L(t_1) = (1-\alpha)/2, \alpha^*_L(t_2) = 1-\alpha$. Critical boundaries are $l_1 = 0.4936$, $u_1 = 2.2976$, $l_2 = u_2 = 2.0980$. $\mathcal{I}_{\max} = 10.30$ which gives a positive result with probability 0.9 when $\theta_0 = \theta_1 = \theta_2 = 1$. Bold-faced entries indicate the "correct" decision given (θ_1, θ_2). Parenthesized values in the **lower table** refer to standard error of conditional bias, obtained numerically. Adjusted conditional bias estimates are a result of double-bootstrap adjustment, and confidence interval points are average upper and lower endpoints based on 10,000 simulations.

Effect Size	Final Decision				Stage 1 Decision							$\mathbb{E}[\mathcal{I}_{\text{term}}]$
$(\theta_0, \theta_1, \theta_2)$	None	H_0	H_1	H_2	Futile	Rej. H_0	Rej. H_1	Rej. H_2	Cont. Ω_0	Cont. Ω_1	Cont. Ω_2	
(0,0,0)	**0.950**	0.017	0.016	0.016	**0.475**	0.010	0.007	0.007	0.087	0.207	0.207	6.9/8.0
(1,1,1)	0.090	**0.710**	0.095	0.095	0.018	**0.472**	0.033	0.033	**0.279**	0.083	0.083	7.1/7.7
(.5,1,0)	0.330	0.178	**0.491**	0.003	0.092	0.099	**0.168**	0.001	0.170	**0.429**	0.040	7.0/8.7
(1,2,0)	0.010	0.304	**0.687**	0.000	0.002	0.260	**0.565**	0.000	0.050	**0.122**	0.001	4.4/6.2

Effect Size	Bias (Std.Err.)			Adj-Bias			CI: $\overline{\text{low}}/\overline{\text{upp}}$			CI: Cov. Prob.		
$(\theta_0, \theta_1, \theta_2)$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
(0,0,0)	0.415 (0.344)	0.274 (0.400)	0.274 (0.400)	-0.010	-0.009	-0.013	-0.051 1.21	-0.334 1.317	-0.337 1.311	0.691	0.903	0.904
(1,1,1)	0.193 (0.340)	0.202 (0.486)	0.202 (0.486)	-0.023	0.010	0.010	0.606 1.963	0.449 2.278	0.442 2.270	0.851	0.881	0.883
(0.5,1,0)	0.354 (0.375)	0.202 (0.486)	0.274 (0.400)	0.014	0.009	-0.012	0.307 1.662	0.444 2.273	-0.337 1.309	0.711	0.882	0.905
(1,2,0)	0.370 (0.338)	0.112 (0.495)	0.274 (0.400)	0.095	-0.005	-0.036	0.811 2.169	1.252 3.209	-0.356 1.265	0.636	0.971	0.944

Table 5.7: GSDS operating characteristics for $\alpha = 0.05$, $f_{01} = 1/4, f_{02} = 3/4, \alpha^*_U(t_1) = \alpha/2, \alpha^*_U(t_2) = \alpha, \alpha^*_L(t_1) = (1-\alpha)/2, \alpha^*_L(t_2) = 1 - \alpha$. Boundaries are $l_1 = 0.4936$, $u_1 = 2.2782$, $l_2 = u_2 = 2.0772$. $\mathcal{I}_{\max} = 10.31$ which gives a positive result with probability 0.9 when $\theta_0 = \theta_1 = \theta_2 = 1$. Bold-faced entries indicate the "correct" decision given (θ_1, θ_2). Parenthesized values in the **lower table** refer to standard error of conditional bias, obtained numerically. Adjusted conditional bias estimates are a result of double-bootstrap adjustment, and confidence interval points are average upper and lower endpoints based on 10,000 simulations.

Effect Size	Final Decision				Stage 1 Decision							$\mathbb{E}[\mathcal{I}_{\text{term}}]$
$(\theta_0, \theta_1, \theta_2)$	None	H_0	H_1	H_2	Futile	Rej. H_0	Rej. H_1	Rej. H_2	Cont. Ω_0	Cont. Ω_1	Cont. Ω_2	
(0,0,0)	**0.950**	0.016	0.016	0.017	**0.475**	0.009	0.008	0.008	0.087	0.206	0.206	6.8/7.9
(1,1,1)	0.093	**0.649**	0.040	0.211	0.018	**0.434**	0.009	0.098	**0.254**	0.043	0.144	7.1/7.6
(.25,1,0)	0.516	0.085	**0.400**	0.006	0.180	0.042	**0.089**	0.003	0.188	**0.422**	0.078	6.8/8.9
(.5,2,0)	0.205	0.208	**0.663**	0.001	0.026	0.119	**0.342**	0.000	0.1802	**0.321**	0.011	5.3/7.9

Effect Size	Bias (Std.Err.)			Adj-Bias			CI: $\overline{\text{low}}/\overline{\text{upp}}$			CI: Cov. Prob.		
$(\theta_0, \theta_1, \theta_2)$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
(0,0,0)	0.398 (0.339)	0.268 (0.535)	0.276 (0.336)	-0.005	-0.015	-0.009	-0.064 1.267	-0.498 1.745	-0.244 1.142	0.706	0.929	0.876
(1,1,1)	0.193 (0.347)	0.275 (0.645)	0.151 (0.406)	-0.012	0.009	0.014	0.598 1.981	0.376 2.847	0.479 2.041	0.853	0.824	0.891
(0.25,1,0)	0.351 (0.356)	0.275 (0.645)	0.276 (0.336)	-0.010	0.007	-0.009	0.095 1.431	0.373 2.876	-0.244 1.144	0.758	0.827	0.880
(0.5,2,0)	0.372 (0.364)	0.307 (0.616)	0.276 (0.336)	0.031	0.041	0.019	0.339 1.673	1.304 3.785	-0.224 1.172	0.674	0.835	0.867

5.9.3 Comparison with FE and HUT

We conclude with a numerical study comparing the operating characteristics of GSDS with those of the methods FE (see Section 4.2) and HUT (see Section 4.3), developed earlier in this thesis. The comparison is carried out in the setting of three disjoint subgroups Ω_1, Ω_2 and Ω_3 ($\mathcal{P} = \{1, 2, 3\}$), where a nesting effect is considered likely, i.e. $\theta_1 \geq \theta_2 \geq \theta_3$, and θ_j is the effect size for Ω_j. In the case of GSDS, this suggests the use of DR-II at the first interim analysis.

Setup

We assume that patient responses are normally distributed with mean μ_j, $j \in \mathcal{P}$ and common variance $\sigma^2 = 5$. Define $\theta_j = \mu_j^E - \mu_j^C$ as the mean difference between treatment and control for population Ω_j, $j \in \mathcal{P}$. The three hypotheses of interest are

$$H_{\{1,2,3\}} : \theta_{\{1,2,3\}} = 0,\ H_{\{1,2\}} : \theta_{\{1,2\}} = 0,\ \text{and}\ H_{\{1\}} : \theta_{\{1\}} = 0.$$

For shorthand notation, we use H_0 and θ_0 to respectively denote the hypothesis $H_{\{1,2,3\}}$ and effect size $\theta_{\{1,2,3\}}$. Let $\theta^* = 1$ represent clinical significance.

Operating characteristics for GSDS and FE can be obtained numerically, while Monte Carlo simulation is conducted for HUT, using 100,000 replications per parameter configuration. All procedures are compared using the same sample size (observed information), and maximum observed information is set equal to $\mathcal{I}_{\max} = \frac{n_{\max}}{4\sigma^2} = 25$, which is equivalent to a maximum sample size of 500 patients. Of the three procedures being investigated, only GSDS allows early rejections. Hence FE and HUT will always use the total allotted sample size (unless stopping early for futility).

Following the blueprint offered by Wang et al. (2009), we consider prevalence levels ranging from low $(f_{01}, f_{02}, f_{03}) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$, to moderate $(f_{01}, f_{02}, f_{03}) = \left(\frac{3}{8}, \frac{1}{8}, \frac{1}{2}\right)$, to high $(f_{01}, f_{02}, f_{03}) = \left(\frac{9}{16}, \frac{3}{16}, \frac{1}{4}\right)$. Interim analysis times are $t = \frac{1}{4}, \frac{1}{2}$ or $\frac{3}{4}$. We consider three main effect size patterns, which respectively correspond to the three patterns identified to be of importance by Wang et al. (2009):

- *Favorable nesting pattern* corresponds to the pattern $\theta_1 > \theta_2 > \theta_3$, or $\theta_0 < \theta_{\{1,2\}} < \theta_{\{1\}}$. In this case, subgroup classification is predictive of treatment response.
- *No nesting pattern* corresponds to the situation where there is no favorable nesting pattern. For example, this is the case when $\theta_0 = 0.8, \theta_{\{1,2\}} = 1.2$, and $\theta_{\{1\}} = 0.7$. Here, subgroup classification is not predictive of treatment response.
- *Homogeneity pattern* is present when treatment effects are roughly equal across all subgroups, i.e. when $\theta_1 \approx \theta_2 \approx \theta_3$.

Both FE and GSDS are use empirical data weights by design and hence their strong control of FWER has been established. When empirical data weights are applied for the HUT design, control of FWER is not guaranteed. We hence investigate empirical Type I error rates to inspect the extent of error rate inflation (if any).

For the GSDS procedure, error spending is set to be proportional to stage length, so $\alpha^*_U(t) = \alpha \cdot t$ and $\alpha^*_L(t) = (1 - \alpha) \cdot t$. We do allow early rejection for GSDS, but as neither FE nor HUT can do so, we also consider the case of no GSDS early rejection ($\alpha^*_U(t) = 0$). For FE, we use two different configurations of local significance levels and enrichment parameters. Namely, $(\alpha_0, \alpha_1, \alpha_2) \in \{(\alpha/3, \alpha/3, \alpha/3), (0.015, 0.007, 0.003)\}$, and $(\gamma_0, \gamma_1, \gamma_2) \in$

$\{(0.5, 0.5, 0.5), (0.3, 0.4, 0.5)\}$. For HUT, we keep θ^- fixed at zero, while $\theta^+ \in \{0, 0.5, 1\}$, and $k^{HUT} \in \{0.5, 1, 2\}$.

Effect of Empirical Data Weights on Type I Error

In Table 5.8 we show simulated Type I error probabilities for the HUT design, based on the use of empirical data weights. The table relies on procedure-specific parameters $\theta^+ = 0$ and $k^{HUT} = 2$. The choice of these parameter values exaggerate the outcome in the sense that small θ^+ and large k^{HUT} will tend to encourage the inclusion of more populations for the second stage. Hence, under this configuration the HUT design will proceed with populations it otherwise might not have, had θ^+ been larger or k^{HUT} smaller. We see that for the cases considered, Type I error probabilities do not exceed $\alpha = 0.025$. Comparison with Type I error rates obtained with pre-specified weights (proportional to interim analysis timing, not reported in a table) indicates that inflation is essentially negligible. In no situation did we observe error rates in excess of the desired FWER of $\alpha = 0.025$.

In general, error rate inflation as compared to pre-specified weights is not common. Wang et al. (2009) report negligible Type I error inflation when evaluating a number of different designs that rely on using conditional power at the interim analysis. They do obtain error rates that exceed the nominal α (a maximum error rate of 0.0263 is reported), but this occurs in designs that include a potential increase in sample size at the interim analysis.

Table 5.8: Empirical Type I error probabilities for the HUT design with $t = \frac{1}{2}$, with and without futility stopping. Probabilities are based on empirical data weights. A total of 100,000 simulation iterations were used, and maximum observed information is $\mathcal{I}_{\max} = 25$. HUT-specific parameters are set as $k = 2$ and $\theta^{+} = 0.0$, so as to imply a *worst-case scenario.* Note that $f_{\{1,2\}} = f_{0,\{1,2\}}$ and $f_{\{1\}} = f_{0,\{1\}}$.

Effect Sizes					With Futility Stopping				No Futility Stopping			
θ_0	$\theta_{\{1,2\}}$	θ_1	θ_2	θ_3	Ω_0	$\Omega_{\{1,2\}}$	Ω_1	T-I Err.	Ω_0	$\Omega_{\{1,2\}}$	Ω_1	T-I Err.
$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$												
0	0	0	0	0	0.0113	0.0089	0.0124	0.0184	0.0111	0.0088	0.0127	0.0183
0.1	0	0	0	0.4	N/A	0.0117	0.0145	0.0179	N/A	0.0116	0.0149	0.0184
0.075	0.1	0	0.4	0	N/A	N/A	0.0166	0.0166	N/A	N/A	0.0163	0.0163
0.275	0.1	0	0.4	0.8	N/A	N/A	0.0213	0.0213	N/A	N/A	0.0206	0.0206
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$												
0	0	0	0	0	0.0106	0.0082	0.0117	0.0198	0.0106	0.0080	0.0116	0.0199
0.2	0	0	0	0.4	N/A	0.0120	0.0154	0.0188	N/A	0.0110	0.0144	0.0180
0.05	0.1	0	0.4	0	N/A	N/A	0.0142	0.0142	N/A	N/A	0.0136	0.0136
0.45	0.1	0	0.4	0.8	N/A	N/A	0.0206	0.0206	N/A	N/A	0.0202	0.0202
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$												
0	0	0	0	0	0.0102	0.0047	0.0105	0.0209	0.0099	0.0053	0.0110	0.0214
0.2	0	0	0	0.4	N/A	0.0074	0.0128	0.0192	N/A	0.0074	0.0133	0.0194
0.1	0.2	0	0.4	0	N/A	N/A	0.0140	0.0140	N/A	N/A	0.0144	0.0144
0.5	0.2	0	0.4	0.8	N/A	N/A	0.0203	0.0203	N/A	N/A	0.0197	0.0197

Power Performance

Tables 5.9–5.14 show power performance for the methods under consideration, using interim timing $t = \frac{1}{2}$. Performance is reported for various configurations of procedure-specific parameters for easy comparison. We consider two examples of each of the three effect size patterns discussed above. For favorable nesting patterns, we use $(\theta_1, \theta_2, \theta_3) = (1, 0.4, 0)$ and $(1.5, 0.2, 0)$. For (unfavorable) patterns with no nesting, we use $(\theta_1, \theta_2, \theta_3) = (0.4, 1.2, 0.4)$ and $(0.2, 1.5, 0.2)$. Finally, the two homogeneity patterns $(\theta_1, \theta_2, \theta_3) = (0.4, 0.4, 0.4)$ and $(0.8, 0.8, 0.8)$ are considered. Tables are divided into three main sections, each corresponding to particular prevalence levels of Ω_1, Ω_2 and Ω_3.

Recall that a clinically significant effect size *worth detecting* has been identified as $\theta^* = 1$. Hence, if a population (composite or not) has effect size "considerably smaller" than 1, a positive finding for said population is not desirable. For example, consider the two homogeneity patterns shown above. The first, $(0.4, 0.4, 0.4)$, represents a scenario where there is a slight and constant effect across all populations. This effect is however not large enough to be considered meaningful (likely any trial would not be powered to detect this effect anyway). The second pattern, $(0.8, 0.8, 0.8)$, is close enough to clinical significance for a trial sponsor to be interested in detecting this effect. A favorable outcome in this scenario is to reject the overall null hypothesis H_0, though positive findings in smaller populations are also of interest.

Impact of Nesting Pattern: When a favorable nesting pattern is present, i.e. $(\theta_1, \theta_2, \theta_3) = (1, 0.4, 0)$ or $(1.5, 0.2, 0)$, it is desirable to find a positive result in Ω_1 only. In other words we want to reject $H_{\{1\}}$ only, and accept all other hypotheses. We see in Tables 5.9 and 5.10 that when the nesting pattern is present,

Table 5.9: Power performance of three two-stage enrichment designs when $(\theta_1, \theta_2, \theta_3) = (1, 0.4, 0)$. As an example, the column "$\Omega_{\{1,2\}}$ (w/other)" gives the probability that only $H_{\{1,2\}}$ is rejected, and the parenthesized value includes events where other hypotheses are also rejected. Prevalence levels are given in the table. Entries are based on $\mathcal{I}_{\max} = 25$ and $t = 1/2$. HUT parameters are $\theta^+ = 0.5$, and $k^{HUT} = 0.5, 1$ and 2, respectively. FE parameters are $(\gamma_0, \gamma_1, \gamma_2) = (0.5, 0.5, 0.5)$ in the first two rows, and (0.3,0.4,0.5) in rows three and four. Local significance levels are $\alpha/3$ in rows one and three, and (0.015,0.007,0.003) in rows two and four. The top GSDS row refers to no early rejection.

Effect Sizes			Procedure	Power Performance			
θ_0	$\theta_{\{1,2\}}$	$\theta_{\{1\}}$		Ω_0	$\Omega_{\{1,2\}}$ (w/other)	$\Omega_{\{1\}}$ (w/other)	Any
$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$							
0.6375	0.85	1	HUT	0.20	0.32 (0.52)	0.42 (0.94)	0.95
				0.32	0.34 (0.66)	0.29 (0.96)	0.96
				0.45	0.32 (0.77)	0.18 (0.96)	0.97
			FE	0.79	0.05 (0.55)	0.01 (0.56)	0.92
				0.84	0.02 (0.55)	0.01 (0.56)	0.92
				0.78	0.05 (0.56)	0.01 (0.57)	0.93
				0.83	0.03 (0.55)	0.01 (0.56)	0.93
			GSDS	0.12	0.33 (N/A)	0.43 (N/A)	0.88
				0.10	0.33 (N/A)	0.43 (N/A)	0.86
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$							
0.425	0.85	1	HUT	0.13	0.33 (0.46)	0.42 (0.88)	0.89
				0.21	0.39 (0.59)	0.31 (0.89)	0.91
				0.29	0.41 (0.68)	0.20 (0.89)	0.91
			FE	0.39	0.14 (0.36)	0.02 (0.39)	0.79
				0.48	0.09 (0.35)	0.01 (0.36)	0.78
				0.39	0.16 (0.38)	0.03 (0.41)	0.82
				0.47	0.11 (0.36)	0.02 (0.38)	0.81
			GSDS	0.09	0.29 (N/A)	0.44 (N/A)	0.82
				0.07	0.14 (N/A)	0.43 (N/A)	0.79
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$							
0.35	0.7	1	HUT	0.11	0.29 (0.39)	0.37 (0.76)	0.81
				0.17	0.33 (0.47)	0.29 (0.77)	0.83
				0.23	0.34 (0.53)	0.21 (0.75)	0.82
			FE	0.26	0.28 (0.50)	0.13 (0.63)	0.72
				0.34	0.20 (0.49)	0.08 (0.57)	0.67
				0.25	0.31 (0.53)	0.14 (0.67)	0.75
				0.33	0.24 (0.51)	0.09 (0.61)	0.71
			GSDS	0.07	0.33 (N/A)	0.34 (N/A)	0.74
				0.05	0.18 (N/A)	0.34 (N/A)	0.71

Table 5.10: Power performance of three two-stage enrichment designs when $(\theta_1, \theta_2, \theta_3) = (1.5, 0.2, 0)$. As an example, the column "$\Omega_{\{1,2\}}$ (w/other)" gives the probability that only $H_{\{1,2\}}$ is rejected, and the parenthesized value includes events where other hypotheses are also rejected. Prevalence levels are given in the table. Entries are based on $\mathcal{I}_{\max} = 25$ and $t = 1/2$. HUT parameters are $\theta^+ = 0.5$, and $k^{HUT} = 0.5, 1$ and 2, respectively. FE parameters are $(\gamma_0, \gamma_1, \gamma_2) = (0.5, 0.5, 0.5)$ in the first two rows, and (0.3,0.4,0.5) in rows three and four. Local significance levels are $\alpha/3$ in rows one and three, and (0.015,0.007,0.003) in rows two and four. The top GSDS row refers to no early rejection.

Effect Sizes			Procedure	Power Performance			
θ_0	$\theta_{\{1,2\}}$	$\theta_{\{1\}}$		Ω_0	$\Omega_{\{1,2\}}$ (w/other)	$\Omega_{\{1\}}$ (w/other)	Any
$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$							
0.881	1.175	1.5	HUT	0.21	0.27 (0.48)	0.52 (1.00)	1.00
				0.34	0.29 (0.63)	0.37 (1.00)	1.00
				0.49	0.29 (0.77)	0.23 (1.00)	1.00
			FE	0.98	0.01 (0.90)	0.00 (0.90)	1.00
				0.99	0.00 (0.90)	0.00 (0.90)	1.00
				0.98	0.01 (0.90)	0.00 (0.90)	1.00
				0.98	0.01 (0.90)	0.00 (0.90)	1.00
			GSDS	0.17	0.25 (N/A)	0.54 (N/A)	0.96
				0.16	0.25 (N/A)	0.54 (N/A)	0.95
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$							
0.588	1.175	1.5	HUT	0.15	0.31 (0.47)	0.53 (0.99)	0.99
				0.25	0.36 (0.62)	0.38 (0.99)	1.00
				0.37	0.39 (0.76)	0.24 (1.00)	1.00
			FE	0.71	0.16 (0.73)	0.01 (0.74)	0.97
				0.78	0.10 (0.72)	0.00 (0.72)	0.97
				0.70	0.17 (0.74)	0.01 (0.75)	0.98
				0.77	0.11 (0.73)	0.00 (0.73)	0.98
			GSDS	0.13	0.23 (N/A)	0.55 (N/A)	0.91
				0.11	0.23 (N/A)	0.55 (N/A)	0.90
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$							
0.425	0.85	1.5	HUT	0.13	0.29 (0.42)	0.55 (0.96)	0.97
				0.21	0.34 (0.54)	0.43 (0.97)	0.98
				0.29	0.37 (0.66)	0.31 (0.97)	0.97
			FE	0.39	0.37 (0.75)	0.17 (0.94)	0.94
				0.48	0.28 (0.75)	0.14 (0.90)	0.91
				0.39	0.40 (0.78)	0.16 (0.95)	0.96
				0.47	0.32 (0.77)	0.14 (0.92)	0.93
			GSDS	0.08	0.25 (N/A)	0.51 (N/A)	0.85
				0.07	0.25 (N/A)	0.51 (N/A)	0.83

Table 5.11: Power performance of three two-stage enrichment designs when $(\theta_1, \theta_2, \theta_3) = (0.4, 1.2, 0.4)$. As an example, the column "$\Omega_{\{1,2\}}$ (w/other)" gives the probability that only $H_{\{1,2\}}$ is rejected, and the parenthesized value includes events where other hypotheses are also rejected. Prevalence levels are given in the table. Entries are based on $\mathcal{I}_{\max} = 25$ and $t = 1/2$. HUT parameters are $\theta^+ = 0.5$, and $k^{HUT} = 0.5, 1$ and 2, respectively. FE parameters are $(\gamma_0, \gamma_1, \gamma_2) = (0.5, 0.5, 0.5)$ in the first two rows, and (0.3,0.4,0.5) in rows three and four. Local significance levels are $\alpha/3$ in rows one and three, and (0.015,0.007,0.003) in rows two and four. The top GSDS row refers to no early rejection.

Effect Sizes			Procedure	Power Performance			
θ_0	$\theta_{\{1,2\}}$	$\theta_{\{1\}}$		Ω_0	$\Omega_{\{1,2\}}$ (w/other)	$\Omega_{\{1\}}$ (w/other)	Any
$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$							
0.55	0.6	0.4	HUT	0.40	0.09 (0.27)	0.02 (0.29)	0.66
				0.50	0.08 (0.30)	0.01 (0.30)	0.70
				0.56	0.05 (0.30)	0.00 (0.31)	0.71
			FE	0.64	0.00 (0.12)	0.00 (0.13)	0.70
				0.72	0.00 (0.12)	0.00 (0.13)	0.74
				0.63	0.01 (0.12)	0.01 (0.13)	0.70
				0.71	0.00 (0.12)	0.00 (0.13)	0.74
			GSDS	0.22	0.36 (N/A)	0.02 (N/A)	0.60
				0.19	0.23 (N/A)	0.02 (N/A)	0.55
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$							
0.5	0.6	0.4	HUT	0.38	0.07 (0.20)	0.02 (0.22)	0.58
				0.46	0.06 (0.21)	0.01 (0.22)	0.63
				0.50	0.04 (0.22)	0.00 (0.23)	0.64
			FE	0.54	0.01 (0.08)	0.01 (0.09)	0.61
				0.63	0.00 (0.08)	0.00 (0.09)	0.67
				0.53	0.01 (0.08)	0.01 (0.09)	0.62
				0.62	0.00 (0.08)	0.00 (0.09)	0.67
			GSDS	0.31	0.23 (N/A)	0.02 (N/A)	0.56
				0.26	0.20 (N/A)	0.02 (N/A)	0.50
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$							
0.6	0.8	0.4	HUT	0.48	0.05 (0.15)	0.01 (0.16)	0.75
				0.58	0.04 (0.16)	0.00 (0.16)	0.83
				0.65	0.03 (0.16)	0.00 (0.17)	0.85
			FE	0.73	0.01 (0.16)	0.00 (0.16)	0.81
				0.79	0.00 (0.16)	0.00 (0.16)	0.84
				0.72	0.01 (0.16)	0.00 (0.17)	0.82
				0.78	0.00 (0.16)	0.00 (0.16)	0.85
			GSDS	0.38	0.35 (N/A)	0.01 (N/A)	0.73
				0.33	0.27 (N/A)	0.01 (N/A)	0.67

Table 5.12: Power performance of three two-stage enrichment designs when $(\theta_1, \theta_2, \theta_3) = (0.2, 1.5, 0.2)$. As an example, the column "$\Omega_{\{1,2\}}$ (w/other)" gives the probability that only $H_{\{1,2\}}$ is rejected, and the parenthesized value includes events where other hypotheses are also rejected. Prevalence levels are given in the table. Entries are based on $\mathcal{I}_{\max} = 25$ and $t = 1/2$. HUT parameters are $\theta^+ = 0.5$, and $k^{HUT} = 0.5, 1$ and 2, respectively. FE parameters are $(\gamma_0, \gamma_1, \gamma_2) = (0.5, 0.5, 0.5)$ in the first two rows, and (0.3,0.4,0.5) in rows three and four. Local significance levels are $\alpha/3$ in rows one and three, and (0.015,0.007,0.003) in rows two and four. The top GSDS row refers to no early rejection.

Effect Sizes			Procedure	Power Performance			
θ_0	$\theta_{\{1,2\}}$	$\theta_{\{1\}}$		Ω_0	$\Omega_{\{1,2\}}$ (w/other)	$\Omega_{\{1\}}$ (w/other)	Any
$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$							
0.444	0.525	0.2	HUT	0.24	0.04 (0.09)	0.00 (0.09)	0.50
				0.31	0.03 (0.09)	0.00 (0.09)	0.54
				0.36	0.02 (0.10)	0.00 (0.10)	0.55
			FE	0.43	0.00 (0.05)	0.00 (0.06)	0.53
				0.52	0.00 (0.05)	0.00 (0.05)	0.57
				0.42	0.01 (0.05)	0.00 (0.06)	0.53
				0.51	0.00 (0.05)	0.00 (0.05)	0.57
			GSDS	0.11	0.41 (N/A)	0.00 (N/A)	0.53
				0.09	0.38 (N/A)	0.00 (N/A)	0.47
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$							
0.363	0.525	0.2	HUT	0.18	0.03 (0.07)	0.00 (0.07)	0.40
				0.23	0.03 (0.07)	0.00 (0.08)	0.45
				0.26	0.03 (0.08)	0.00 (0.08)	0.45
			FE	0.28	0.01 (0.04)	0.00 (0.04)	0.41
				0.36	0.00 (0.04)	0.00 (0.04)	0.44
				0.28	0.01 (0.04)	0.00 (0.04)	0.42
				0.35	0.01 (0.04)	0.00 (0.04)	0.45
			GSDS	0.13	0.31 (N/A)	0.01 (N/A)	0.45
				0.11	0.28 (N/A)	0.01 (N/A)	0.39
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$							
0.525	0.85	0.2	HUT	0.31	0.03 (0.06)	0.00 (0.06)	0.71
				0.40	0.02 (0.06)	0.00 (0.06)	0.84
				0.48	0.02 (0.06)	0.00 (0.06)	0.85
			FE	0.59	0.00 (0.07)	0.00 (0.07)	0.80
				0.67	0.00 (0.07)	0.00 (0.07)	0.81
				0.58	0.01 (0.07)	0.00 (0.07)	0.81
				0.66	0.00 (0.07)	0.00 (0.07)	0.83
			GSDS	0.21	0.56 (N/A)	0.00 (N/A)	0.77
				0.17	0.55 (N/A)	0.00 (N/A)	0.72

Table 5.13: Power performance of three two-stage enrichment designs when $(\theta_1, \theta_2, \theta_3) = (0.4, 0.4, 0.4)$. As an example, the column "$\Omega_{\{1,2\}}$ (w/other)" gives the probability that only $H_{\{1,2\}}$ is rejected, and the parenthesized value includes events where other hypotheses are also rejected. Prevalence levels are given in the table. Entries are based on $\mathcal{I}_{\max} = 25$ and $t = 1/2$. HUT parameters are $\theta^+ = 0.5$, and $k^{HUT} = 0.5, 1$ and 2, respectively. FE parameters are $(\gamma_0, \gamma_1, \gamma_2) = (0.5, 0.5, 0.5)$ in the first two rows, and (0.3,0.4,0.5) in rows three and four. Local significance levels are $\alpha/3$ in rows one and three, and (0.015,0.007,0.003) in rows two and four. The top GSDS row refers to no early rejection.

Effect Sizes			Procedure	Power Performance			
θ_0	$\theta_{\{1,2\}}$	$\theta_{\{1\}}$		Ω_0	$\Omega_{\{1,2\}}$ (w/other)	$\Omega_{\{1\}}$ (w/other)	Any
$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$							
0.4	0.4	0.4	HUT	0.20	0.05 (0.17)	0.09 (0.26)	0.37
				0.27	0.05 (0.21)	0.06 (0.27)	0.41
				0.32	0.04 (0.24)	0.03 (0.28)	0.42
			FE	0.35	0.01 (0.08)	0.02 (0.11)	0.40
				0.43	0.00 (0.09)	0.01 (0.10)	0.45
				0.34	0.01 (0.08)	0.02 (0.11)	0.41
				0.42	0.00 (0.09)	0.01 (0.10)	0.45
			GSDS	0.14	0.12 (N/A)	0.10 (N/A)	0.36
				0.12	0.09 (N/A)	0.09 (N/A)	0.31
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$							
0.4	0.4	0.4	HUT	0.21	0.04 (0.12)	0.07 (0.19)	0.34
				0.29	0.04 (0.15)	0.05 (0.21)	0.39
				0.33	0.03 (0.17)	0.03 (0.20)	0.41
			FE	0.35	0.01 (0.05)	0.02 (0.07)	0.40
				0.43	0.00 (0.06)	0.01 (0.07)	0.46
				0.34	0.01 (0.05)	0.02 (0.08)	0.40
			E	0.42	0.00 (0.06)	0.01 (0.07)	0.46
			GSDS	0.24	0.07 (N/A)	0.06 (N/A)	0.37
				0.20	0.09 (N/A)	0.06 (N/A)	0.32
$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$							
0.4	0.4	0.4	HUT	0.22	0.03 (0.09)	0.06 (0.15)	0.34
				0.29	0.03 (0.11)	0.04 (0.15)	0.40
				0.33	0.02 (0.12)	0.03 (0.15)	0.41
			FE	0.35	0.01 (0.10)	0.02 (0.13)	0.40
				0.43	0.01 (0.11)	0.01 (0.12)	0.46
				0.34	0.02 (0.10)	0.02 (0.13)	0.40
				0.42	0.01 (0.11)	0.01 (0.12)	0.46
			GSDS	0.23	0.09 (N/A)	0.04 (N/A)	0.37
				0.19	0.10 (N/A)	0.04 (N/A)	0.31

Table 5.14: Power performance of three two-stage enrichment designs when $(\theta_1, \theta_2, \theta_3) = (0.8, 0.8, 0.8)$. As an example, the column "$\Omega_{\{1,2\}}$ (w/other)" gives the probability that only $H_{\{1,2\}}$ is rejected, and the parenthesized value includes events where other hypotheses are also rejected. Prevalence levels are given in the table. Entries are based on $\mathcal{I}_{\max} = 25$ and $t = 1/2$. HUT parameters are $\theta^+ = 0.5$, and $k^{HUT} = 0.5, 1$ and 2, respectively. FE parameters are $(\gamma_0, \gamma_1, \gamma_2) = (0.5, 0.5, 0.5)$ in the first two rows, and (0.3,0.4,0.5) in rows three and four. Local significance levels are $\alpha/3$ in rows one and three, and (0.015,0.007,0.003) in rows two and four. The top GSDS row refers to no early rejection.

Effect Sizes			Procedure	Power Performance			
θ_0	$\theta_{\{1,2\}}$	$\theta_{\{1\}}$		Ω_0	$\Omega_{\{1,2\}}$ (w/other)	$\Omega_{\{1\}}$ (w/other)	Any
	$f_{\{1,2\}} = 3/4, f_{\{1\}} = 9/16$						
0.8	0.8	0.8	HUT	0.71	0.11 (0.73)	0.10 (0.84)	0.94
				0.82	0.08 (0.80)	0.04 (0.84)	0.95
				0.89	0.05 (0.83)	0.02 (0.85)	0.96
			FE	0.95	0.00 (0.40)	0.00 (0.40)	0.95
				0.97	0.00 (0.40)	0.00 (0.40)	0.97
				0.94	0.00 (0.40)	0.00 (0.40)	0.95
				0.96	0.00 (0.40)	0.00 (0.40)	0.97
			GSDS	0.56	0.17 (N/A)	0.08 (N/A)	0.81
				0.52	0.17 (N/A)	0.08 (N/A)	0.76
	$f_{\{1,2\}} = 1/2, f_{\{1\}} = 3/8$						
0.8	0.8	0.8	HUT	0.77	0.05 (0.59)	0.09 (0.68)	0.91
				0.87	0.04 (0.64)	0.03 (0.68)	0.95
				0.92	0.02 (0.66)	0.01 (0.68)	0.96
			FE	0.95	0.00 (0.27)	0.00 (0.28)	0.95
				0.97	0.00 (0.27)	0.00 (0.28)	0.97
				0.94	0.00 (0.27)	0.00 (0.28)	0.95
				0.96	0.00 (0.27)	0.00 (0.28)	0.97
			GSDS	0.78	0.06 (N/A)	0.04 (N/A)	0.88
				0.74	0.06 (N/A)	0.04 (N/A)	0.84
	$f_{\{1,2\}} = 1/2, f_{\{1\}} = 1/4$						
0.8	0.8	0.8	HUT	0.79	0.04 (0.47)	0.05 (0.52)	0.91
				0.88	0.03 (0.49)	0.02 (0.51)	0.95
				0.92	0.02 (0.49)	0.01 (0.51)	0.96
			FE	0.95	0.00 (0.48)	0.00 (0.49)	0.95
				0.97	0.00 (0.49)	0.00 (0.49)	0.97
				0.94	0.00 (0.48)	0.00 (0.50)	0.95
				0.96	0.00 (0.49)	0.00 (0.50)	0.97
			GSDS	0.78	0.08 (N/A)	0.02 (N/A)	0.88
				0.74	0.08 (N/A)	0.02 (N/A)	0.83

GSDS arrives at the correct conclusion (pass treatment for Ω_1 only) with greater probability than FE (always) and HUT (usually). HUT performs well in this scenario if $k^{HUT} = 0.5$, as this penalizes false positives more so than false negatives. The weakness of FE, discussed in Chapter 3, is present in this setting. This is more obvious when prevalence levels are moderate to high (first two sections of the two tables); we can see that FE tends to reach a positive conclusion for Ω_0 rather than identify strong effects in subgroups. This can be countered by making γ_0^{FE} small, and we consider the values 0.3 and 0.5, though lower values might be useful especially if prevalence of Ω_1 and Ω_2 is high.

By design, GSDS can only make inference on a single population at the end of the trial. For example, it is not possible for GSDS to reject both $H_{\{1\}}$ and $H_{\{1,2\}}$, something both FE and HUT can easily do. This represents a sort of trade-off that must be carefully considered by a trial sponsor. GSDS will reach the correct conclusion with higher probability in comparison to the two other designs, and has smaller expected sample size if early stopping for rejection is allowed. On the other hand, there is no inbuilt mechanism to "rescue" the trial at the final analysis if results are not as good as hoped. In particular, there is no unused Type I error that may be spent on examining smaller populations than the one selected at the first interim analysis. Whether the design can be modified to allow for testing of smaller populations while maintaining its simplicity is a topic for further research. We discuss some possible remedies in Section 6.2.

If the prevalence levels of Ω_1 and Ω_2 are high, it may appear as if a positive result in $\Omega_{\{1,2\}}$ would also be desirable. This occurs for example when $(\theta_1, \theta_2, \theta_3) = (1.5, 0.2, 0)$ and $(f_{01}, f_{02}, f_{03}) = (9/16, 3/16, 1/4)$, see Table 5.10. In this case, $\theta_{\{1,2\}} = 1.175$, which exceeds the effect size considered to be of clinical significance.

However, we believe that the correct conclusion should still exclude Ω_2, as the high effect size in $\Omega_{\{1,2\}}$ is primarily a result of $\theta_1 = 1.5$ and the high prevalence of Ω_1 in $\Omega_{\{1,2\}}$ ($f_{\{1,2\},1} = 0.75$).

Overall, while GSDS reaches a positive conclusion (i.e. reject any null hypothesis) with lower probability than either FE or HUT, it does perform better in terms of rejecting only the hypotheses that *should* be rejected. Power performance of HUT and FE is also fairly dependent on procedure-specific parameters (local significance levels and enrichment parameters for FE; θ^+ and k^{HUT} for HUT), while GSDS only requires specification of the two spending functions α_L^* and α_U^*. These functions are naturally defined as proportional to the interim analysis timing. As a result, GSDS is reasonably simple to implement and its performance does not overly rely on the specification of a number of parameters. This is desirable, as procedure-specific parameters can substantially influence the probability of a positive result depending on the true effect size pattern and prevalence levels.

Effect of No Nesting Pattern: We see in Tables 5.11 and 5.12 power performance when there is no nesting pattern. Here, the effect size is only large enough for population Ω_2, while effects in Ω_1 and Ω_3 are not worth detecting. None of the designs under consideration are particularly useful for this scenario. GSDS could be applied, but DR-I (see Section 5.2) would be more appropriate than the rule used in this example (DR-II, which is used because of the perceived nesting structure). Depending on prevalence levels, the most desirable conclusion may be rejection of $H_{\{1,2\}}$, in particular when $(f_{01}, f_{02}, f_{03}) = (1/4, 1/4, 1/2)$, in which case $\theta_{\{1,2\}} = 0.8$ in Table 5.11 and 0.85 in Table 5.12. The GSDS procedure achieves highest rejection probabilities for $H_{\{1,2\}}$ in all cases that are considered. As expected though, the overall performance of the three designs is uninspiring.

Effect of Homogeneity Pattern: Tables 5.13 and 5.14 show performance in the case of equal effect size across all subgroups. In this type of scenario, the correct conclusion is to reject Ω_0, though secondary findings involving (possibly composite) subgroups are also valid. When all $\theta_j = 0.4$, we expect that all designs are underpowered, and this is confirmed in Table 5.13. In this case, it can be argued that the "best" conclusion is simply a negative trial, as effect sizes are not close to a clinically meaningful difference. The FE and HUT designs achieve higher rejection probabilities for Ω_0 than GSDS does.

When all $\theta_j = 0.8$, it is desirable that the trial ends with a positive finding, preferably for Ω_0. Inspecting Table 5.14, we again see that FE achieves this with high probability, dominating the other procedures in the cases considered. This finding is with the caveat that by design, FE tends to have high power for H_0, even when there is heterogeneity among the subgroups. HUT can perform well in the homogeneity setting, though power is quite dependant on the value chosen for k^{HUT}. Our results indicate that $k^{HUT} = 2$ leads to good performance (using $k^{HUT} = 2$ means false negatives are penalized heavily, which makes subgroup elimination unlikely at the interim analysis). We note that GSDS is dominated in this setting, but comes closest to the other procedures when prevalence levels are low rather than high.

Impact of Early Rejection in GSDS: In Tables 5.9–5.14 we have included power performance for GSDS with $\alpha_U^*(t) = t\alpha$ and $\alpha_U^*(t) = 0$, i.e. both early rejection and no early rejection. In all cases, we observe that power is slightly lower when early rejection is allowed. The difference is essentially negligible when a favorable nesting pattern is present, but more pronounced when treatment effects are equal across the subgroups. Hence there is a trade-off to consider for a trial sponsor; if early findings warrant terminating the trial with a positive finding, early rejection can mean substantial savings in time and financial resources. On the other hand, it may be necessary to budget for a slightly larger maximum sample size due to the fact that power tends to be lower when early rejection is possible.

Impact of Interim Timing: Figure 5.5 shows the impact of interim timing on power performance for GSDS. (Timing properties for HUT and FE were considered in Chapter 3.) We see that when a favorable nesting pattern is present, the procedure benefits from a late interim analysis. Recall that lower limits are $l_1\sqrt{\Delta_{1j}}$, $j \in \mathcal{P}$, and hence they increase proportionally to the square root of accumulated information. When effect sizes are small or non-existent, a longer first stage hence decreases the probability that an ineffective population be passed on to stage two. However, for the homogeneity pattern $(\theta_1, \theta_2, \theta_3) = (0.8, 0.8, 0.8)$ (red line in Figure 5.5), we see the opposite effect.

When a favorable nesting pattern is present, Wang et al. (2009) report that the earlier an interim analysis is performed, the higher the power performance for all subgroups and here we see the opposite effect. This is in contrast to GSDS, and the difference can be explained by examining the procedure used to eliminate populations; GSDS employs a sort of “bottom-up” approach, whereby the population believed to be least responsive is first up for elimination. In particular, the decision on whether or not to eliminate a particular subgroup is based purely on results from that subgroup. Wang et al. (2009), on the other hand, employ a “top-down” approach, where the overall results are checked first. If these are not strong enough, Ω_3 is eliminated, and results involving only Ω_1 and Ω_2 are considered, and so forth. Hence, a long first stage implies that the non-centrality parameter for Z_{10} (first stage, overall population) is large relative to when the first stage is shorter.

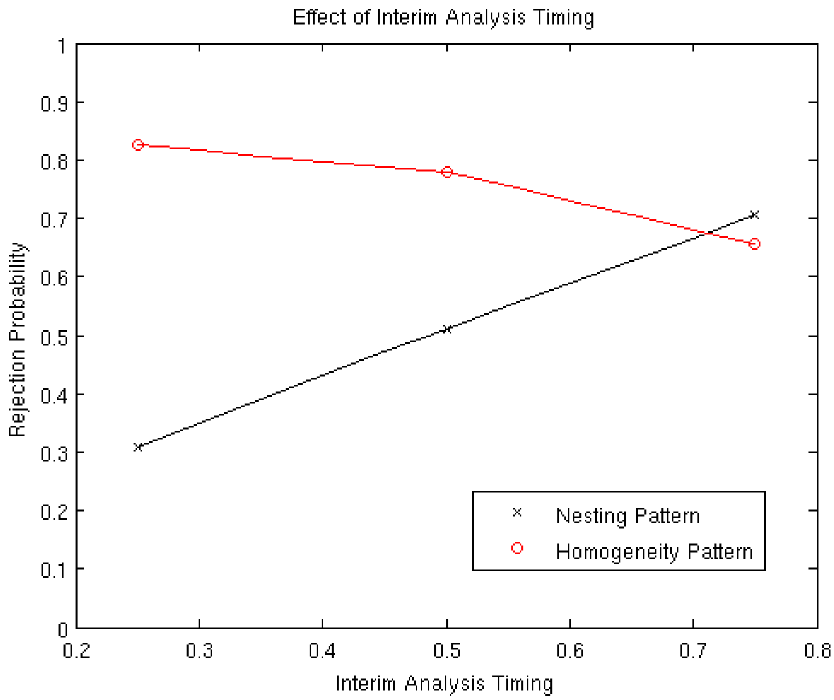


Figure 5.5: Rejection probabilities for the GSDS procedure over t values 0.25, 0.5 and 0.75. For the nesting pattern, power performance is for rejection of $\Omega_{\{1\}}$ only. For the homogeneity pattern, power performance is for rejection of Ω_0.

This makes detection of heterogeneity more difficult, and the incorrect decision to proceed with Ω_0 is probable. As discussed in the previous paragraph, GSDS is well equipped to detect this favorable nesting pattern, and does so with greater probability as the first stage is lengthened.

Chapter 6

Discussion

6.1 Conclusions

We have considered the problem of designing and analyzing clinical trials that prospectively incorporate analysis of subgroup-specific effects. This issue is very prevalent in modern day clinical trials, and several motivating examples were discussed in Chapter 1. We face many challenges with subgroup analysis, including lack of study power, issues with multiplicity, as well as interpretability and generalizability of study results. The small size of many subgroups can mean that prohibitively large sample sizes are needed to guarantee desired power, and hence trials must be carefully designed to detect effects restricted to certain subgroups. In addition, a study that entertains multiple questions must provably bound the probability that any false positive decision is made. This can become quite complicated when there are many subgroups to chose from. In Section 1.2.2, we discussed in detail these issues and others that should be taken into consideration when conducting subgroup analysis.

In Chapter 3, we considered the case where there is one subgroup of interest. In this scenario, it is preferred to detect an overall effect if one is present; if not, a subgroup effect should be detected with high probability. By prospectively specifying the subgroup of interest, it is possible to design trials that allow confirmatory evaluation of treatment effect in this subgroup, while protecting FWER strongly at the desired level and also maintaining the integrity of the trial. Several novel designs were proposed, and others were adapted from the literature for comparison. We saw that adaptive designs that incorporate an interim analysis can do particularly well in this setting, specifically if only the subgroup is responsive. In contrast, traditional designs such as a fixed Hochberg-Simes procedure or a multiplicity adjusted O'Brien and Fleming group sequential trial are not suited to detect heterogeneity. Procedures that showed particular promise allow adaptive enrichment at interim (e.g. FE or HUT), whereby all resources can be focused on the subgroup for the second stage. In Chapter 4, the procedures AFP, FE and HUT were extended to allow consideration of any number of subgroups.

As part of the analysis conducted in Chapter 3, we also investigated the effect of subgroup prevalence and interim analysis timing on the power performance, as well as the quality of interim analysis decisions. As expected, power for the subgroup was observed to increase with its prevalence, while power to detect an overall positive effect also relies on the effect size in the subgroup complement. There is no obvious "correct" time for an interim analysis, though our conclusions indicate that t between 1/4 and 1/2 is appropriate. For the scenarios that we analyzed, there is little difference in power performance as t varies, though obviously the interim analysis is more robust as t increases.

An important observation made in our research is that the decision on whether or not to enrich the subgroup of interest should be based on (or at least incorporate) the analysis of treatment effect in the subgroup complement. It is possible to see a strong overall effect, while at the same time the subgroup complement is completely nonresponsive. In such cases, the observed positive overall effect is mainly driven by very strong results in the subgroup of interest. Basing interim decisions only on the overall statistic Z_{10} can result in designs that are biased towards positive results for the complete populations, and do not detect heterogeneity except under very favorable circumstances. FE and CP are the two proposed procedures to which this applies. However, this weakness is less obvious when the interim analysis is conducted early, and the prevalence of the subgroup of interest is relatively low.

In Chapter 5 we proposed a confirmatory adaptive multi-stage group sequential design (GSDS) that only eliminates populations that show early signs of being non-responsive. At the end of the first stage, results from each prospectively specified subgroup are examined, and those falling below a certain threshold are discarded for the remainder of the trial. The remaining populations are then pooled to create a single overall population of interest that is tested at subsequent interim analyses. The pooling approach is novel, and has certain strengths as well as drawbacks. For instance, combining the remaining populations is a simple way to deal with issues of multiplicity, which can be very challenging when conducting a group sequential trial. One of the strengths of group sequential trials is the high probability that the trial can be terminated before the final analysis with a positive conclusion. The GSDS design thus combines this strength with the possibility to detect subgroup heterogeneity while avoiding issues that can arise when more flexible designs are used (Burman and Sonesson, 2006). A weakness of the pooling approach is that once a stopping boundary is crossed, no further testing is allowed (all α has been spent). This can be problematic if the trial was negative, but the combined population consists of multiple prospectively specified subgroups, some of which may actually be responsive to treatment.

The GSDS procedure was compared to the FE and HUT procedures in terms of power performance when there were three subgroups of interest, and no procedure was dominated over all cases considered. We found that when a favorable nesting pattern is present, the GSDS procedure reaches the *correct conclusion only* with greater probability than FE and HUT, while the latter two procedures are more powerful when treatment effect is homogeneous across all populations. GSDS does have the advantage of easily incorporating early stopping for rejection, and hence has a smaller expected sample size than the other two designs.

In conclusion, we have presented several confirmatory adaptive clinical trial designs that are specifically tailored to allow analysis of subgroups. When effects are confined to a subgroups, procedures that incorporate at least one interim analysis were seen to achieve substantial gains in power over non-adaptive designs. In our examples, no adaptive design outperformed another for all configurations, nor was any design outperformed for all configurations. The FE procedure, while powerful overall, can struggle to detect heterogeneity if subgroup prevalence is high, or if the interim analysis is unsuitably timed. The HUT procedure performs well in a variety of scenarios, but power performance can be highly dependent on procedure-specific parameters such as k^{HUT} and θ^{+}. Further, while using empirical data weights is desirable to increase efficiency, protection of Type I error is not guaranteed and hence their use may be unacceptable in the eyes of regulatory authorities. If empirical data weights are not used, then equally informative observations are weighted unequally at the final analysis. The GSDS procedure, though adaptive in nature, is essentially a fixed design that requires minimal specification of procedure-specific parameters. Once spending functions are chosen, critical boundaries can be obtained in a straightforward manner, and the trial then has a clear and unambiguous decision path. Further, unlike procedures considered in Chapters 3 and 4, GSDS can incorporate more than one interim analysis, any of which can result in trial termination due either to rejection or acceptance. We believe that this flexibility, coupled with its relative simplicity, makes the GSDS design a suitable procedure of choice for confirmatory trials where subgroup analysis is desired.

6.2 Future Research

In Chapter 3 we observed that the FE procedure can struggle to detect heterogeneity, instead attributing positive results to an overall effect. Hence, FE power for H_0 is usually higher than power for a subgroup, unless the subgroup is relatively small or if the interim analysis is conducted early. In the single-subgroup setting, we can counter this using an approach similar to that given by Brannath et al. (2009). Let $\gamma_2 \in [0, 1]$, and let $C_{\gamma_2} = F^{-1}(1 - \gamma_2)$ where F is the cumulative distribution function of the test statistic Z (see notation in Section 3.2.2). Now the first stage analysis FE procedure can be stated as follows:

1. If $Z_{10} < C_{\gamma_0}$ and $Z_{11} < C_{\gamma_1}$, then stop the study for futility.
2. If $Z_{10} < C_{\gamma_0}$ and $Z_{11} \geq C_{\gamma_1}$, then proceed to stage two and enrich the subgroup Ω_1.
3. If $Z_{10} \geq C_{\gamma_0}$ and $Z_{11} \geq C_{\gamma_1}$, but $Z_{12} < C_{\gamma_2}$, then also enrich to Ω_1 for stage two.
4. If $Z_{10} \geq C_{\gamma_0}$, $Z_{11} \geq C_{\gamma_1}$, and $Z_{12} \geq C_{\gamma_2}$, then proceed to stage two with the full population.
5. If $Z_{10} \geq C_{\gamma_0}$ and $Z_{11} < C_{\gamma_1}$, then proceed to stage two with the full population.

Steps 3–5 contain the modified decision process that is proposed, where we require some consistency of findings in the subgroup complement in order to allow the procedure to go forward with the full population. A new value for $\tilde{\alpha}_1$ will need to be computed, and when there is only one subgroup of interest the analysis involved is manageable. Incorporating steps 3–5 will affect power for H_0 negatively, even when there is no heterogeneity among subgroups. However, the purpose of a method such as FE is to detect subgroup-specific effects, and the proposed changes will help accomplish this.

The HUT and HPP designs presented in Chapter 3 rely on Bayesian computational methods when conducting the interim analysis. A Phase III trial needs to provably protect FWER strongly at level α, and this can be problematic to show for a pure Bayesian design. However, the relatively strong performance of HUT and HPP suggests that more effort should be put into the development of hybrid Bayesian designs for confirmatory Phase III trials. In particular, exchangeability was a key assumption in our setup, which is appropriate when there is no natural ordering of subgroups. When this is not the case, however, a general method that accounts for patient heterogeneity is still needed. Wathen et al. (2008) have proposed a class of model-based Bayesian designs for Phase II trials, specifically tailored to handle differing prior beliefs about particular subgroups, but to our knowledge no such designs exist yet for use in Phase III trials.

The GSDS procedure, proposed in Chapter 5, presents a new approach to combine "traditional" group sequential trials with the problem of detecting heterogeneity among subgroups. We presented the procedure in terms of the efficient score statistic, relying on the normal approximation which holds for a variety of endpoints. However, when endpoints are binary and sample sizes are small, or success probabilities p are small, this approximation is inaccurate. Moreover, observed information depends on p, and in our examples we chose the conservative estimate by using $\bar{p} = \frac{1}{2}$. To remedy this, we could use exact methods for Bernoulli data, where the number of successes for each population at a given stage are distributed as a binomial random variable. For the first stage, the conditional distribution of the cumulative number of successes for retained populations Ω_j, $j \in \mathcal{P}^*$, can be obtained with convolution. Its exact form will depend on the decision rule used at the first analysis (nested or not).

We list other extensions/modifications of GSDS that might be of interest:

1. Include a prospectively specified option to delay population elimination. That is, stage $1 < k^* < K$ is predetermined as the stage at which some populations may be dropped. Prior to stage k^*, the trial will run as a one-population trial, testing only H_0. If stage k^* is reached without rejection or acceptance of H_0, each population can be examined individually in the same fashion as described in Section 5.2. Hence, after stage k^* the trial would behave exactly as the original GSDS design.

2. Another option is to allow elimination of populations in more than one stage. For example, say we have three populations Ω_1, Ω_2 and Ω_3 and a three stage trial is planned. At the first interim analysis, Ω_3 is dropped, but the other two populations look promising. If, at the end of the second stage, only Ω_1 looks responsive, it is desirable to have the option to drop Ω_2 as well. The current formulation of GSDS does not account for this, and the added complexity may make the resulting analysis more computationally intensive. Approximate boundaries could still be easily obtained by use of simulation.

3. As currently stated, if the GSDS accepts $H_{\mathcal{P}^*}$ at a given stage, no more testing may take place. However, if $\Omega_{\mathcal{P}^*}$ was composed of multiple populations and there is reason to believe some were incorrectly chosen after the first analysis, we would like for there to be some flexibility to test smaller populations, as a sort of "fallback" option. In such cases, it may be possible to use up unspent α and test hypotheses corresponding to the remaining pop-

ulations using a simple multiplicity adjustment. Test statistics of interest can be computed using a combination rule such as the weighted inverse normal method, where combination weights are proportional to the originally planned information increments. Then the closed testing principle would be applied, taking all original hypotheses into consideration (even those corresponding to dropped populations). If the procedure terminated at stage T, then the remaining unspent α can be expressed in terms of the upper spending function $\alpha^*_U(\cdot)$, as $\sum_{i=T+1}^{K} \alpha^*_U(i) - \alpha^*_U(T)$. Note that if acceptance occurred at the final analysis, all α has been spent. P-values could be adjusted using, for example, Hochberg's method, and tests carried out using the remaining α as the effective FWER. Further investigation is required to assess the operating characteristics (e.g. possible Type I error inflation) of this type of "rescue" testing.

4. It might be of interest to not force pooling of chosen populations, and to test these individually. In this case, we are essentially running a number of parallel and independent group sequential trials, and the GSDS procedure would no longer apply. Sample size requirements would certainly be cause for concern, and while obtaining integral representations of stage-wise probabilities (as we did for GSDS) is straightforward, accounting for multiplicity is not.

5. The procedure is developed under the assumption that observation variance σ^2 is known. If σ^2 is not known, then observed information is not known; it's estimate depends on the estimate of σ. In this case tests would be specified in terms of t-distributed test statistics, see for example (Jennison and Turnbull, 2000, Ch. 4.4).

BIBLIOGRAPHY

Alosh, M. and Huque, M. F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*, 28:3–23.

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

Atkinson, A., Colburn, W., DeGruttola, V., DeMets, D., Downing, G., Hoth, D., Oates, J., Peck, C., Schooley, R., Spilker, B., et al. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95.

Barker, A., Sigman, C., Kelloff, G., Hylton, N., Berry, D., and Esserman, L. (2009). I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, 18(14):1833–1848.

Bauer, P. and Köhne, K. (1994). Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics*, 50(4):1029–1041.

Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* John Wiley & Sons, Inc.

Begley, S. (2010, January 29). The Depressing News About Antidepressants. *Newsweek.* `http://www.newsweek.com/2010/01/28/the-depressing-news-about-antidepressants.html`.

Begley, S. (2010, September 07). Curing Cancer. *Newsweek*. http://www.newsweek.com/2010/09/07/what-we-can-learn-from-curable-cancers.html.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697.

Berry, S. M., Carlin, B. P., Lee, J., and Müller, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL.

Bollag, G., Hirth, P., Tsai, J., Zhang, J., Ibrahim, P. N., Cho, H., Spevak, W., Zhang, C., Zhang, Y., Habets, G., et al. (2010). Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAD-mutant melanoma. *Nature*, 467:596–599.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive Combination Tests. *Journal of the American Statistical Association*, 97(457):236–244.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, 28:1445–1463.

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28:586–604.

Bretz, F., Schmidli, H., König, F., Racine, A., and Maurer, W. (2006). Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts. *Biometrical Journal*, 48(4):623–634.

Burman, C.-F. and Sonesson, C. (2006). Are Flexible Designs Sound. *Biometrics*, 62:664–683.

Chang, M. (2010). *Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms, and Case Studies.* Chapman & Hall/CRC, Boca Raton, FL.

Cui, L., Hung, H., and Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 55(3):853–857.

Davies, H., Bignell, G., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M., Bottomley, W., et al. (2002). Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954.

Dixon, D. O. and Simon, R. (1991). Bayesian subset analysis. *Biometrics*, 47(3):871–881.

Dmitrienko, A., Bretz, F., Westfall, P. H., Troendle, J., Wiens, B. L., Tamhane, A. C., and Hsu, J. C. (2010). Multiple Testing Methodology. In Dmitrienko, A., Tamhane, A. C., and Bretz, F., editors, *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

European Agency for the Evaluation of Medicinal Products (2002). Points to Consider on Multiplicity Issues in Clinical Trials. `http://www.tga.gov.au/docs/pdf/euguide/ewp/090899en.pdf`.

Flaherty, K., Puzanov, I., Kim, K., Ribas, A., McArthur, G., Sosman, J., O'Dwyer, P., Lee, R., Grippo, J., Nolop, K., et al. (2010). Inhibition of mutated, activated BRAF in metastatic melanoma. *New England Journal of Medicine*, 363(9):809–819.

Fleming, T. (2006). Standard versus adaptive monitoring procedures: a commentary. *Statistics in Medicine*, 25(19):3305–3312.

Fournier, J., DeRubeis, R., Hollon, S., Dimidjian, S., Amsterdam, J., Shelton, R., and Fawcett, J. (2010). Antidepressant drug effects and depression severity:

a patient-level meta-analysis. *Journal of the American Medical Association*, 303(1):47.

Freidlin, B., McShane, L., and Korn, E. (2010). Randomized clinical trials with biomarkers: design issues. *JNCI Journal of the National Cancer Institute*, 102(3):152–160.

Freidlin, B. and Simon, R. (2005a). Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing a Gene Expression Signature for Sensitive Patients. *Clinical Cancer Research*, 11:7872–7878.

Freidlin, B. and Simon, R. (2005b). Evaluation of randomized discontinuation design. *Journal of Clinical Oncology*, 23(22):5094.

Gallo, P. (2006). Operational challenges in adaptive design implementation. *Pharmaceutical Statistics*, 5(2):119–124.

Genz, A. and Bretz, F. (2002). Methods for the Computation of Multivariate t-Probabilities. *Journal of Computational and Graphical Statistics*, 11:950–971.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158.

Harmon, A. (2010, February 21). A Roller Coaster Chase for a Cure. *The New York Times*. `http://www.nytimes.com/2010/02/22/health/research/22trial.html`.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386.

Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika*, 76(3):624–625.

Hung, H. M. J., O'Neill, R. T., Wang, S.-J., and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal*, 48(4):565–573.

Huque, M. F. and Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference*, 138:321–335.

Huque, M. F. and Röhmel, J. (2010). Multiplicity Problems in Clinical Trials: A Regulatory Perspective. In Dmitrienko, A., Tamhane, A. C., and Bretz, F., editors, *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*, 2:349–360.

Jänne, P. and Johnson, B. (2006). Effect of Epidermal Growth Factor Receptor Tyrosine Kinase Domain Mutations on the Outcome of Patients with Non–Small Cell Lung Cancer Treated with Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors. *Clinical Cancer Research*, 12(14):4416s–4420s.

Jennison, C. and Turnbull, B. (2006a). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Opportunities and limitations. *Biometrical Journal*, 48(4):650–655.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22:971–993.

Jennison, C. and Turnbull, B. W. (2006b). Adaptive and nonadaptive group sequential tests. *Biometrika*, 93(1):1–21.

Jennison, C. and Turnbull, B. W. (2007). Adaptive Seamless Designs: Selection and Prospective Testing of Hypotheses. *Journal of Biopharmaceutical Statistics*, 17(6):1135–1161.

Jiang, W., Freidlin, B., and Simon, R. (2007). Biomarker-Adaptive Treshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect. *Journal of the National Cancer Institute*, 99(13):1036–1043.

Kirsch, I., Deacon, B., Huedo-Medina, T., Scoboria, A., Moore, T., and Johnson, B. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Medicine*, 5(2):260–268.

Koenig, F., Brannath, W., Bretz, F., and Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics In Medicine*, 27:1612–1625.

Kolata, G. (2010, April 19). Cancer Fight: Unclear Tests for New Drug. *The New York Times*. `http://www.nytimes.com/2010/04/20/health/research/20cancer.html?emc=eta1`.

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

Lan, K. and DeMets, D. (1989). Group sequential procedures: calendar versus information time. *Statistics in Medicine*, 8(10):1191–1198.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Liu, Q., Proschan, M. A., and Pledger, G. W. (2002). A Unified Theory of Two-Stage Adaptive Designs. *Journal of the American Statistical Association*, 97(460):1034–1041.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., and Krams, M. (2006). Adaptive seamless phase II/III designs–background, operational aspects, and examples. *Drug Information Journal*, 40(4):463–473.

Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine*, 28:329–339.

Mandrekar, S. J. and Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: one size does not fit all. *Journal of Biopharmaceutical Statistics*, 19:530–542.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special references to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Maurer, W., Hothorn, L., and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der chemisch-pharmazeutischen Industrie*, 6:3–18.

Maynes, D., Hunt, K., Pockaj, B., Gray, R., Tong, W., Bothe, M., Dueck, A., and Northfelt, D. (2010). Are HER2-positive breast cancer and BRCA mutation-associated breast cancer mutually exclusive diseases? Evidence from the Mayo Clinic Arizona Cohort. In *ASCO Meeting Abstracts*, volume 28, pages e21075, `http://www.asco.org/ASCOv2/Meetings/Abstracts?&vmview=abst_detail_view&confID=74&abstractID=51692`.

Mehta, C. and Patel, N. (2006). Adaptive, group sequential and decision theoretic approaches to sample size determination. *Statistics in medicine*, 25(19):3250–3269.

Mook, S., Veer, L., Emiel, J., Piccart-Gebhart, M., and Cardoso, F. (2007). Individualization of Therapy Using Mammaprint®: from Development to the MINDACT Trial. *Cancer Genomics-Proteomics*, 4(3):147–155.

Moyé, L. A. (1998). P-Value Interpretation and Alpha Allocation in Clinical Trials. *Annals of Epidemiology*, 8:351–357.

Moyé, L. A. (2000). Alpha calculus in clinical trials: considerations and commentary for the new millenium. *Statistics in Medicine*, 19:767–779.

Moyé, L. A. and Deswal, A. (2001). Trials within Trials: Confirmatory Subgroup Analyses in Controlled Clinical Experiments. *Controlled clinical trials*, 22(6):605–619.

Müller, H.-H. and Schäfer, H. (2001). Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. *Biometrics*, 57:886–891.

Nahta, R. and Esteva, F. J. (2003). HER-2-Targeted Therapy: Lessons Learned and Future Directions. *Clinical Cancer Research*, 9:5078–5084.

O'Brien, P. and Fleming, T. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556.

Paik, S., Kim, C., Jeong, J., Geyer, C., Romond, E., Mejia-Mejia, O., Mamounas, E., Wickerham, D., Costantino, J., and Wolmark, N. (2007). Benefit from adjuvant trastuzumab may not be confined to patients with IHC 3+ and/or FISH-positive tumors: Central testing results from NSABP B-31. In *ASCO Annual Meeting Proceedings*, volume 25, page 511. http://meeting.ascopubs.org/cgi/content/abstract/25/18_suppl/511.

Perez, E., Romond, E., Suman, V., Jeong, J., Davidson, N., Geyer, C., Martino, S., Mamounas, E., Kauffman, P., Wolmark, N., et al. (2007). Updated results of the combined analysis of NCCTG N9831 and NSABP B-31 adjuvant chemotherapy with/without trastuzumab in patients with HER2-positive breast cancer. In *ASCO Annual Meeting Proceedings*, volume 25, page 512. http://meeting.ascopubs.org/cgi/content/abstract/25/18_suppl/512.

Perez, E., Suman, V., Davidson, N., Martino, S., Kaufman, P., Lingle, W., Flynn, P., Ingle, J., Visscher, D., and Jenkins, R. (2006). HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. *Journal of Clinical Oncology*, 24(19):3032–3038.

Petrucelli, N., Daly, M., Culver, J., Levy-Lahad, E., and Feldman, G. (1997). BRCA1 and BRCA2 hereditary breast/ovarian cancer. *GeneReviews at GeneTests: Medical Genetics Information Resource (database online). Copyright, University of Washington, Seattle*, 2007.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:2917–2930.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed Extension of Studies Based on Conditional Power. *Biometrics*, 51:1315–1324.

Romond, E., Perez, E., Bryant, J., Suman, V., Geyer Jr, C., Davidson, N., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P., et al. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine*, 353(16):1673–1684.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2):220–238.

Sargent, D. J., Conley, B. A., Allegra, C., and Collette, L. (2005). Clinical Trial Designs for Predictive Marker Validation in Cancer Treatment Trials. *Journal of Clinical Oncology*, 23(9):2020–2027.

Schmidli, H., Bretz, F., Racine, A., and Maurer, W. (2006). Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: Applications and Practical Considerations. *Biometrical Journal*, 48(4):635–643.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Simon, R. (2002). Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine*, 21:2909–2916.

Simon, R. (2005). Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. *Journal of Clinical Oncology*, 23(29):1–10.

Simon, R., Dixon, D., and Freidlin, B. (1995). A Bayesian model for evaluating specificity of treatment effects in clinical trials. In Thall, P., editor, *Recent Advances in Clinical Trial Design and Analysis*. Kluwer Academic Publications: Norwell, MA.

Simon, R. and Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, 10(20):6759.

Simon, R. and Wang, S.-J. (2006). Use of genomic signatures in therapeutics development in oncology and other diseases. *The Pharmacogenomics Journal*, 6(3):166–173.

Song, Y. and Chi, G. Y. H. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in medicine*, 26(19):3535–3549.

Speed, T. (2010, December). Terence's Stuff: Enduring Values. *IMS Bulletin*.

Spiegelhalter, D., Abrams, K., and Myles, J. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Wiley.

Stallard, N. and Facey, K. (1996). Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics*, 6(3):361–373.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*, 22(5):689–703.

Strassburger, K., Bretz, F., and Hochberg, Y. (2004). Compatible confidence intervals for intersection union tests involving two hypotheses. *Institute of Mathematical Sciences*, 47:129–142.

Temple, R. J. (1994). Special study designs: early escape, enrichment, studies in non-responders. *Communications in Statistics-Theory and Methods*, 23(2):499–531.

Temple, R. J. (2005). Enrichment Designs: Efficiency in Development of Cancer Treatments. *Journal of Clinical Oncology*, 23(22):4838–4839.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90(2):367–378.

US Food and Drug Administration (1998). Statistical principles for clinical trials (ICH E-9). In *International Conference on Harmonization*. U.S. Food and Drug Administration, DHHS.

US Food and Drug Administration (2010). Guidance for Industry – Adaptive Design Clinical Trials for Drugs and Biologics. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf`.

Wang, R., Lagakos, S., Ware, J., Hunter, D., and Drazen, J. (2007a). Statistics in medicine–reporting of subgroup analyses in clinical trials. *The New England journal of medicine*, 357(21):2189–2194.

Wang, S.-J. (2007). Biomarker as a classifier in pharmacogenomics clinical trials: a tribute to 30th anniversary of PSI. *Pharmaceutical Statistics*, 6(4):283–296.

Wang, S.-J., Hung, H. J., and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, 51(2):358–374.

Wang, S.-J., O'Neill, R. T., and Hung, H. J. (2007b). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*, 6:227–244.

Wang, Y. and Leung, D. (1997). Bias reduction via resampling for estimation following sequential tests. *Sequential Analysis*, 16(3):249–267.

Wathen, J. K., Thall, P. F., Cook, J. D., and Estey, E. H. (2008). Accounting for patient heterogeneity in phase II clinical trials. *Statistics in Medicine*, 27:2802–2815.

Westfall, P. and Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99(1):25–40.

Wiens, B. L. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2:211–215.

Wiens, B. L. and Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*, 15:929–942.

Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association*, 266(1):93.

Zhou, X., Liu, S., Kim, E. S., Herbst, R. S., and Lee, J. J. (2008). Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine. *Clinical Trials*, 5:181–193.

Zuber, E., Brannath, W., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2006). Phase II/III seamless adaptive designs with Bayesian decision tools for an efficient development of a targeted therapy in oncology. Technical report, Department of Statistics and Decision Support Systems, University of Vienna.