SIMULTANEOUS POPULATION AND DOSE SELECTION IN CLINICAL TRIALS AND CLUSTER VALIDATION

by

Siyu Li

B.S., Biotechnology, Wuhan University, 2005

M.A., Applied Statistics, University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Siyu Li

It was defended on

Jan 30th, 2014

and approved by

Allan R. Sampson, Ph.D., Department of Statistics, Professor

Yu Cheng, Ph.D., Department of Statistics, Assistant Professor

Leon J. Gleser, Ph.D., Department of Statistics, Professor

Abdus S. Wahed, Ph.D., Department of Biostatistics, Associate Professor

Dissertation Director: Allan R. Sampson, Ph.D., Department of Statistics, Professor

SIMULTANEOUS POPULATION AND DOSE SELECTION IN CLINICAL TRIALS AND CLUSTER VALIDATION

Siyu Li, PhD

University of Pittsburgh, 2014

In clinical trials, the population of interest may be heterogeneous with regard to a subject's protein expression level, genotype, or other characteristics, e.g., age or initial disease severity. In particular, there can exist a subpopulation of subjects with certain characteristics that are more sensitive to the targeted agents. Wang et al. [68] suggested a two-stage design involving subpopulation enrichment along with a sample size adaptation in the second stage when evaluating the treatment effects on the overall population and the subpopulations.

An important component of drug development is to select the minimum effective dose (MED). Multiple comparisons and adaptive designs have been used for dose selection, typically in Phase 2 clinical trials.

In this research, we consider Phase 2 clinical trials with multiple populations and multiple doses. We propose methodologies for both non-adaptive and adaptive designs to select the most desired dose and population to enter the Phase 3 confirmative clinical trials. A testing scheme is established under the closed testing principle to strongly protect the familywise type I error rate for the population and MED choice for both non-adaptive and adaptive designs. Flexible test orderings are considered in order to achieve the largest power for a variety of study goals.

In related research for post-mortem tissue studies where we again study the heterogeneity of a population, we externally validate a previous subpopulation finding in a schizophrenia population. Previous research of ours had suggested a subpopulation of all individuals diagnosed with schizophrenia [66]. This subpopulation was termed the low GABA marker

(LGM) cluster. A new study was undertaken to validate these findings. In our research we first extend the classification approach proposed by Kapp and Tibshirani [27] and apply it to the validating data set. Then we apply the clustering analysis, as used in the previous research, on the validating data set and the combination of the defining and validating data sets to again demonstrate that the LGM finding is valid.

Keywords: population selection, dose selection, partial enrichment, adaptive design, cluster validation.

TABLE OF CONTENTS

1.0 INTRODUCTION FOR DOSE AND POPULATION SELECTION . . 1

1.1 Background and Literature Review 1

1.1.1 Subpopulation Study 1

1.1.2 Dose Selection 5

1.2 Research Overview 7

2.0 OBJECTIVES AND CHALLENGES FOR DOSE AND POPULATION SELECTION 9

2.1 Goal 9

2.2 Challenges 10

3.0 NON-ADAPTIVE DESIGN FOR TWO DOSES AND TWO POPULATIONS 12

3.1 The Problem Set Up 12

3.1.1 Notation and The Design 12

3.1.2 The Assumptions 14

3.2 The Proposed Testing Scheme 16

3.2.1 Closed Testing Procedures 16

3.2.2 Details for One Testing Scheme as an Example 17

3.2.2.1 Illustration Details 18

3.2.2.2 The Corresponding Decision Rule 19

3.2.2.3 Discussion on the Test Ordering 20

3.2.3 General Testing Schemes and Decision Rules 24

3.2.3.1 Possible Testing Schemes for Two Doses and Two Populations 24

3.2.3.2 General Decision Rules 24

3.3 Testing Individual Hypotheses 25

3.3.1 Data Structure 25

3.3.1.1 Unbiased Estimators of Drug Effects 26

3.3.1.2 Test Statistics and Their Covariance 27

3.3.2 Follmann's test 30

3.4 Simulation Studies on Finding the Appropriate Test Ordering 32

3.4.1 Selection of Test Ordering Depends on the Prior Beliefs of the Drug Effects 33

3.4.2 Selection of Test Ordering Depends on the Study Goal 40

3.5 Extending the 2 Population and 2 Dose Case to Multiple Population and Multiple Dose Cases 40

4.0 ADAPTIVE DESIGNS 52

4.1 Adaptive Design with Second Stage Sample Size Adjustment 52

4.1.1 The Adaptive Design 52

4.1.2 The Final Statistics 54

4.1.3 Final Analysis and Strong Control of Type I Error 56

4.1.4 The Adaptation Rule 58

4.2 The Advanced Adaptive Design 60

4.2.1 The Design and the Final Statistics 60

4.2.2 Final Analysis 62

4.2.2.1 Numerical Integration 63

4.2.2.2 Approximation to Follmann's Test 64

4.2.2.3 Obtaining Critical Values by Simulation in Follmann's Procedure 66

4.3 Simulation Studies 67

4.3.1 Simulation Studies for the Adaptive Design with Sample Size Adjustment 67

4.3.2 Simulation Studies for the Advanced Adaptive Design 71

5.0 CLUSTER VALIDATION 77

5.1 Introduction 77
5.2 Motivating Data 77
5.3 Cluster Validation using Kapp and Tibshirani's Classification Approach . . 80
5.3.1 Literature Review 80
5.3.1.1 Classification 81
5.3.1.2 Cluster Quality Measure 81
5.3.1.3 Generating the Null Distribution 82
5.3.1.4 Conclusion and Interpret the Results 82
5.3.2 Methods 83
5.3.3 Classification and Cluster Validation Result 86
5.4 Using Cluster Analysis to Validate the Previous Findings 87
5.4.1 On the Validating Data Set N=40 87
5.4.2 On the Combined Data Set N=124 88
5.5 Summary of Findings 91
6.0 CONCLUSIONS AND FUTURE RESEARCH 93
6.1 Conclusions 93
6.2 Future Research 94
APPENDIX. NON-ADAPTIVE DESIGN FOR MULTIPLE DOSES AND TWO NESTED POPULATIONS 98
A.1 3 population and 3 dose case 98
A.1.0.1 Unbiased Estimators of Drug Effects 98
A.1.1 Test Statistics and The Covariance 100
BIBLIOGRAPHY 103

LIST OF TABLES

1 Non-Adaptive Study Design Sample Sizes: 2 by 2 Case 14

2 Outcomes of the Procedure and Decision Rule for Section 3.2.2.1 20

3 All Possible Testing Schemes: 2 by 2 case 24

4 Simulation Parameters - Design: $\boldsymbol{\Delta}$'s in Scenario 1 34

5 Simulation Parameters - Design: $\boldsymbol{\Delta}$'s in Scenario 2 34

6 Simulation Parameters - Design: $\boldsymbol{\Delta}$'s in Scenario 3 35

7 Simulation Result: Scenario 1 ($\boldsymbol{\Delta} = [0.22, 0.33, 0.4, 0.6]'$) 36

8 Simulation Result: Scenario 2 ($\boldsymbol{\Delta} = [0.055, 0.33, 0.1, 0.6]'$) 37

9 Simulation Result: Scenario 3 ($\boldsymbol{\Delta} = [0.32, 0.26, 0.5, 0.2]'$) 38

10 Study Design Sample Sizes: 3 by 3 Case 43

11 Outcomes of the Procedure and Decision Rule for Section 3.5 51

12 First Stage Adaptive Design Sample Sizes: 2 by 2 Case 53

13 Second Stage Adaptive Design Sample Sizes: 2 by 2 Case 54

14 Second Stage Advanced Adaptive Design Sample Sizes: 2 by 2 Case 61

15 Drug Effects for Simulation Studies for the Adaptive Designs 68

16 Finding the Best Ordering for $\boldsymbol{\Delta} = [0.18, 0.27, 0.3, 0.45]'$. 70

17 Determining the Second Stage Sample Size 71

18 Finding Critical Values: $c^{\star}$'s 74

19 Finding N_0^{II} and g_0^{II} 76

20 Summary of Clustering Result on the Defining Data (N=84) 79

21 Classification Results for Random Samples 85

22 Classification Results for the New Data Set 87

23 Clustering Analysis Result on the Validating Data (N=40) 90

24 Summary of Findings (N=124) 92

25 Outcomes of the Procedure and Decision Rule for Parallel Subpopulations. . . 97

LIST OF FIGURES

1 Testing Scheme and Decision Rule in One Testing Order for Illustration Purpose 21

2 Depiction of Three Nested Populations 42

3 Two Cluster Identified in Volk et al. (2012) 79

4 Clustering Analysis Result on the Validating Data (N=40) 89

5 Clustering Analysis Result on the Validating Data (N=124) 91

6 Discrete Parallel Subopulations 95

1.0 INTRODUCTION FOR DOSE AND POPULATION SELECTION

1.1 BACKGROUND AND LITERATURE REVIEW

1.1.1 Subpopulation Study

With the rapid development of molecular biotechnologies and genomic technologies, such as IHC (ImmunoHistoChemistry) test, FISH (Fluorescence In Situ Hybridization) test, SNP (Single Nucleotide Polymorphism) genotyping and microarrays, etc., researchers are more aware that populations of interest in clinical trials are not homogeneous. Subjects are heterogeneous with regard to differing levels of protein expression and various genotypes that they carry. Certain recent studies have shown that there can exist a subpopulation of subjects with certain characteristics that are more sensitive to targeted agents. For example, breast cancer subjects with c-erbB-2 gene overexpression had significantly longer survival times under a high-dose regimen of adjuvant chemotherapy [43]. Metastatic colorectal cancer subjects with wild-type K-ras gene are more sensitive to cetuximab treatment compared to the subjects with a mutation in K-ras [30]. Furthermore, in general, disease severity levels and the baseline properties further contribute to the heterogeneity of the treatment responses of the population of interest. For example, antidepressant drug effects are related to initial severity, with the benefit increasing with the initial severity of depression [31]; [17].

There are marketed drugs that are specifically targeting subpopulations of sensitive subjects. For example, Iressa® (gefitinib) is used for treatment of adults with non-small cell lung cancer with activating mutations in EGFR-TK (Epidermal Growth Factor Receptor-Tyrosine Kinase). Zelboraf® (vemurafenib) effectively treats metastatic melanoma subjects who have BRAF V600E gene mutation. Herceptin® (trastuzumab) is indicated to actively treat metastatic breast cancer subjects who over-express HER2 (Human Epidermal growth factor Receptor 2) protein [3] [4]. Erbitux$^{®}$ (cetuximab) is an effective treatment for colorectal cancer that has metastasized based on the absence of K-ras gene mutation [36] [12].

The sensitive subpopulations are identified based on classifiers. In some settings, the classifiers can be fixed at the beginning of the study. The new techniques for gene sequencing have greatly helped physicians to know subjects' genotypes and protein expression levels. For example, the number of HER2 genes in the cancer cells can be measured by the FISH test; the amount of HER2 protein on the surface of cancer cells can be measured by IHC test; and K-ras mutations can be detected by simple, robust, and sensitive gene sequencing methods [44]. The U.S. Food and Drug Administration (FDA) has approved the therascreen KRAS RGQ PCR Kit to obtain information about the K-ras mutation in metastasized colorectal cancer subjects, which is the first genetic test that has been approved by the FDA. Classifiers based on baseline properties, such as age, disease severity and blood pressure can be obtained before treatment as well.

In other settings, the classifiers can also be adaptively defined, which means they are developed during the trial based on the data available at an interim analysis [72]. Freidlin and Simon [18] proposed the signature design, which is conducted in two stages and the classifiers are determined based on the Stage I data. The treatment effects are tested at the final analysis for the overall population as well as for the selected sensitive subpopulation accrued in Stage II. Their design is intended for high-dimensional data, such as microarrays, where thousands of genes were measured but only a few might be used to identify a sensitive population. Jiang et al. [26] proposed a similar design where the classifier is based on continuous biomarkers, for example, protein expression levels and blood pressure. The procedure estimates a cutoff value of the biomarker based on Stage I data and consequently defines the sensitive subpopulation. The treatment effects are evaluated on both the overall population and the sensitive subpopulation at the end of the study.

In this dissertation, we consider biopharmaceutical clinical trials where the sensitive subpopulations are identified by fixed classifiers, which are well defined before the study.

The true proportion of a sensitive subpopulation to the overall disease population cannot be known precisely. However, using numerous datasets from various sources, these proportions can be in many cases be accurately estimated. In the United States, there are HMO's (Health Maintenance Organizations) and the VA (Veterans Administration) that maintain large pertinent computerized databases. The vast data in these managed care database systems are rich sources to study the characteristics, proportions and distributions of possible subpopulations [33]. The WHO (World Health Organization) also maintains Global InfoBases for worldwide population epidemiology data and for chronic diseases and their risk factors, which can serve as a good source to estimate the proportion of a subpopulation based on baseline properties [11] [46]. For example, the proportion (global and for each area) of subpopulations with diabetes and the proportion (global and for each area) of subpopulation with hypertension have been obtained from the WHO Global InfoBse [1] [28]. Meta-analysis can also be used to estimate the proportions of subpopulations by combining information from multiple studies. For example, the obesity proportion is obtained based on databases from various sources which cover approximately 88% of the world population [29]. Another example would be the proportion of a specific gene expression or gene mutation. Numerous studies have been conducted on the subpopulation with HER2 overexpression. Based on such databases and studies, it is reasonable to assume that the proportions of the subpopulations defined by widely-studied classifiers are constants. This assumption is used in the proposed designs we develop.

Traditionally, the treatment effect is evaluated on either an unselected overall population or a sensitive subpopulation in a clinical trial, with the intended population being defined in the protocol [41] [56] [19] [22]. If the subpopulation, in which it is more likely to show an effective drug effect, is identified a priori and investigated, then such a subpopulation is called an enriched population. There are various designs for such populations and studies which are generally termed enrichment designs [61]. Sometimes the assessment of the treatment effect in the subpopulation is *post hoc*, where the subpopulation analysis is conducted after the the clinical trials find no positive overall treatment effects [10] [47]. The FDA rejects such *post hoc* subpopulation analyses due to the risk of spurious results [38]. Moreover, when multiple comparisons in subpopulation analyses are not appropriately performed, the probability of false positive findings can be substantial, and therefore may lead to approval of drugs that are not effective [21] [67]. Methods have been proposed to assess the overall treatment effect and the treatment effect on the sensitive subpopulation in one trial [72] [26] [58].

In order to accelerate drug development, the FDA is encouraging development of adaptive designs with more flexibility, which promise quicker results and smaller trials [65]. Adaptive designs have been used for assessing whether there is a drug effect in the overall population or in the subpopulations [54]. However, due to multiple hypotheses being potentially tested for different populations and the adaptive nature of the study, one major concern is how to protect the FamilyWise Type I Error Rate (FWER) in the strong sense. The strong control of FWER is defined by Hochberg and Tamhane [23] as “the probability of making one or more false discoveries, or type I errors among all the hypotheses when performing multiple hypotheses tests”. The FDA (2010) stated in the draft guidance for adaptive design clinical trials for drugs and biologics that “The chief concerns with these [adaptive] designs are control of the study-wide Type I error rate ...”. Russek-Cohen and Simmon [54] presented a novel two-stage adaptive design that incorporated a test for a subpopulation. However, they did not prove strong control of FWER for their design. Rosenblum and Van Der Laan [50] proposed a procedure based on Russek-Cohen and Simmon’s work, and they modified the design such that FWER was strongly controlled. Zhao et al. [72] proposed an adaptive design called the feedback procedure in clinical trials with a sensitive subpopulation, and showed their procedure strongly controlled for FWER. Song and Chi [58] proposed a method for testing both the overall and subpopulation hypotheses with strong control of FWER and also having an optimal power property. Wang et al. [69] proposed a design where the non-sensitive subpopulation was dropped after Stage I and the hypothesis was focused on a single population or subpopulation. Wang et al. [68] considered a setting with two fixed binary classifiers. The subpopulations were defined in the protocol, and were nested, i.e., subpopulation 1 is nested in subpopulation 2, and subpopulation 2 is nested in the overall population. The authors proposed a two-stage adaptive enrichment design involving a second stage sample size adaptation.

Beyond the difficulty of strong controlling the FWER, there are other challenges for these proceeding noted adaptive enrichment designs. One is that the results are hard to interpret in that the drug effects are evaluated on certain subpopulations. Also, because of the enrichment of the subpopulation, the treatment effect for the overall population is usually estimated with bias [68]. The FDA (2010) stated that “The chief concerns with these [adaptive] designs are ... minimization of the impact of any adaptation-associated statistical or operational bias on the estimates of treatment effects, and the interpretability of trial results”. For example, the study proposed in Wang et al. [68] failed to provide unbiased estimators of the drug effects for the overall population if the subpopulations were enriched in Stage II. The challenges also come from the lack of testing the interactions of drug effect within the subpopulations and the overall population. Simon and Wang [57] suggested that if effectiveness was established only on the subpopulation, then the conclusion of the study should avoid the overall population.

1.1.2 Dose Selection

Clinical trials that involve new drug development typically have four phases, commonly known as Phases 1, 2, 3 and 4. After the pharmacodynamics and pharmacokinetics of the drug components have been studied, the candidate drug will be tested on humans for its efficacy and safety in Phase 1 clinical trials. The most common objective of traditional Phase 2 clinical trials is to find a dose (doses) of a drug candidate that is (are) both efficacious and safe. Phase 2 also defines the population on which to study drug. These studies are often designed with small or moderate sample sizes, and they are usually short-term. The dose (doses) selected in Phase 2 is (are) used in the Phase 3 clinical trials. Phase 3 clinical trials are confirmatory studies and the objective of Phase 3 trials is to verify the established findings from the earlier stages. Certain Phase 3 trials can be long-term that can last up to several years and large numbers of subjects being recruited. If the drug candidate is shown to be efficacious and safe through Phase 1, 2 and 3 of the clinical trials, a new drug application will be filed to the FDA for the approval of the drug. After the drug goes on the market, there can be post-marketing studies, which are called Phase 4 clinical trial. [64]

Phase 2 trials are usually designed to compare one or more doses of the drug candidate against placebo (or active control that we want to show superiority to). Typically, these studies are designed with parallel treatment groups with fixed doses and placebo. In common cases, the goal is to find the Minimum Effective Dose (MED) of the candidate drug.

MED is defined as ”the lowest dose producing a clinically important response that can be declared statistically, significantly different from the placebo response” (See Ruberg [52] [53]). Multiple comparison methods or modeling approaches are typically used to analyze the trial results. Various multiple comparison procedures have been developed based on contrasts of the treatment group means [51] [15] [34]. Step-down multiple comparison procedures are widely used based on the closure principle proposed by Marcus et al. [39], thereby preserving the FWER. Multiple comparison procedures treat dose as a qualitative factor, and no assumption about the underlying dose-response model needs to be made. Therefore, multiple comparison procedures are relatively robust to the underlying dose-response distribution. Model-based approaches, on the other hand, assume a parametric model for the relationship between response and dose, where dose is considered as a quantitative factor. Sometimes, the toxicity is further evaluated in Phase 2 trial, and model-based approaches can be used for dose-finding based on efficacy-toxicity trade-offs [62].

There is substantial literature proposing methods on adaptive designs, which modify the trial design and/or statistical procedures during the study based on the observed interim data [8]. Adaptive dose finding designs are widely used in Phase 2 clinical trials to identify MED and/or MTD. Bauer and Rohmel [6] suggested an adaptive design for establishing the relationship between response and dose. Zhang et al. [71] proposed an adaptive dose-finding design that incorporates both efficacy and toxicity assessment. Sampson and Sill [55] proposed an adaptive design concerning dropping the inferior treatment groups. Sample size re-estimation designs are another widely studied adaptive designs. Proschan and Hunsberger [48] suggested a method of re-estimating the sample size of Stage II based on conditional power. Based on Proschan and Hunsberger’s work, Liu and Chi [37] developed an adaptive sample size adjustment design dealing with smaller than anticipated effect size. Hung et al. [24] proposed a sample size modification based on an interim review of effect size.

Although there is previous research focusing on selecting the sensitive populations or finding the effective/safe doses, however, there appears to be little research dealing with simultaneously dose and population selection. In this dissertation, we want to propose non-adaptive and adaptive designs of simultaneous population and dose selection in one study.

1.2 RESEARCH OVERVIEW

In our research, we consider settings where the classifier is pre-specified and the subpopulations are nested, as in Wang et al. [68]. We focus on one subpopulation cases in this dissertation while briefly introducing the extension to multiple subpopulations cases. Moreover, we consider trials where the follow-up time is relatively short so that adaptive designs are feasible. We consider evaluating treatment effects for each does on every subpopulation as well as the overall population. Our research focuses on designing studies to evaluate the treatment effects of each dose on both the subpopulations and the overall population, and finding the most desirable dose and population combination, while controlling FWER's. While our main purpose is developing adaptive designs to accomplish this, we also examine fixed sample size designs.

In this dissertation, we first introduce the objectives of these dose and population selection studies in Section 2.1. Then we discuss the challenges of designing such studies in Section 2.2.

In Chapter 3, we focus on the two dose and two population case, where the possible sensitive subpopulation is nested in the overall population. We propose a non-adaptive design to simultaneously pick the desired population and the MED, and we also provide the statistical methodologies for analyzing the data collected from such designs. The design of the study is introduced in Section 3.1.1, and the assumptions are discussed in Section 3.1.2. In order to strongly control the familywise type I error, we construct the testing schemes under the principle of closed testing, which is carefully described in Section 3.2.1. There are various testing schemes that can be constructed under the principle of closed testing procedures. We illustrate one such testing scheme and its corresponding decision rule as an example in Section 3.2.2. We also propose general testing schemes and their decision rules in Section 3.2.3. After we build a closed testing scheme, we illustrate in Section 3.3 how each hypothesis in the testing scheme can be tested using ideas of Follmann [16]. Because there are various orderings of the proposed testing schemes, we show how simulation studies allow us to choose an appropriate test ordering based on the study objectives and the beliefs of the drug effects in Section 3.4 . In Section 3.5, we show how to extend the two population and two dose cases to three population and three dose cases.

In Chapter 4, for the two dose and two population settings, we propose two adaptive designs. The first adaptive design (Section 4.1) considers a second stage sample size adjustment and the second adaptive design (Section 4.2) allows both a second stage sample size adjustment and a second stage subpopulation sampling proportion adjustment. Both adaptive designs are fully developed and analyses that strongly control FWER are given. Limited simulation results are presented in Section 4.3 to show the characteristics of the adaptive designs.

In Chapter 5, we present auxiliary material concerning identifying possible subpopulations based on post-mortem studies in schizophrenia. Previously, we have identified a cluster of schizophrenia subjects that consistently express lower GABA marker mRNA levels [66]. A second dataset was to be made available due to these findings, and this motivates us to consider approaches to see if a previously identified cluster is still present in a new dataset. In other words, we want to show whether the previously found cluster is valid. A review of previous study and study goal is provided in Section 5.1. We review the motivating data in Section 5.2. Cluster validation in new data sets is discussed by Kapp and Tibshirani [27], which we review, as well as other related literature (Section 5.3.1). We propose to extend Kapp and Tibshirani's classification approach to make the procedure more appropriate for our purposes (Section 5.3.2). The proposed cluster validation procedures is applied to our motivating data (Section 5.3.3). We also directly apply the cluster analysis approach, as used for the previous data set, to the validating data set or the combination of the motivating and validating data sets to again show the cluster findings (Section 5.4). Summary of findings are provided in Section 5.5.

In Chapter 6, we provide conclusions and possible future research directions.

2.0 OBJECTIVES AND CHALLENGES FOR DOSE AND POPULATION SELECTION

2.1 GOAL

Consider a Phase 2 clinical trial where the drug effects of multiple dose levels on multiple populations are to be evaluated. After the trial, a decision is made regarding which dose and which population proceeds to the Phase 3 clinical trials.

Previously proposed methods for designing Phase 2 clinical trials and analyzing the collected data focused on either finding a dose in one homogeneous population or finding a population with the fixed dose in a single trial. We want to develop methodologies of selecting dose and population simultaneously within one trial.

The goals of such studies vary, depending on what the biopharmaceutical companies want to achieve, e.g., aiming at approval of the drug no matter the population, or marketing to the largest population, all in the context of the drug's properties, e.g., severe side effects when doses are too high. Here, we describe three reasonable company objectives.

1. **Identify any dose and population combination.**

   When there is no concern about adverse effects of the highest dose under study, it is reasonable for the pharmaceutical company to consider the trial a success if the treatment effect on any dose and population combination is shown to be effective.

2. **Identify the largest population where there is at least one dose effective and the MED for that population.**

   Sometimes, the pharmaceutical companies want to have as large a market as reasonable. As a consequence, our primary goal is to identify the largest population possible where there is at least one effective dose for this population. Our secondary goal is that for this population, we want to find the minimum effective dose. This is because the side effects are expected to be lower at lower doses, as well as less severe.

3. **Identity the lowest dose which is effective on at least one population and the largest population corresponding to this dose.**

   Sometimes, the drug might have severe side effect at high doses for the overall population or a subpopulation. For example, the anti-coagulant drug, Coumadin$^{®}$ (warfarin), which reduces blood clotting, is reported to have severe adverse effect in the elderly people. Hylek et al. [25] reported that subjects over 80 years old are especially susceptible to bleeding complications. Thus for such drugs, we'd like to avoid the high doses, and prefer low doses even if the low doses only work on a sensitive subpopulation. Therefore, the primary goal is to find the lowest dose which works for at least one population, and the secondary goal is to find the largest population for this dose.

2.2 CHALLENGES

A major challenge is how to find the most desired population and dose corresponding to the study goal. The study goal will be described in details in Section 2.1, from which we will see that the study goal varies from study to study. For example, the primary goal can be finding the largest population or finding the lowest dose. Obviously, there is no unique testing procedure that is best (largest power for the desired dose and population) for every study goal. We want to address this challenge by designing flexible testing schemes such that the testing schemes are easy to be modified in order to perform best for each specific study goal.

Another challenge of such multiple dose and multiple population studies is how to strongly control the familywise Type I error rate. In a study, the treatment effect of each dose in each population is compared with the controls from the same population. For an L dose (excluding control) and M populations problem, there are $L * M$ contrasts to be tested, i.e., comparing the drug responses of Dose i in Population j with controls in Population j, where $i = 1, 2, \cdots, L,\ j = 1, 2, \cdots, M$. Appropriate multiple comparison procedures are required in order to test all the treatment effects without inflating Type I error.

Sometimes the proportion of the subpopulation to the overall population is small. For moderate sample sizes, there are relatively few subjects that are from the subpopulation. Thus, we don't have large power for testing the treatment effects on the subpopulation. Enrichment designs have been proposed previously, where all subjects are selected from the subpopulation due to small sample sizes. However, the enrichment studies only focus on the subpopulation of interest, and the treatment effect of the overall population is not evaluated. One possible solution is that we may partially enrich the subpopulation, which means we pick a larger proportion of subjects from the subpopulation than the true proportion, so that there will be enough subjects from the subpopulation. However, the sample of a partial enrichment design is not representative of the true population, and we won't be able to obtain an unbiased estimator of the drug effect for the overall population. In this dissertation, we want to partially enrich the subpopulation, while maintaining unbiased estimators of the drug effects for the subpopulation and the overall population.

It is a challenge how to conduct the study adaptively. Adaptive designs are more flexible, with quicker results, smaller sample sizes or larger power. In these population and dose selection studies, possible adaptations include increasing the second stage sample size, increasing the proportion of subjects selected from the subpopulation, and dropping a dose/doses after examining the interim data. When adaptive designs are conducted, it is important to pay attention to controlling the study-wide type I error.

3.0 NON-ADAPTIVE DESIGN FOR TWO DOSES AND TWO POPULATIONS

3.1 THE PROBLEM SET UP

3.1.1 Notation and The Design

Consider one binary indicator I. The indicator can be a genotype (e.g., KRAS gene mutation versus no mutation), or protein expression levels (e.g., over-expression of HER2 protein versus normal expression), or a baseline covariate (e.g., age < 80 versus age ≥ 80), or the initial disease severity (e.g., severe depression versus mild or moderate depression), etc. The study subjects can be classified prior to sampling into one of the two mutually exclusive subsets I^+ and I^-. In the study, we are interested in assessing the drug effects in two populations: the overall population and the subpopulation with positive indicator, I^+.

Denote the two populations of interest by G_A and G_S, where G_A is the overall population, and G_S is the subpopulation with I^+. Also denote G_{S^-} as the compliment population with negative indicator I^-. The subpopulation G_S and the compliment population G_{S^-} are both nested in the overall population G_A, and the compliment population consists of all subjects that are contained in G_A but not contained in G_S. Denote the populations by G_l, where $l = A, S, S^-$.

Suppose the proportion of the subpopulation G_S out of the overall population G_A is known, as described in Chapter 1, and is denoted by f. Since the subpopulation and the complimentary population are mutually exclusive, the true proportion of the complimentary population G_{S^-} is $(1-f)$. Suppose in our study design, the proportion of subjects in the sample that will be selected from the subpopulation G_S is chosen to be g, $0 < g < 1$. Therefore, the proportion of subjects that will be selected from the complimentary population G_{S^-} is $1-g$. When $g = f$, that indicates that the sample has "true" proportion of subjects that are from each population; $g > f$ indicates that the subpopulation with I^+ is partially enriched; and $g < f$ indicates that the complimentary population with I^- is partially enriched.

Usually, there are more than 2 dose levels being considered in clinical trials. Furthermore, there might be more than 2 population levels. However, for simplicity and illustrative purpose, we are going to consider 2 population and 2 dose case to demonstrate our proposed methodologies for trial design and data analysis. Extending our 2 population and 2 dose case to multiple population and multiple dose cases is notationally complex, while the methodologies of study design and data analysis are similar. The extension will be discussed in Section 3.5.

Suppose the 2 dose levels of interest are Low and High. There is a control group as well, receiving placebo (or active comparator that we want to show superiority to). The subjects within each population are randomly assigned to one of the three treatment arms: low dose, high dose, or placebo. Denote the treatments by m, where $m = L, H, c$.

Denote the total sample size for the study by N_{Total}. Thus, we have stratified sampling, i.e., gN_{Total} subjects are sampled from the subpopulation G_S, and $(1-g)N_{Total}$ subjects sampled from the complimentary population G_{S^-}.

For simplicity, assume that subjects within a population are randomly assigned to high dose, low dose, or placebo with equal numbers in each. Denote the total number of subjects receiving each treatment by N, where $N = N_{Total}/3$, so that $gN_{Total}/3 = gN$ subjects in the subpopulation group G_S will be assigned to receive high dose, low dose, and placebo, respectively. Similarly, $(1-g)N$ subjects in the complimentary population group G_{S^-} will be assigned to receive high dose, low dose, and placebo, respectively. The following table summarizes the experimental design with respect to the sample size of stratified sampling.

Denote the response of each subject in the study by $X_{G_l,m,i}$, where $l = S, S^-$; $m = L, H, c$; $i = 1, 2, \cdots, gN$, if $l = S$; $i = 1, 2, \cdots, (1-g)N$, if $l = S^-$.

Denote by $\mu_{G_l,m}$ the population l true mean response at dose m, where $l = A, S, S^-$; $m = L, H, c$. Denote the true drug effect of dose m relative to the control group in population

Table 1: Non-Adaptive Study Design Sample Sizes: 2 by 2 Case

	Doses			
	Low	High	Control	Total
Subpopulation G_S	gN	gN	gN	gN_{Total}
Complimentary Pop. G_{S^-}	$(1-g)N$	$(1-g)N$	$(1-g)N$	$(1-g)N_{Total}$
Overall Pop. G_A	N	N	N	N_{Total}

Note: N_{Total} is the total sample size, N is the sample size for each dose, and g is the sampling proportion of subjects from subpopulation.

l by $\Delta_{G_l,m}$,

$$\Delta_{G_l,m} = \mu_{G_l,m} - \mu_{G_l,c}, \tag{3.1}$$

where $l = A, S, S^-$; $m = L, H$.

3.1.2 The Assumptions

In this dissertation, we make the following assumptions:

1. All responses $X_{G_l,m,i}$ from the subpopulation and complimentary population are mutually independent. Thus the subpopulation sample mean responses $\bar{X}_{G_S,L}$, $\bar{X}_{G_S,H}$, $\bar{X}_{G_S,c}$, $\bar{X}_{G_{S^-},L}$, $\bar{X}_{G_{S^-},H}$, and $\bar{X}_{G_{S^-},c}$ are mutually independent.
2. $X_{G_l,m,i}$ are normally distributed with mean $\mu_{G_l,m}$ and a common variance σ^2. Due to the generally sufficiently large sample sizes in each group, we assume throughout this dissertation that the variance is known,

$$X_{G_l,m,i} \sim \mathcal{N}(\mu_{G_l,m}, \sigma^2), \quad \sigma^2 \textit{ is known},$$

where $l = S, S^-$; $m = L, H, c$; $i = \begin{cases} 1, 2, \cdots, gN, & \text{if } \ l = S; \\ 1, 2, \cdots, (1-g)N, & \text{if } \ l = S^-. \end{cases}$

We can easily obtain the distributions of sample mean responses for each treatment m, where $m = L, H, c$, for the subpopulation and the complimentary population, namely,

$$\bar{X}_{G_S,m} \sim \mathcal{N}(\mu_{G_S,m}, \frac{\sigma^2}{gN}),$$

$$\bar{X}_{G_{S^-},m} \sim \mathcal{N}(\mu_{G_{S^-},m}, \frac{\sigma^2}{(1-g)N}),$$

where $m = L, H$.

3. We assume that the drug has nonnegative effect for both doses, which is that we assume that the mean response of subjects $\mu_{G_l,m}$ in population G_l with dose m is not smaller than the mean response of subjects $\mu_{G_l,c}$ in that population with placebo, so that

$$\Delta_{G_l,m} = \mu_{G_l,m} - \mu_{G_l,c} \geq 0,$$

where $l = A, S, S^-$; $m = L, H$.

4. As discussed in Section 1.1.1, we assume that the proportion f of the subpopulation G_S is known. Therefore, we have that for $m = L, H, c$,

$$\mu_{G_A,m} = f\mu_{G_S,m} + (1-f)\mu_{G_{S^-},m},$$

and for $m = L, H$,

$$\begin{aligned}
\Delta_{G_A,m} &= \mu_{G_A,m} - \mu_{G_A,c} \\
&= [f\mu_{G_S,m} + (1-f)\mu_{G_{S^-},m}] - [f\mu_{G_S,c} + (1-f)\mu_{G_{S^-},c}] \\
&= f(\mu_{G_S,m} - \mu_{G_S,c}) + (1-f)(\mu_{G_{S^-},m} - \mu_{G_{S^-},c}) \\
&= f\Delta_{G_S,m} + (1-f)\Delta_{G_{S^-},m}
\end{aligned}$$

3.2 THE PROPOSED TESTING SCHEME

In the clinical trials, we are interested in evaluating the drug effects on four groups: {G_A, Low}, {G_A, High}, {G_S, Low}, {G_S, High}, which are low dose in the overall population, high dose in the overall population, low dose in the subpopulation, and high dose in the subpopulation, respectively. Thus we want to construct a testing procedure that tests for the drug effects $\mathbf{\Delta}$, where $\mathbf{\Delta} = (\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H})'$.

First of all, the testing scheme must be flexible in order to meet various goals of the company as described in Section 2.1. Moreover, as discussed in Section 2.2, it is very important that the Type I FWER is strongly controlled using our proposed testing scheme since multiple comparisons are conducted.

In this section, we will first introduce the concept of closed testing procedures proposed by Marcus, Peritz, and Gabriel [39]. Then we will introduce our proposed testing scheme which is constructed under the principle of closed testing.

3.2.1 Closed Testing Procedures

Marcus, Peritz, and Gabriel [39] proposed the concept of closed testing procedures, which is a property of multiple comparison testing procedures that consists of a set of hypotheses that are closed under intersection. Closed testing procedures control the familywise type I error rate at level α in the strong sense.

Familywise type I Error Rate (FWER) is defined as the probability of rejecting one or more true null hypotheses when performing multiple hypotheses tests, i.e., FWER = P(Rejecting at least one true null hypothesis). A procedure controls the familywise error rate in the weak sense if the probability of rejecting one or more true null hypotheses when all null hypotheses are true is less than or equal to the significance level α, i.e., P(Rejecting at least one true null hypothesis | All null hypotheses are true) $\leq \alpha$. A procedure controls the FWER in the strong sense if the probability of rejecting one or more true null hypotheses under any configuration of the true and non-true null hypotheses is less than the significance level α, i.e., P(Rejecting at least one true null hypothesis | Any configuration of true and non-true hypotheses) $\leq \alpha$.

A closed testing procedure requires a set of hypotheses that are closed under intersection. Before we introduce the details of closed testing procedures, we first elaborate on "closed under intersection". Suppose we have a random sample following a distribution with unknown parameters $\vec{\theta}$: $\vec{\theta} \in \Omega$, where Ω is the parameter space. Suppose we want to test a set of hypotheses, with $H_0^{(1)}, H_0^{(2)}, \cdots, H_0^{(K)}$ as the null hypotheses and $H_a^{(1)}, H_a^{(2)}, \cdots, H_a^{(K)}$ as the alternatives. Each hypothesis $H^{(i)}$ tests whether the parameters $\vec{\theta}$ are in the null parameter space ω_i: $\vec{\theta} \in \omega_i$, for $i = 1, 2, \cdots, K$. Since Ω is the parameter space, it is obvious that each null parameter space ω_i is contained in Ω: $\omega_i \subseteq \Omega$. Let W be the set of all null parameter spaces: $W = \{\omega_i\}$. Then W is closed under intersection if for any elements ω_i, ω_j in W, the intersection of the two elements is also in W: $\forall \omega_i$, $\omega_j \in W$ implies $\omega_i \cap \omega_j \in W$, where $i, j = 1, 2, \cdots, K$.

The closed testing procedure requires a set of hypotheses, i.e., $H^{(1)}, H^{(2)}, \cdots, H^{(K)}$ that are closed under intersection, then a null hypothesis $H_0^{(i)}$ can be rejected at level α if and only if ω_i and all null parameter spaces w's that are included in w_i and belong to W are tested and rejected at local significance level α. Local significance means that each of these hypotheses is tested at its own nominal significance level α using any valid statistical procedure.

For example, suppose there are a set of hypotheses $H^{(1)}, H^{(2)}, H^{(3)}$ to be tested. The closed testing procedure allows the rejection of any one of these hypotheses, say $H^{(1)}$, if all possible intersection hypotheses involving $H^{(1)}$, i.e., $H^{(1)}, H^{(1)} \cap H^{(2)}, H^{(1)} \cap H^{(3)}, H^{(1)} \cap H^{(2)} \cap H^{(3)}$, can be each rejected by using valid α level tests.

Marcus et al. [39] provided a proof that any closed testing procedure protects the familywise error rate in the strong sense.

3.2.2 Details for One Testing Scheme as an Example

In this section, we construct the testing scheme which we show to be a closed testing procedure. For the two population and two dose case, our testing scheme consists of 4 individual hypotheses.

There are 4 individual hypotheses within each testing scheme. There is one common hypothesis to be tested , which is to test whether the 4 drug effects all equal to 0, i.e., $\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$. Let $c_1, c_2, c_3, c_4 \in \{\{G_A, L\}, \{G_A, H\}, \{G_S, L\}, \{G_S, H\}\}$ and $c_1 \neq c_2 \neq c_3 \neq c_4$, i.e., each c_i is a population and dose combination and c_1, c_2, c_3, c_4 represent different combinations. The other three hypotheses test whether $\Delta_{c_1} = \Delta_{c_2} = \Delta_{c_3} = 0$, $\Delta_{c_1} = \Delta_{c_2} = 0$, and $\Delta_{c_1} = 0$, respectively. Each corresponding alternative hypothesis states that there is at least one positive drug effect among the populations and doses considered in the null hypothesis. Note that c_1, c_2, c_3, c_4 can be chosen in multiple ways from $\{\{G_A, L\}, \{G_A, H\}, \{G_S, L\}, \{G_S, H\}\}$, so that there are 4! possible outcomes for choosing c_1, c_2, c_3, c_4 and hence there are 4! possible testing schemes. We denote each of these possible testing schemes by a test ordering.

The testing scheme allows for various orderings in order to achieve desired power characteristics, and the orderings are determined based on the previous knowledge or prior beliefs about the parameters (drug effects and variances) and the goal of the trial. We will show in Section 3.4 how to choose an ordering for our testing scheme under various circumstances.

Now, we show the details for one possible ordering of the testing scheme as an illustration. Specifically, we show how the decision rule is made, and how the familywise type I error rate is strongly protected. Other orderings and their decision rules similarly control the familywise type I error rate in the strong sense. The corresponding decision rules are made for each ordering accordingly.

3.2.2.1 Illustration Details

Suppose in a clinical trial, our goal is to find the largest population where there is at least one effective dose and for this population find the Minimum Effect Dose (MED). To achieve this goal, we consider the following testing scheme where the null hypotheses are

$$H_0^{(4)} : \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0, \tag{3.3a}$$

$$H_0^{(3)} : \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0, \tag{3.3b}$$

$$H_0^{(2)} : \Delta_{G_A,L} = \Delta_{G_A,H} = 0, \tag{3.3c}$$

$$H_0^{(1)} : \Delta_{G_A,L} = 0, \tag{3.3d}$$

where $\Delta_{G_l,m}$ is given by equation (3.1).

Each corresponding alternative hypothesis states that there is at least one positive drug effect among the populations and doses considered in the null hypothesis:

$$H_a^{(4)} : \Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0 \quad or \quad \Delta_{G_S,L} > 0 \quad or \quad \Delta_{G_S,H} > 0, \tag{3.4a}$$

$$H_a^{(3)} : \Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0 \quad or \quad \Delta_{G_S,L} > 0, \tag{3.4b}$$

$$H_a^{(2)} : \Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0, \tag{3.4c}$$

$$H_a^{(1)} : \Delta_{G_A,L} > 0. \tag{3.4d}$$

We test the family of hypotheses in a step down manner from the highest level until we accept one null hypothesis. To be specific, $H^{(4)}$ is tested first. If $H_0^{(4)}$ is rejected, then $H^{(3)}$ is to be tested next; otherwise, stop testing. If $H_0^{(3)}$ is rejected, then continue to test $H^{(2)}$; otherwise, stop testing. Similarly, If $H_0^{(2)}$ is rejected, then continue to test $H^{(1)}$; otherwise, stop.

Note that the family of hypotheses listed in equation (3.3) and (3.4) is a step down procedure, and due to the property that a step down testing procedure is a closed testing procedure, strong control of FWER is guaranteed in our scheme.

3.2.2.2 The Corresponding Decision Rule

We establish a one to one correspondence between the chosen population and its corresponding MED and all possible outcomes of true and false null hypotheses. After the hypotheses are tested, we make the conclusions based on the testing results according to the following rules.

- If $H_0^{(4)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$, then there is no population that has an effective dose.
- If $H_0^{(4)}$ is false, i.e., $\Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0 \quad or \quad \Delta_{G_S,L} > 0 \quad or \quad \Delta_{G_S,H} > 0$, and if $H_0^{(3)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$, then we conclude $\Delta_{G_S,H} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation and the MED for this population is the high dose. We denote this choice by $\{G_S, \text{High}\}$.

- If $H_0^{(4)}$ is false, and $H_0^{(3)}$ is false, i.e., $\Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0 \quad or \quad \Delta_{G_S,L} > 0$, and if $H_0^{(2)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,H} = 0$, then we conclude $\Delta_{G_S,L} > 0$ and it is possible that $\Delta_{G_S,H} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation and the MED for this population is the low dose. We denote this choice by $\{G_S$, Low$\}$.
- If $H_0^{(4)}$ is false, $H_0^{(3)}$ is false, and $H_0^{(2)}$ is false, i.e., $\Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0$, and if $H_0^{(1)}$ is true, i.e., $\Delta_{G_A,L} = 0$, then we conclude $\Delta_{G_A,H} > 0$ and it is possible that $\Delta_{G_S,L} > 0, \Delta_{G_S,H} > 0$. We thereby establish the largest population that has a positive dose effect is the overall population and the MED for this population is the high dose. We denote this choice by $\{G_A$, High$\}$.
- If $H_0^{(4)}$ is false, $H_0^{(3)}$ is false, $H_0^{(2)}$ is false, and $H_0^{(1)}$ is false, we conclude $\Delta_{G_A,L} > 0$ and it is possible that $\Delta_{G_A,H} > 0$, $\Delta_{G_S,L} > 0$, $\Delta_{G_S,H} > 0$. We establish the largest population that has a positive dose effect is the overall population and the MED for this population is the low dose. We denote this choice by $\{G_A$, Low$\}$.

For this illustration, Table 2 summarizes all the possible outcomes and the corresponding conclusions and Figure 1 is an alternative representation.

Table 2: Outcomes of the Procedure and Decision Rule for Section 3.2.2.1

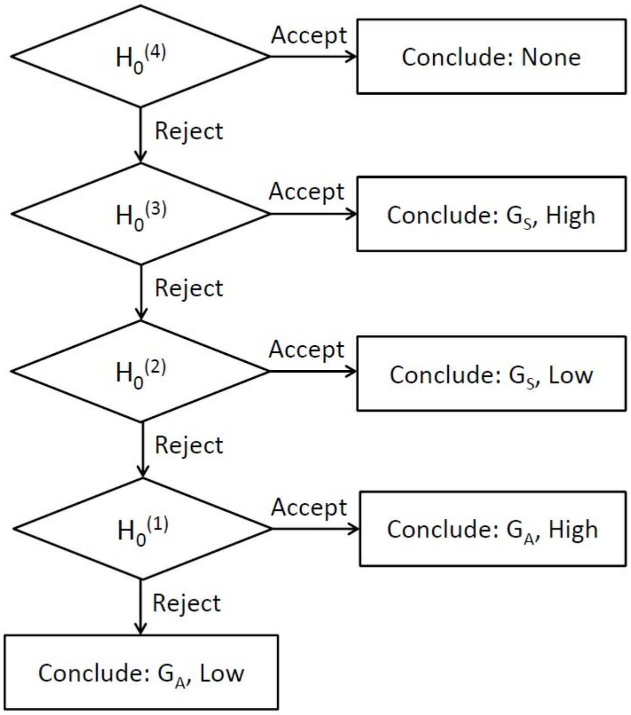
$H_0^{(4)}$	$H_0^{(3)}$	$H_0^{(2)}$	$H_0^{(1)}$	$\Delta_{G_A,L}$	$\Delta_{G_A,H}$	$\Delta_{G_S,L}$	$\Delta_{G_S,H}$	Conclusion
ACC	NT	NT	NT	0	0	0	0	None
REJ	ACC	NT	NT	0	0	0	> 0	G_S, High
REJ	REJ	ACC	NT	0	0	> 0	UKN	G_S, Low
REJ	REJ	REJ	ACC	0	> 0	UKN	UKN	G_A, High
REJ	REJ	REJ	REJ	> 0	UKN	UKN	UKN	G_A, Low

Note: ACC=Accept, REJ=Reject, NT=Not Tested, UKN=Unknown.

3.2.2.3 Discussion on the Test Ordering

There are totally 24 possible test orderings in 2 population and 2 dose case. Some of the test ordering work well for certain study goals under certain prior knowledge or beliefs of the drug effects, while other test orderings may not even make sense.

Figure 1: Testing Scheme and Decision Rule in One Testing Order for Illustration Purpose



In Section 3.2.2.1 and Section 3.2.2.2, we demonstrated a testing scheme, its test ordering and the corresponding decision rule for illustrative purpose. In simulation studies in Section 3.4, we will evaluate how this test ordering performs in several different scenarios of prior knowledge or beliefs of drug effects with respect to varying study objectives .

It makes sense that the test ordering proposed in Section 3.2.2.1 performs well for certain study objective and certain prior beliefs of the drug effects. For example, suppose the primary objective of a study is to obtain the largest population where at least one dose is effective, the secondary objective is to find the MED for that population, and the drug effects are positive for all dose and population combinations. This means that our most desired population and dose combination is $\{G_A$, Low$\}$. If we could not conclude $\{G_A$, Low$\}$, then the next most desired population and dose combination is $\{G_A$, High$\}$. Only when we could not conclude any effective dose on the overall population, we may make conclusions about the subpopulation. The decision rule for the test ordering in Section 3.2.2.1 tells us the followings:

1. If $H_0^{(4)}$ is rejected and $H_0^{(3)}$ is accepted, we conclude $\{G_S$, High$\}$. This is reasonable because accepting $H_0^{(3)}$ indicates that the drug effects of $\{G_A$, Low$\}$, $\{G_A$, High$\}$, $\{G_S$, Low$\}$ are all zero, and $\{G_S$, High$\}$ is the only population and dose combination that is effective.
2. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, and $H_0^{(2)}$ is accepted, we conclude $\{G_S$, Low$\}$ while it is possible that $\{G_S$, High$\}$ is effective. The conclusion makes sense in that within one population, we want to find the minimum effective dose. Consequently, if we know $\{G_S$, Low$\}$ is effective, we conclude $\{G_S$, Low$\}$ no matter $\{G_S$, High$\}$ is effective or not.
3. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, $H_0^{(2)}$ is rejected, and $H_0^{(1)}$ is accepted, we conclude $\{G_A$, High$\}$ while it is possible that $\{G_S$, Low$\}$ and $\{G_S$, High$\}$ are effective. The conclusion makes sense in that we want to find the largest population in this study. Consequently, if we know $\{G_A$, High$\}$ is effective, we conclude it no matter $\{G_S$, High$\}$ or $\{G_S$, Low$\}$ is effective or not.
4. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, $H_0^{(2)}$ is rejected, and $H_0^{(1)}$ is rejected, we conclude $\{G_A$, Low$\}$ while it is possible that $\{G_A$, High$\}$, $\{G_S$, Low$\}$, and $\{G_S$, High$\}$ are effective. The conclusion is reasonable in that $\{G_A$, Low$\}$ is the most desirable population and

dose combination, if it is effective, we conclude it no matter whether the other three combinations are effective or not.

However, for other study objectives, this test ordering may not be appropriate. For example, suppose now the primary goal of our study is to obtain the lowest dose which is effective on at least one population, and the secondary goal is to find the largest population where this dose is effective. This means that our most desired population and dose combination is $\{G_A$, Low$\}$. If we could not conclude $\{G_A$, Low$\}$, then the next most desired population and dose combination is $\{G_S$, Low$\}$. Only when we could not conclude the low dose effective on either overall population or the subpopulation, we consider making conclusions about the high dose. The test ordering indicated in Section 3.2.2.1 is not appropriate in that when $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, $H_0^{(2)}$ is rejected, and $H_0^{(1)}$ is accepted, according to the decision rule, we conclude $\{G_A$, High$\}$ while it is possible that $\{G_S$, Low$\}$ and $\{G_S$, High$\}$ are effective. The conclusion is not appropriate here because we want to find the lowest dose. If both $\{G_A$, High$\}$ and $\{G_S$, Low$\}$ are effective here, we are not making the correct conclusion.

Moreover, for other prior beliefs of the drug effects, this test ordering may not be appropriate either. For example, suppose the drug has no or little effect at the low dose. This means that our most desired population and dose combination is $\{G_A$, High$\}$. If we could not conclude $\{G_A$, High$\}$, then the next most desired population and dose combination is $\{G_S$, High$\}$. In this case, the test ordering illustrated in Section 3.2.2.1 may not be appropriate since we won't have sufficient power to reject $H_0^{(2)}$ when $\Delta_{G_A,L}$ and $\Delta_{G_S,L}$ both equal to zero. Other test orderings may work better under this scenario. Simulation and more discussion will be provided in Section 3.4.

Therefore, the selected ordering should depend on prior knowledge or beliefs of drug effects and the study objective, and there is no unique test ordering that fits all studies. It is important that we find the most appropriate test ordering for each study using simulation. In the following section, we propose general testing schemes and how to develop the decision rule for each test ordering, and in Section 3.4, we discuss in detail how to choose the test ordering for each scenario of drug effects with respect to the study objective.

3.2.3 General Testing Schemes and Decision Rules

3.2.3.1 Possible Testing Schemes for Two Doses and Two Populations

As mentioned before, the ordering of the testing scheme is flexible. In the two dose and two population trial setting, there are four combinations of doses and populations, i.e., $\{G_A$, Low$\}$, $\{G_A$, High$\}$, $\{G_S$, Low$\}$, and $\{G_S$, High$\}$, yielding $4! = 24$ possible sets of hypotheses, each of which is closed under intersection. The null hypotheses of all possible testing schemes are listed in Table 3 and the alternative hypotheses are the corresponding one sided alternatives that at least one of the drug effects is/are strictly greater than zero. We test the 4 hypotheses in each set in a step down manner, and there is a decision rule for each testing scheme. Hence the strong control of FWER for each family in Table 3 is guaranteed .

Table 3: All Possible Testing Schemes: 2 by 2 case

	$H_0^{(4)}$	$H_0^{(3)}$	$H_0^{(2)}$	$H_0^{(1)}$
1	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = 0$	$\Delta_{G_A,L} = 0$
2	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = 0$	$\Delta_{G_A,H} = 0$
3	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,L} = \Delta_{S,L} = 0$	$\Delta_{G_A,L} = 0$
4	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = 0$	$\Delta_{G_S,L} = 0$
5	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,H} = 0$
6	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_S,L} = 0$
7	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = 0$	$\Delta_{G_A,L} = 0$
8	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = 0$	$\Delta_{G_A,H} = 0$
9	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = 0$
10	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{A,L} = \Delta_{G_S,H} = 0$	$\Delta_{S,H} = 0$
11	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = 0$
12	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{S,H} = 0$	$\Delta_{G_S,H} = 0$
13	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{S,L} = 0$	$\Delta_{G_A,L} = 0$
14	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = 0$	$\Delta_{G_S,L} = 0$
15	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{S,H} = 0$	$\Delta_{G_A,L} = 0$
16	$\Delta_{A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_S,H} = 0$
17	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_S,L} = 0$
18	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,L} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_S,H} = 0$
19	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{S,L} = \Delta_{G_S,H} = 0$	$\Delta_{A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_A,H} = 0$
20	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = 0$	$\Delta_{G_S,L} = 0$
21	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = 0$
22	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$	$\Delta_{G_A,H} = \Delta_{S,H} = 0$	$\Delta_{S,H} = 0$
23	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{S,L} = \Delta_{S,H} = 0$	$\Delta_{G_A,H} = \Delta_{S,L} = \Delta_{S,H} = 0$	$\Delta_{S,L} = \Delta_{S,H} = 0$	$\Delta_{S,L} = 0$
24	$\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{S,L} = \Delta_{S,H} = 0$	$\Delta_{G_A,H} = \Delta_{S,L} = \Delta_{S,H} = 0$	$\Delta_{S,L} = \Delta_{S,H} = 0$	$\Delta_{S,H} = 0$

3.2.3.2 General Decision Rules

Decision rules for these testing schemes are not hard to develop. To be general, we follow the rules described as below:

1. If $H_0^{(4)}$ is accepted, it is indicated that the drug effects on all populations and doses are zero. Thus we conclude there is no effective dose on either the overall population or the subpopulation. If $H_0^{(4)}$ is rejected, then we move on and test $H_0^{(3)}$.
2. If $H_0^{(4)}$ is rejected and $H_0^{(3)}$ is accepted, we conclude the dose and population that is tested in $H_0^{(4)}$ but not tested in $H_0^{(3)}$. If $H_0^{(4)}$ is rejected and $H_0^{(3)}$ is rejected, we continue to test $H_0^{(2)}$.
3. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, and $H_0^{(2)}$ is accepted, we conclude the dose and population that is tested in $H_0^{(3)}$ but not tested in $H_0^{(2)}$. It is possible that the drug effect for the dose and population that is tested in $H_0^{(4)}$ but not tested in $H_0^{(3)}$ is positive, but we don't have evidence to conclude it. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, and $H_0^{(2)}$ is rejected, we continue to test $H_0^{(1)}$.
4. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, $H_0^{(2)}$ is rejected, and $H_0^{(1)}$ is accepted, we conclude the dose and population that is tested in $H_0^{(2)}$ but not tested in $H_0^{(1)}$. It is possible that the drug effect for the dose and population that is tested in $H_0^{(4)}$ but not tested in $H_0^{(3)}$ is positive. It is also possible that the drug effect for the dose and population that is tested in $H_0^{(3)}$ but not tested in $H_0^{(2)}$ is positive, but we don't conclude them. If $H_0^{(4)}$ is rejected, $H_0^{(3)}$ is rejected, and $H_0^{(2)}$ is rejected, and $H_0^{(1)}$ is rejected, we conclude the dose and population tested in $H_0^{(1)}$.

3.3 TESTING INDIVIDUAL HYPOTHESES

3.3.1 Data Structure

In the first part of this section, we explicitly describe how the estimates of drug effects are obtained, and how the test statistics and covariance among the test statistics are computed. In the second part, we will describe how to perform the hypothesis testing.

3.3.1.1 Unbiased Estimators of Drug Effects

The estimator of the mean response at each dose in the sub or complimentary populations is the corresponding sample mean.

$$\hat{\mu}_{G_l,m} = \bar{X}_{G_l,m}, \quad \text{for} \quad l = S, S^-; m = L, H, c.$$

Since the proportion, f, of the subpopulation to the true population is assumed known, we estimate the mean responses in the overall population by the weighted averages,

$$\hat{\mu}_{G_A,m} = f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m}, \quad \text{for} \quad m = L, H, c.$$

Then the drug effects $\Delta_{G_l,m}$ are estimated by:

$$\hat{\Delta}_{G_l,m} = \hat{\mu}_{G_l,m} - \hat{\mu}_{G_l,c} = \begin{cases} f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m} - f\bar{X}_{G_S,c} - (1-f)\bar{X}_{G_{S^-},c}, & l = A; \\ \bar{X}_{G_l,m} - \bar{X}_{G_l,c}, & l = S, S^-. \end{cases}$$

Since, $\bar{X}_{G_l,m}$ is the unbiased estimator of $\mu_{G_l,m}$ for the sub and complimentary population, we have that,

$$E(\hat{\mu}_{G_A,m}) = E[f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m}] = f\mu_{G_S,m} + (1-f)\mu_{G_{S^-},m} = \mu_{G_A,m}, \text{ for } m = L, H, c,$$

are also unbiased estimators. Therefore, $\hat{\Delta}_{G_l,m}$ are unbiased estimators for $\Delta_{G_l,m}$, in that,

$$E(\hat{\Delta}_{G_l,m}) = E[\hat{\mu}_{G_l,m} - \hat{\mu}_{G_l,c}] = \mu_{G_l,m} - \mu_{G_l,c} = \Delta_{G_l,m}, \text{ for } l = A, S; m = L, H.$$

3.3.1.2 Test Statistics and Their Covariance

Next, we calculate test statistics for the drug effects $\mathbf{\Delta} = [\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H}]'$. Since the responses are normally distributed, and the variance is known, $\mathbf{Z}$ test statistics are used. The test statistics of the drug effects at dose for the subpopulation are computed as

$$Z_{G_S,m} = \frac{\hat{\Delta}_{G_S,m}}{\sqrt{\sigma^2_{(\hat{\Delta}_{G_S,m})}}} = \frac{\bar{X}_{G_S,m} - \bar{X}_{G_S,c}}{\sqrt{\sigma^2_{(\bar{X}_{G_S,m} - \bar{X}_{G_S,c})}}} = \frac{\bar{X}_{G_S,m} - \bar{X}_{G_S,c}}{\sqrt{\sigma^2(\frac{1}{gN} + \frac{1}{gN})}} = \frac{\bar{X}_{G_S,m} - \bar{X}_{G_S,c}}{\sqrt{\frac{2\sigma^2}{gN}}}, \tag{3.5}$$

for $m = L, H$.

The test statistics of the drug effects for the overall population are computed as, for $m = L, H$,

$$\begin{aligned} Z_{G_A,m} &= \frac{\hat{\Delta}_{G_A,m}}{\sqrt{\sigma^2_{(\hat{\Delta}_{G_A,m})}}} \\ &= \frac{f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m} - f\bar{X}_{G_S,c} - (1-f)\bar{X}_{G_{S^-},c}}{\sqrt{\sigma^2_{f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m} - f\bar{X}_{G_S,c} - (1-f)\bar{X}_{G_{S^-},c}}}} \\ &= \frac{f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m} - f\bar{X}_{G_S,c} - (1-f)\bar{X}_{G_{S^-},c}}{\sqrt{f^2\frac{\sigma^2}{gN} + (1-f)^2\frac{\sigma^2}{(1-g)N} + f^2\frac{\sigma^2}{gN} + (1-f)^2\frac{\sigma^2}{(1-g)N}}} \\ &= \frac{f\bar{X}_{G_S,m} + (1-f)\bar{X}_{G_{S^-},m} - f\bar{X}_{G_S,c} - (1-f)\bar{X}_{G_{S^-},c}}{\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}}. \end{aligned} \tag{3.6}$$

Since the sample means $\bar{X}_{G_S,L}$, $\bar{X}_{G_S,H}$, $\bar{X}_{G_S,c}$, $\bar{X}_{G_{S^-},L}$, $\bar{X}_{G_{S^-},H}$, and $\bar{X}_{G_{S^-},c}$ are mutually independent (Section 3.1.2), it follows that

$$cov(\bar{X}_{G_l,m}, \bar{X}_{G_{l'},m'}) = 0,$$

for $l, l' = S, S^-$; $m, m' = L, H, c$; $\{l, m\} \neq \{l', m'\}$.

The variance of each $Z_{G_l,m}$ is 1, and the covariances among the Z's are computed as,

$$\begin{aligned} cov(Z_{G_A,L}, Z_{G_A,H}) &= \frac{f^2 var(\bar{X}_{G_S,c}) + (1-f)^2 var(\bar{X}_{G_{S^-},c})}{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})} \\ &= \frac{f^2\frac{\sigma^2}{gN} + (1-f)^2\frac{\sigma^2}{(1-g)N}}{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})} = \frac{1}{2}, \end{aligned} \tag{3.7}$$

$$\begin{aligned} cov(Z_{G_A,L}, Z_{G_S,L}) &= \frac{fvar(\bar{X}_{G_S,L}) + fvar(\bar{X}_{G_S,c})}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \frac{f\frac{\sigma^2}{gN} + f\frac{\sigma^2}{gN}}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \sqrt{\frac{1}{1 + \frac{(1-f)^2 g}{f^2(1-g)}}}, \end{aligned} \tag{3.8}$$

$$\begin{aligned} cov(Z_{G_A,L}, Z_{G_S,H}) &= \frac{fvar(\bar{X}_{G_S,c})}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \frac{f\frac{\sigma^2}{gN}gN}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \frac{1}{2}\sqrt{\frac{1}{1 + \frac{(1-f)^2 g}{f^2(1-g)}}}, \end{aligned} \tag{3.9}$$

$$\begin{aligned} cov(Z_{G_A,H}, Z_{G_S,L}) &= \frac{fvar(\bar{X}_{S,c})}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \frac{f\frac{\sigma^2}{gN}gN}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \frac{1}{2}\sqrt{\frac{1}{1 + \frac{(1-f)^2 g}{f^2(1-g)}}}, \end{aligned} \tag{3.10}$$

$$\begin{aligned} cov(Z_{G_A,H}, Z_{G_S,H}) &= \frac{fvar(\bar{X}_{G_S,H}) + fvar(\bar{X}_{G_S,c})}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \frac{f\frac{\sigma^2}{gN} + f\frac{\sigma^2}{gN}}{\sqrt{\frac{2\sigma^2}{gN}}\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ &= \sqrt{\frac{1}{1 + \frac{(1-f)^2 g}{f^2(1-g)}}}, \end{aligned} \tag{3.11}$$

and

$$cov(Z_{G_S,L}, Z_{G_S,H}) = \frac{gN}{2\sigma^2} var(\bar{X}_{S,c}) = \frac{gN}{2\sigma^2}\frac{\sigma^2}{gN} = \frac{1}{2}. \tag{3.12}$$

Let $\mathbf{Z} = [Z_{G_A,L}, Z_{G_A,H}, Z_{G_S,L}, Z_{G_S,H}]'$, so that normality (see Section 3.1.2) yields that

$$\mathbf{Z} \sim \mathcal{MVN}(\mu_{\mathbf{Z}}, \mathbf{\Sigma}),$$

where

$$\mu_{\mathbf{Z}} = \begin{pmatrix} \frac{\Delta_{G_A,L}}{\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g}+\frac{(1-f)^2}{1-g})}} \\ \frac{\Delta_{G_A,H}}{\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g}+\frac{(1-f)^2}{1-g})}} \\ \frac{\Delta_{G_S,L}}{\sqrt{\frac{2\sigma^2}{gN}}} \\ \frac{\Delta_{G_S,H}}{\sqrt{\frac{2\sigma^2}{gN}}} \end{pmatrix}, \tag{3.13a}$$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \frac{1}{2} & D & \frac{D}{2} \\ \frac{1}{2} & 1 & \frac{D}{2} & D \\ D & \frac{D}{2} & 1 & \frac{1}{2} \\ \frac{D}{2} & D & \frac{1}{2} & 1 \end{pmatrix}, \tag{3.13b}$$

where $D = \sqrt{\frac{1}{1+\frac{(1-f)^2 g}{f^2(1-g)}}}$.

Under the null hypothesis $H_0^{(4)} : \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$,

$$\mathbf{Z} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}).$$

3.3.2 Follmann's test

In this section, we explore the testing method for each of the hypotheses in (3.3) and (3.4). Based on our testing scheme, the following hypothesis is the general format of each test:

$$H_0 : \Delta_i = 0 \ (i = 1, 2, \cdots, p),$$
$$H_a : \Delta_i \geq 0 \ (i = 1, 2, \cdots, p), \text{ with strict inequality for at least one value of } i,$$

where we observe a p-dimensional normal random vector for p population and dose combinations with mean $[\Delta_1, \cdots, \Delta_p]'$ and known covariance.

The techniques related to test these hypotheses have been extensively considered in the literature, which considers random samples $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n$ that are independent identically multivariate normal distributed with means $\vec{\nu}$ and covariance $\mathbf{\Psi}$. The parameter space Ω for $\vec{\nu}$ and $\mathbf{\Psi}$ is $\Omega = \{\nu_i \geq 0$ for every i, $\mathbf{\Psi}$ is positive definite$\}$, and the considered testing is $H_0 : \vec{\nu} = \mathbf{0}$ versus $H_a : \nu_i > 0$, for some i. For a univariate normal distribution, the hypothesis reduces to $H_0 : \nu = 0$ vs $H_a : \nu > 0$, which is a one-sided one sample Z test for known variance, and a one-sided one sample t test for unknown variance. In the multivariate cases, clearly the two sided Hotelling's T^2 test and the analogue χ^2 test for known $\mathbf{\Psi}$ are not appropriate for this one sided hypothesis [20] [42]. To handle the one sided problem, Kudo [32] derived a likelihood ratio test for the hypothesis with multivariate "one-sided" alternative assuming a known covariance matrix. Perlman [45] derived the corresponding likelihood ratio test assuming an unknown covariance. However, these likelihood ratio tests are theoretical and very difficult to evaluate for application purposes. Tang, Gnecco, and Geller [60] proposed an approximation to the likelihood ratio test, which is simpler but still practically difficult to implement.

Follmann [16] more recently proposed a simpler test (which we term Follmann's procedure), for the one-sided multivariate test given a multivariate normal population assuming known covariance matrix. He showed that his procedure protected the type I error rate, and had comparable power to the exact likelihood test. Follmann [16] also provided tight bounds on the power of his procedure, which makes calculating power possible. This makes Follmann's procedure very useful in our testing procedure.

Consider random samples $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n$ that are independent identically p-dimensional multivariate normal distributed with means $\vec{\nu}$ and covariance $\mathbf{\Psi}$. Let $\bar{V}_i$ be the average of all values in vector $\mathbf{V}_i$, where $i = 1, 2, ..., p$. Let $\bar{\mathbf{V}} \equiv (\bar{V}_1, \cdots, \bar{V}_p)'$. Follmann's test, when assuming $\mathbf{\Psi}$ is known, rejects at level α as long as a certain quadratic form of the sample mean vector exceeds a suitable 2α critical value and the sum of all the elements of the mean vector exceeds zero. Specifically, Follmann rejects H_0 if

$$n\bar{\mathbf{V}}'\mathbf{\Psi}^{-\mathbf{1}}\bar{\mathbf{V}} > \chi^2_{p,\,2\alpha} \quad \textit{and} \quad \sum_{i=1}^{p} \bar{V}_i > 0.$$

This test procedure is an analogue to a two sided multivariate likelihood ratio test with the average of the component means greater than zero. Follmann terms this a χ^2_+ test.

For each of the hypotheses, we apply Follmann's procedure to test our drug effects, and reject the null hypothesis if

$$\mathbf{Z}'\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{Z} > \chi^2_{p,\,2\alpha} \quad \textit{and} \quad \sum_{i=1}^{p} Z_i > 0.$$

where $\mathbf{Z}$ is given by (3.5) (3.6) and $\mathbf{\Sigma}$ by (3.13).

Again Follmann's procedure protects the type I error rate at α, so that each hypothesis in the testing scheme tested is a level α test. It now follows from Section 3.2.1 that our hypothesis testing scheme using Follmann's method strongly controls the type I error rate.

Now let's look at an example of how to apply Follmann's procedure for an individual null hypothesis. Suppose we are going to test $H_0^{(3)} : \Delta_{A,L} = \Delta_{A,H} = \Delta_{S,L} = 0$ versus $H_a^{(3)} : \Delta_{A,L} > 0 \; \textit{or} \; \Delta_{A,H} > 0 \; \textit{or} \; \Delta_{S,L} > 0$. From (3.13), we can determine the distribution of $\mathbf{Z_3}$, where $\mathbf{Z_3} = [Z_{A,L}, Z_{A,H}, Z_{S,L}]'$.

$$\mu_{\mathbf{Z_3}} = \begin{pmatrix} \frac{\Delta_{A,L}}{\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ \frac{\Delta_{A,H}}{\sqrt{\frac{2\sigma^2}{N}(\frac{f^2}{g} + \frac{(1-f)^2}{1-g})}} \\ \frac{\Delta_{S,L}}{\sqrt{\frac{2\sigma^2}{gN}}} \end{pmatrix}, \tag{3.15a}$$

$$\mathbf{\Sigma_3} = \begin{pmatrix} 1 & \frac{1}{2} & D \\ \frac{1}{2} & 1 & \frac{D}{2} \\ D & \frac{D}{2} & 1 \end{pmatrix}, \tag{3.15b}$$

where $D = \sqrt{\frac{1}{1+\frac{(1-f)^2 g}{f^2(1-g)}}}$.

Under the null hypothesis,

$$\mathbf{Z_3} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma_3}).$$

Therefore, we can directly apply Follmann's procedure based on the values of $\mathbf{Z_3}$ and its corresponding covariance matrix $\mathbf{\Sigma_3}$, which is to reject $H_0^{(3)}$ when $\mathbf{Z_3'}\mathbf{\Sigma_3^{-1}}\mathbf{Z_3} > \chi^2_{3,\, 2\alpha}$ and $Z_{A,L} + Z_{A,H} + Z_{S,L} > 0$. Similarly, we are able to apply Follmann's procedure for each individual hypothesis in our testing scheme.

3.4 SIMULATION STUDIES ON FINDING THE APPROPRIATE TEST ORDERING

When applying our proposed procedure to clinical trials, one needs to decide which ordering of the testing scheme to use before starting the trial. The choice of the ordering of the testing scheme depends on the goal of the study, and requires pilot data or prior beliefs of the drug effects of each dose on each population. We want to choose an ordering of the testing scheme so that we can achieve reasonably large probability (power) to conclude the results that are desired.

The power of our proposed procedure varies with regard to different study goals. In our setting, power is not a unique probability, but instead depends on what we want to achieve in a clinical trial. Furthermore, we could look at several powers at the same time. For example, consider the study goal where our primary goal is to find the largest population that has at least one dose effective, and our secondary goal is to find the MED for this population. In addition, suppose we strongly believe that the drug has positive drug effects on both the overall population and subpopulation for both high dose and low dose. Then we suggest that we have two powers to look at. The primary power we want to look at is the probability of concluding the overall population, i.e., $Power_1 = Prob(concluding \{G_A, L\}\ or\ \{G_A, H\})$, and the secondary power is the probability of concluding the low dose for the overall population, i.e., $Power_2 = Prob(concluding\{G_A, L\})$.

Clearly there is no general best ordering for any specific study goal. We need to find the most appropriate test ordering using simulation results based on our prior knowledge or beliefs of the drug effects, and to compare multiple powers at the same time in order to choose the test ordering. When more than one power needs to be evaluated, we want all the powers to be reasonably good.

When applying our proposed design in practice, simulation should be conducted before the study begins to obtain powers for the relevant subset of the 24 test orderings based on the prior knowledge or beliefs of the drug effects. Then the powers are compared among all test orderings, and the test ordering that leads to the desired powers is selected. In our proceeding example, we first compare the simulated primary powers and find the large powers. However, it is not certain that the test ordering with the largest primary power is most appropriate. Instead, we need to also compare the secondary powers. If several primary powers are comparable, then we want the secondary power to be large as well.

In this dissertation, we show in our simulation study that under the same study goal, different test ordering should be chosen for different drug effects. We also show that for the same scenario of drug effects, the appropriate test ordering depends on the study goal.

3.4.1 Selection of Test Ordering Depends on the Prior Beliefs of the Drug Effects

Suppose the primary goal of the sponsor is to find the largest population where there is at least one effective dose, and the secondary goal is to find the lowest effect dose for this largest population. We present simulation results under three scenarios of drug effects for this goal.

In the first scenario, the drug has positive effects for both populations and both doses, where the high dose has a larger effect size than the low dose and the drug effect size is larger for the subpopulation than the overall population. The drug effect sizes are listed in Table 4. Suppose $f = g = 0.4$, then $\boldsymbol{\Delta} = [\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H}]' = [0.22, 0.33, 0.4, 0.6]'$.

In the second scenario, only the high dose of the drug has positive effects for both populations, where the drug effect size of the high dose is larger for the subpopulation than the overall population and the low dose has little drug effects for both populations. The drug effect sizes are listed in Table 5. Again, suppose $f = g = 0.4$ as in the first scenario, then $\boldsymbol{\Delta} = [\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H}]' = [0.055, 0.33, 0.1, 0.6]'$.

In the third scenario, the drug has positive effects for both populations and both doses. However, the high dose has larger effect for the overall population than the subpopulation, while the low dose has larger effect for the subpopulation than the overall population. This scenario resembles the case where the subpopulation is more sensitive to the low dose of the drug. The drug effect sizes are listed in Table 6. Again, suppose $f = g = 0.4$ as in the first scenario, then $\boldsymbol{\Delta} = [\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H}]' = [0.32, 0.26, 0.5, 0.2]'$.

Table 4: Simulation Parameters - Design: $\boldsymbol{\Delta}$'s in Scenario 1

	Low	High
G_A	0.22	0.33
G_S	0.4	0.6
G_{S^-}	0.1	0.15

Table 5: Simulation Parameters - Design: $\boldsymbol{\Delta}$'s in Scenario 2

	Low	High
G_A	0.055	0.33
G_S	0.1	0.6
G_{S^-}	0.025	0.15

Simulations are conducted to obtain powers for all 24 test orderings based on each scenario of the drug effects. Each simulation study uses 10,000 iterations. The sample size chosen for each treatment arm is 200. The results are shown in Tables 7-9. The 24 orderings are listed in Table 3. The conclusions of all test orderings are listed in the five columns on the right of the table. For example, under scenario 1 (see Table 7), the five values on the right of the first row are the probabilities of concluding the corresponding dose and

Table 6: Simulation Parameters - Design: $\boldsymbol{\Delta}$'s in Scenario 3

	Low	High
G_A	0.32	0.26
G_S	0.5	0.2
G_{S^-}	0.2	0.3

population when using ordering 1 , i.e., the probability of concluding $\{G_A, L\}$ is 0.57; the probability of concluding $\{G_A, H\}$ is 0.26; the probability of concluding $\{G_S, L\}$ is 0.04; the probability of concluding $\{G_S, H\}$ is 0.05; and the probability of concluding no effective dose on any population is 0.08. The first four columns, "G_A, G_S, L, H", are the summary probabilities of concluding overall population (low dose and high dose), subpopulation (low dose and high dose), low dose (overall population and subpopulation) and high dose (overall population and subpopulation), respectively. These probabilities are calculated from the five columns on the right. For example, under Scenario 1 (see Table 7), the four values on the left of the first row are the probabilities of concluding the corresponding population or dose using ordering 1, i.e., the probability of concluding the overall population ("G_A") is 0.83, where $P(G_A) = P(\{G_A, L\}) + P(\{G_A, H\} = 0.57 + 0.26 = 0.83)$; the probability of concluding the subpopulation ("G_S") is 0.09, where $P(G_S) = P(\{G_S, L\}) + P(\{G_S, H\} = 0.04 + 0.05 = 0.09)$; the probability of concluding the low dose ("L") is 0.61, where $P(L) = P(\{G_A, L\}) + P(\{G_S, L\} = 0.57 + 0.04 = 0.61)$; and the probability of concluding the high dose ("H") is 0.31, where $P(H) = P(\{G_A, H\}) + P(\{G_S, H\} = 0.26 + 0.05 = 0.31)$.

The primary power we want to compare for Scenario 1 and Scenario 3 is the probability of concluding the overall population, i.e., $Power_1$ = Prob(G_A) = Prob(concluding $\{G_A, L\}$ *or* $\{G_A, H\}$), and the secondary power can be the probability of concluding the low dose for the overall population, i.e., $Power_2$ = Prob(concluding $\{G_A, L\}$). For Scenario 2, since the low dose has little effects on two populations, the primary goal is to conclude the high dose and the secondary goal is to conclude the overall population for high dose.

Table 7: Simulation Result: Scenario 1 ($\boldsymbol{\Delta} = [0.22, 0.33, 0.4, 0.6]'$)

ordering	G_A	G_S	L	H	G_A, L	G_A, H	G_S, L	G_S, H	None
1	0.83	0.09	0.61	0.31	0.57	0.26	0.04	0.05	0.08
2	0.83	0.09	0.05	0.87	0.01	0.82	0.04	0.05	0.08
3	0.75	0.17	0.64	0.28	0.53	0.23	0.12	0.05	0.08
4	0.26	0.66	0.64	0.28	0.04	0.23	0.61	0.05	0.08
5	0.85	0.08	0.03	0.89	0.01	0.84	0.02	0.05	0.08
6	0.18	0.74	0.69	0.23	0.01	0.18	0.68	0.05	0.08
7	0.84	0.08	0.58	0.34	0.58	0.27	0.01	0.07	0.08
8	0.84	0.08	0.01	0.91	0.01	0.84	0.01	0.07	0.08
9	0.59	0.33	0.59	0.33	0.58	0.01	0.01	0.32	0.08
10	0.01	0.91	0.01	0.91	0.00	0.01	0.01	0.90	0.08
11	0.87	0.05	0.01	0.91	0.00	0.87	0.01	0.04	0.08
12	0.01	0.91	0.01	0.91	0.00	0.01	0.01	0.90	0.08
13	0.54	0.38	0.65	0.27	0.53	0.01	0.12	0.26	0.08
14	0.05	0.87	0.65	0.27	0.04	0.01	0.61	0.26	0.08
15	0.60	0.32	0.59	0.33	0.58	0.01	0.00	0.32	0.08
16	0.02	0.90	0.01	0.91	0.00	0.01	0.00	0.90	0.08
17	0.02	0.90	0.70	0.22	0.01	0.01	0.69	0.21	0.08
18	0.02	0.90	0.01	0.91	0.01	0.01	0.00	0.90	0.08
19	0.86	0.06	0.03	0.89	0.00	0.86	0.03	0.03	0.08
20	0.20	0.73	0.70	0.22	0.00	0.19	0.69	0.03	0.08
21	0.87	0.05	0.01	0.91	0.00	0.87	0.00	0.04	0.08
22	0.01	0.91	0.01	0.91	0.00	0.01	0.00	0.91	0.08
23	0.01	0.91	0.70	0.22	0.00	0.01	0.70	0.21	0.08
24	0.01	0.91	0.01	0.91	0.00	0.01	0.00	0.90	0.08

Table 8: Simulation Result: Scenario 2 ($\mathbf{\Delta} = [0.055, 0.33, 0.1, 0.6]'$)

Ordering	G_A	L	G_S	H	G_A, L	G_A, H	G_S, L	G_S, H	None
1	0.80	0.09	0.14	0.85	0.08	0.73	0.01	0.13	0.06
2	0.80	0.02	0.14	0.92	0.01	0.80	0.01	0.13	0.06
3	0.78	0.08	0.16	0.86	0.05	0.74	0.03	0.13	0.06
4	0.76	0.08	0.18	0.86	0.02	0.74	0.05	0.13	0.06
5	0.81	0.03	0.13	0.91	0.03	0.78	0.00	0.13	0.06
6	0.73	0.12	0.21	0.82	0.03	0.70	0.08	0.13	0.06
7	0.87	0.09	0.07	0.85	0.08	0.79	0.01	0.06	0.06
8	0.87	0.02	0.07	0.92	0.01	0.85	0.01	0.06	0.06
9	0.10	0.09	0.84	0.85	0.08	0.02	0.01	0.84	0.06
10	0.02	0.01	0.92	0.93	0.00	0.02	0.01	0.91	0.06
11	0.89	0.02	0.05	0.92	0.01	0.88	0.01	0.04	0.06
12	0.02	0.02	0.92	0.92	0.01	0.01	0.01	0.91	0.06
13	0.07	0.08	0.87	0.86	0.05	0.03	0.03	0.84	0.06
14	0.05	0.08	0.89	0.86	0.02	0.03	0.05	0.84	0.06
15	0.10	0.09	0.84	0.85	0.08	0.03	0.01	0.82	0.06
16	0.03	0.01	0.91	0.93	0.00	0.03	0.01	0.90	0.06
17	0.03	0.09	0.91	0.85	0.00	0.03	0.09	0.83	0.06
18	0.03	0.01	0.91	0.93	0.00	0.03	0.00	0.91	0.06
19	0.83	0.01	0.11	0.93	0.01	0.82	0.00	0.11	0.06
20	0.75	0.09	0.19	0.85	0.01	0.74	0.08	0.11	0.06
21	0.88	0.02	0.06	0.92	0.01	0.88	0.01	0.04	0.06
22	0.01	0.02	0.93	0.92	0.01	0.01	0.01	0.92	0.06
23	0.02	0.09	0.92	0.85	0.01	0.01	0.09	0.84	0.06
24	0.02	0.01	0.92	0.93	0.01	0.01	0.01	0.92	0.06

Table 9: Simulation Result: Scenario 3 ($\mathbf{\Delta} = [0.32, 0.26, 0.5, 0.2]'$)

Ordering	G_A	G_S	L	H	G_A, L	G_A, H	G_S, L	G_S, H	None
1	0.83	0.08	0.85	0.06	0.81	0.02	0.04	0.04	0.09
2	0.83	0.08	0.17	0.74	0.13	0.70	0.04	0.04	0.09
3	0.84	0.07	0.85	0.06	0.82	0.02	0.03	0.04	0.09
4	0.06	0.85	0.85	0.06	0.03	0.02	0.81	0.04	0.09
5	0.72	0.19	0.17	0.74	0.02	0.70	0.15	0.04	0.09
6	0.05	0.86	0.83	0.08	0.02	0.04	0.82	0.04	0.09
7	0.81	0.10	0.88	0.03	0.79	0.02	0.09	0.01	0.09
8	0.81	0.10	0.21	0.70	0.12	0.68	0.09	0.01	0.09
9	0.81	0.09	0.86	0.05	0.77	0.04	0.09	0.00	0.09
10	0.59	0.32	0.64	0.27	0.55	0.04	0.09	0.23	0.09
11	0.81	0.10	0.28	0.63	0.19	0.61	0.09	0.01	0.09
12	0.60	0.31	0.28	0.63	0.19	0.41	0.09	0.22	0.09
13	0.87	0.04	0.83	0.07	0.81	0.06	0.03	0.01	0.09
14	0.09	0.81	0.83	0.07	0.03	0.06	0.80	0.01	0.09
15	0.84	0.06	0.84	0.07	0.78	0.06	0.06	0.01	0.09
16	0.62	0.29	0.61	0.29	0.56	0.06	0.06	0.23	0.09
17	0.13	0.78	0.84	0.07	0.06	0.06	0.78	0.01	0.09
18	0.13	0.78	0.62	0.29	0.06	0.06	0.55	0.23	0.09
19	0.72	0.19	0.17	0.74	0.01	0.71	0.16	0.03	0.09
20	0.05	0.86	0.84	0.07	0.01	0.04	0.83	0.03	0.09
21	0.63	0.28	0.27	0.64	0.01	0.62	0.26	0.01	0.09
22	0.42	0.48	0.27	0.64	0.01	0.42	0.26	0.22	0.09
23	0.11	0.80	0.80	0.11	0.01	0.10	0.79	0.01	0.09
24	0.11	0.80	0.58	0.33	0.01	0.10	0.57	0.23	0.09

Therefore, the primary power we want to look at for Scenario 2 is the probability of concluding $\{G_A, H\}$ or $\{G_S, H\}$, i.e., $Power_1 = \text{Prob}(H) = \text{Prob}(\text{concluding } \{G_A, H\} \textit{ or } \{G_S, H\})$. The secondary power is the probability of concluding the high dose on the overall population, i.e., $Power_2 = \text{Prob}(\text{concluding } \{G_A, H\})$. This makes sense because we want to get the drug approved on the largest population as possible.

Let's look at the simulation results for each scenario. For Scenario 1, when comparing the primary power, we find that ordering 21 has the largest primary power, while ordering 1, 2, 5, 7, 8, 11, 19 have comparably good primary power (marked in blue or red in Table 7). When comparing the secondary power, we find that ordering ordering 2, 5, 8, 11, 19 and 21 (marked in red in Table 7) all have very low secondary power, which means that it is more likely to conclude for $\{G_A, H\}$ in these orderings. Therefore, in this scenario, ordering 7 is the most appropriate testing ordering to achieve our study goal, while ordering 1 seems to be an appropriate testing ordering as well (marked in blue in Table 7). These test orderings make sense because there is reasonably large drug effects for both low dose and high dose on the overall population. As a consequence, there is enough power to reject $H_0^{(3)}$, $H_0^{(2)}$ and $H_0^{(1)}$ in orderings 1 and 7, and therefore lead to the conclusion of the overall population and low dose.

Consider the primary power and the secondary power for Scenario 2. The orderings that have comparably large primary power are marked in blue or red in Table 8. Some of these test orderings don't lead to reasonable secondary power (marked in red in Table 8). Ordering 11 has nice primary and secondary power, and it is considered the most appropriate test ordering for this scenario and study goal (marked in blue). Orderings 8, 19 and 21 are also considered appropriate testing orderings in this case (marked in blue in Table 8).

Finally, we look at both primary power and secondary power for Scenario 3. In this case, it is straight forward that ordering 13 is the most appropriate test ordering (marked in blue in Table 9). When there is relatively large drug effects for the low dose on the overall and subpopulation, the power is sufficiently large for rejecting $H_0^{(3)}$, $H_0^{(2)}$ and $H_0^{(1)}$ in ordering 13 and therefore lead to conclusion of the low dose for the overall population.

In summary, when the prior beliefs of the drug effects vary, the most appropriate test orderings vary even when the study goal is the same.

3.4.2 Selection of Test Ordering Depends on the Study Goal

Suppose there are severe adverse effects of the drug at higher doses. Furthermore, suppose we only consider drug effects as described in Scenario 1 (Section 3.4.1). Suppose that the study goal were to change. Then which test ordering should we choose? Consider the study goal where the primary goal is to find the lowest dose which is effective for at least one population, and the secondary goal is to find the largest population for that lowest dose. Then the primary power to be considered is the probability to conclude the low dose, i.e., Prob(“L”), and the secondary power to be considered is the probability to conclude the low dose and the overall population, i.e., Prob($\{G_A, L\}$), since the pharmaceutical company wants the drug get approved for the largest population. After comparing the primary power and the secondary power from Table 7, we find that ordering 13 is the most appropriate ordering with P(“L”)=0.65, and P($\{G_A, L\}$)=0.53.

We can see that under the same drug effects Scenario 1, when the goal of the studies vary, the most appropriate test orderings also vary even when the prior beliefs of the drug effects are the same.

The simulation studies are intended to emphasize that there is no general best ordering, and that simulation studies must be done in the design phase in order to find the best test ordering for that setting.

3.5 EXTENDING THE 2 POPULATION AND 2 DOSE CASE TO MULTIPLE POPULATION AND MULTIPLE DOSE CASES

Very often there are three doses considered in clinical trials: Low, Medium and High. But clearly there can be more than three doses. Also, there might be more than 2 levels of populations. Wang et al. [68] considered 3 levels of populations, where the populations are nested. Furthermore, Wang et al. [68] proposed enrichment designs assuming nested multiple populations. In our design, we also consider nested multiple populations.

Extending the 2 population and 2 dose case to multiple population and multiple dose cases is notationally complex, while the idea of the methodologies for the study design and data analysis are the same. In this section, we will show how to extend our 2 population and 2 dose case to 3 population and 3 dose case, so that the formulation of the methodology for even larger numbers of doses and populations could be extended similarly.

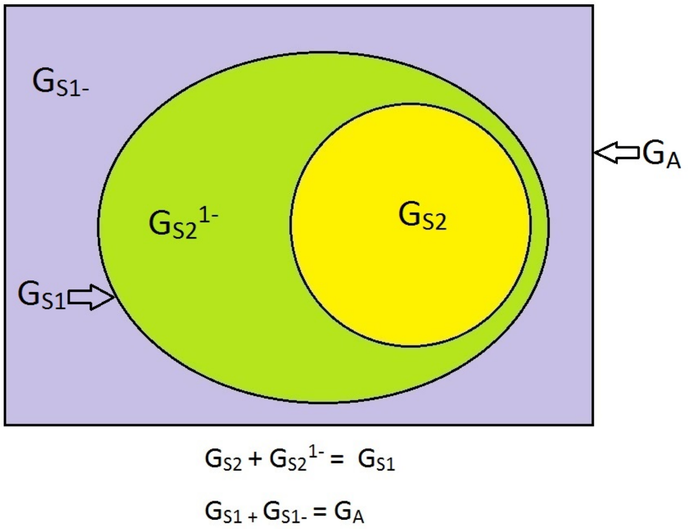
Consider two binary indicator I_1, I_2. The study subjects can be classified into one of the mutually exclusive subsets I_1^+ or I_1^-; or the mutually exclusive sets I_2^+ or I_2^-. In the study, we are interested in assessing the drug effects in three populations: the overall population, the subpopulation with I_1^+, and the subpopulation with I_1^+ and I_2^+. For example, the first classifier could be HER2 protein expression level (over-expression versus normal expression); and the second classifier could be subject's menopause status (during or post menopause versus before menopause). The populations of interest are the overall population, the subjects with HER2 over-expression, and the subjects who over-express HER2 and are during or after their menopause.

Following Wang et al.'s [68] notation, we denote the three populations of interest by G_A, G_{S1} and G_{S2}, where G_A is the overall population, G_{S1} is the subpopulation with I_1^+, and G_{S2} is the subpopulation with $I_1^+ \bigcap I_2^+$. Denote G_{S1^-} as the population with negative indicator I_1^-, which is the complimentary population of G_{S1}. The subpopulation G_{S1} and the compliment population G_{S1^-} are both nested in G_A, i.e., G_{S1^-} consists of all subjects that are contained in G_A but not contained in G_{S1}. Denote $G_{S2^{1-}}$ as the population with indicator $I_1^+ \bigcap I_2^-$ in G_{S1}, which is the complimentary population of G_{S2} within G_{S1} (Figure 2). Denote the populations by G_l, where $l = A, S1, S2, S1^-, S2^{1-}$.

Suppose the true proportion of G_{S1} out of G_A and the true proportion of G_{S2} out of G_{S1} are known, as described in Chapter 1, and are denoted by f_1 and f_2, respectively. Since G_{S1} and G_{S1^-} are mutually exclusive, the true proportion of G_{S1^-} out of G_A is $(1 - f_1)$. Since G_{S2} and $G_{S2^{1-}}$ are mutually exclusive in G_{S1}, the true proportion of G_{S2} out of G_A is $f_1 f_2$, and the true proportion of $G_{S2^{1-}}$ out of G_A is $f_1(1 - f_2)$.

Suppose in our study design, the proportion of subjects in the sample that will be selected from G_{S1} out of G_A is g_1, and the proportion of subjects in the sample that will be selected from G_{S2} out of G_{S1} is g_2, $0 < g_1, g_2 < 1$. Therefore, the proportion of subjects that will

Figure 2: Depiction of Three Nested Populations



be selected from G_{S1^-} is $1 - g_1$, the proportion of subjects that will be selected from G_{S2} is $g_1 g_2$, and the proportion of subjects that will be selected from $G_{S2^{1-}}$ is $g_1(1 - g_2)$.

When $g_1 = f_1$ and $g_2 = f_2$, this indicates that the sample has "true" proportion of subjects that are from each population; $g_1 > f_1$ indicates that the subpopulation with I^+ (G_{S1}) is partially enriched with respect to G_A; and $g_1 g_2 > f_1 f_2$ indicates that the subpopulation with I_1^+ and I_2^+ (G_{S2}) is partially enriched with respect to G_A and also G_{S1}; otherwise, there is no enrichment.

Suppose there are 3 dose levels of interest, denoted as Low, Medium and High. There is a control group as well, receiving placebo (or active comparator if we want to show superiority). Denote the treatments by m, where $m = L, M, H, c$.

Denote the total sample size for the study by N_{Total}. Thus, we have stratified sampling, i.e., $g_1 g_2 N_{Total}$ subjects are sampled from G_{S2}, $g_1(1-g_2)N_{Total}$ subjects sampled from $G_{S2^{1-}}$, and $(1 - g_1)N_{Total}$ subjects sampled from G_{S1^-}.

For simplicity, assume that subjects within a population are randomly assigned to high dose, low dose, or placebo with equal numbers to each treatment. Denote the total number of subjects receiving each treatment by N, where $N = N_{Total}/4$, so that $g_1 g_2 N$ subjects in

G_{S2} will be assigned to receive high dose, medium dose, low dose, and placebo, respectively; $g_1(1-g_2)N$ subjects in $G_{S2^{1-}}$ will be assigned to each dose group; similarly, $(1-g_1)N$ subjects in G_{S1^-} will be assigned to each dose group. Table 10 summarizes the experimental design with respect to the sample sizes for the stratified sampling.

Table 10: Study Design Sample Sizes: 3 by 3 Case

	Doses				
Population	Low	Medium	High	Control	Total
G_{S1^-}	$(1-g_1)N$	$(1-g_1)N$	$(1-g_1)N$	$(1-g_1)N$	$(1-g_1)N_{Total}$
$G_{S2^{1-}}$	$g_1(1-g_2)N$	$g_1(1-g_2)N$	$g_1(1-g_2)N$	$g_1(1-g_2)N$	$g_1(1-g_2)N_{Total}$
G_{S2}	g_1g_2N	g_1g_2N	g_1g_2N	g_1g_2N	$g_1g_2N_{Total}$
G_{S1}	g_1N	g_1N	g_1N	g_1N	g_1N_{Total}
G_A	N	N	N	N	N_{Total}

Note: N_{Total} is the total sample size; N is the sample size for each dose; g_1 is the sampling proportion of subjects from G_{S1} out of G_A; and g_2 is the sampling proportion of subjects from G_{S2} out of G_{S1}.

Denote the drug response of each subject in the study by $X_{G_l,m,i}$, where $l = S2, S2^{1-}, S1^-$; $m = L, M, H, c$; $i = 1, 2, \cdots, g_1g_2N$, if $l = S2$; $i = 1, 2, \cdots, g_1(1-g_2)N$, if $l = S2^{1-}$; $i = 1, 2, \cdots, (1-g_1)N$, if $l = S1^-$.

Denote by $\mu_{G_l,m}$ the true drug response mean for population l at dose m, where $l = A, S1, S1^-, S2, S2^{1-}$; $m = L, M, H, c$. Denote by $\Delta_{G_l,m}$ the true drug effect of dose m relative to the control group in population l ,

$$\Delta_{G_l,m} = \mu_{G_l,m} - \mu_{G_l,c},$$

where $l = A, S1, S1^-, S2, S2^{1-}$; $m = L, M, H$.

Like the 2 population and 2 dose case, we make the following assumptions:

1. All responses $X_{G_l,m,i}$ from the mutually exclusive populations are mutually independent. Thus the sample mean responses $\bar{X}_{G_{S1^-},L}$, $\bar{X}_{G_{S1^-},M}$, $\bar{X}_{G_{S1^-},H}$, $\bar{X}_{G_{S2^{1-}},L}$, $\bar{X}_{G_{S2^{1-}},M}$, $\bar{X}_{G_{S2^{1-}},H}$, $\bar{X}_{G_{S2},L}$, $\bar{X}_{G_{S2},M}$, and $\bar{X}_{G_{S2},H}$ are mutually independent.

2. $X_{G_l,m,i}$'s are normally distributed with mean $\mu_{l,m}$ and a common variance σ^2. Due to the generally sufficiently large sample sizes in each group, we continue to assume throughout this dissertation that the variance is known, that is,

$$X_{G_l,m,i} \sim \mathcal{N}(\mu_{G_l,m}, \sigma^2), \quad \sigma^2 \textit{ is known},$$

where $l = S2, S2^{1-}, S1^{-}$; $m = L, M, H, c$; $i = \begin{cases} 1, 2, \cdots, (1-g_1)N, & \text{if } \ l = S1^{-}; \\ 1, 2, \cdots, g_1(1-g_2)N, & \text{if } \ l = S2^{1-}; \\ 1, 2, \cdots, g_1 g_2 N, & \text{if } \ l = S2. \end{cases}$

We can easily obtain the distributions of sample mean responses for each treatment m, where $m = L, H, c$, for the subpopulation and the complimentary population, namely,

$$\bar{X}_{G_{S1-},m} \sim \mathcal{N}(\mu_{G_{S1-},m}, \frac{\sigma^2}{(1-g_1)N}),$$

$$\bar{X}_{G_{S2^{1-}},m} \sim \mathcal{N}(\mu_{G_{S2^{1-}},m}, \frac{\sigma^2}{g_1(1-g_2)N}),$$

$$\bar{X}_{G_{S2},m} \sim \mathcal{N}(\mu_{G_{S2},m}, \frac{\sigma^2}{g_1 g_2 N}).$$

3. We assume that the drug has nonnegative effects for all three doses, that is,

$$\Delta_{G_l,m} = \mu_{G_l,m} - \mu_{G_l,c} \geq 0,$$

where $l = A, S1, S2$; $m = L, M, H$.

4. As discussed in Section 1.1.1, we assume that the proportion f_1 of G_{S1} out of G_A, and the proportion f_2 of G_{S2} out of G_{S1} are known. Therefore, we have that for $m = L, M, H, c$, the population mean response for G_{S1} is

$$\mu_{G_{S1},m} = f_2 \mu_{G_{S2},m} + (1-f_2)\mu_{G_{S2^{1-}},m},$$

and the population mean response for G_A is

$$\begin{aligned} \mu_{G_A,m} &= f_1 \mu_{G_{S1},m} + (1-f_1)\mu_{G_{S1^-},m} \\ &= f_1[f_2 \mu_{G_{S2},m} + (1-f_2)\mu_{G_{S2^{1-}},m}] + (1-f_1)\mu_{G_{S1^-},m}. \end{aligned}$$

For $m = L, M, H$, the drug effect for G_{S1} is

$$\Delta_{G_{S1},m} = f_2\Delta_{G_{S2},m} + (1 - f_2)\Delta_{G_{S2^{1-}},m},$$

and the drug effect for G_A is

$$\Delta_{G_A,m} =f_1 f_2\Delta_{G_{S2},m} + f_1(1 - f_2)\Delta_{G_{S2^{1-}},m} + (1 - f_1)\Delta_{G_{S1^{-}},m}.$$

When there are three nested populations and three doses in a clinical trial, we are interested in evaluating the drug effects on nine groups: $\{G_A$, Low$\}$, $\{G_A$, Medium$\}$, $\{G_A$, High$\}$, $\{G_{S1}$, Low$\}$, $\{G_{S1}$, Medium$\}$, $\{G_{S1}$, High$\}$, $\{G_{S2}$, Low$\}$, $\{G_{S2}$, Medium$\}$, $\{G_{S2}$, High$\}$, which are low dose in the overall population, medium dose in the overall population, high dose in the overall population, low dose in the population with I_1^+, medium dose in the population with I_1^+, high dose in the population with I_1^+, low dose in the population with $I_1^+ \bigcap I_2^+$, medium dose in the population with $I_1^+ \bigcap I_2^+$, high dose in the population with $I_1^+ \bigcap I_2^+$, respectively. Thus we want to construct a testing procedure that tests for the drug effects $\mathbf{\Delta}$, where $\mathbf{\Delta} = (\Delta_{G_A,L}, \Delta_{G_A,M}, \Delta_{G_A,H}, \Delta_{G_{S1},L}, \Delta_{G_{S1},M}, \Delta_{G_{S1},H}, \Delta_{G_{S2},L}, \Delta_{G_{S2},M}, \Delta_{G_{S2},H})'$ and where the testing scheme meets the goals of the company as described in Section 2.1.

Again, we construct the testing scheme under the principle of closed testing procedures. For the three nested population and three dose case, our testing scheme consists of 9 individual hypotheses, and therefore there are $9! = 362,880$ in the analogue of the testing scheme in Table 3 for our 3×3 case. We do illustrate one testing scheme as an example. Due to huge possible test orderings, we do not perform extensive comparative simulations for 3 population and 3 dose case as we did for Table 7, 8, 9.

Suppose in a clinical trial, our goal is to find the largest population where there is at least one effective dose and for this population find the Minimum Effect Dose (MED). To achieve this goal, we consider the following testing scheme where the null hypotheses are as follows

$$
\begin{aligned}
H_0^{(9)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} \\
&= \Delta_{G_{S2},M} = \Delta_{G_{S2},H} = 0, && \text{(3.19a)} \\
H_0^{(8)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} \\
&= \Delta_{G_{S2},M} = 0, && \text{(3.19b)} \\
H_0^{(7)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} = 0, && \text{(3.19c)} \\
H_0^{(6)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = 0, && \text{(3.19d)} \\
H_0^{(5)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = 0, && \text{(3.19e)} \\
H_0^{(4)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = 0, && \text{(3.19f)} \\
H_0^{(3)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = 0, && \text{(3.19g)} \\
H_0^{(2)} :&\Delta_{G_A,L} = \Delta_{G_A,M} = 0, && \text{(3.19h)} \\
H_0^{(1)} :&\Delta_{G_A,L} = 0, && \text{(3.19i)}
\end{aligned}
$$

and where Δ_{lm} is given previously in this section.

Each corresponding alternative hypothesis states that there is at least one positive drug effect among the populations and doses considered in the null hypothesis:

$$
\begin{aligned}
H_a^{(9)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0 \text{ or } \Delta_{G_{S1},L} > 0 \text{ or } \Delta_{G_{S1},M} > 0 \\
& \text{or } \Delta_{G_{S1},H} > 0 \text{ or } \Delta_{G_{S2},L} > 0 \text{ or } \Delta_{G_{S2},M} > 0 \text{ or } \Delta_{G_{S2},H} > 0, && (3.20a) \\
H_a^{(8)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0 \text{ or } \Delta_{G_{S1},L} > 0 \text{ or } \Delta_{G_{S1},M} > 0 \\
& \text{or } \Delta_{G_{S1},H} > 0 \text{ or } \Delta_{G_{S2},L} > 0 \text{ or } \Delta_{G_{S2},M} > 0, && (3.20b) \\
H_a^{(7)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0 \text{ or } \Delta_{G_{S1},L} > 0 \text{ or } \Delta_{G_{S1},M} > 0 \\
& \text{or } \Delta_{G_{S1},H} > 0 \text{ or } \Delta_{G_{S2},L} > 0, && (3.20c) \\
H_a^{(6)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0 \text{ or } \Delta_{G_{S1},L} > 0 \text{ or } \Delta_{G_{S1},M} > 0 \\
& \text{or } \Delta_{G_{S1},H} > 0, && (3.20d) \\
H_a^{(5)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0 \text{ or } \Delta_{G_{S1},L} > 0 \text{ or } \Delta_{G_{S1},M} > 0, && (3.20e) \\
H_a^{(4)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0 \text{ or } \Delta_{G_{S1},L} > 0, && (3.20f) \\
H_a^{(3)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0 \text{ or } \Delta_{G_A,H} > 0, && (3.20g) \\
H_a^{(2)} :& \Delta_{G_A,L} > 0 \text{ or } \Delta_{G_A,M} > 0, && (3.20h) \\
H_a^{(1)} :& \Delta_{G_A,L} > 0. && (3.20i)
\end{aligned}
$$

We test the family of hypotheses in a step down manner from the highest level until we accept one null hypothesis. To be specific, $H^{(9)}$ is tested first. If $H_0^{(9)}$ is rejected, then $H^{(8)}$ is to be tested next; otherwise, stop testing. If $H_0^{(8)}$ is rejected, then continue to test $H^{(7)}$; otherwise, stop testing. Similarly, we continue each individual test in a step down manner until one hypothesis is accepted or $H^{(1)}$ is rejected. Controlling each individual test at level α yields overall strong control of type I error at the α-level.

After the hypotheses are tested, we make the conclusions based on the testing results according to the following rules.

- If $H_0^{(9)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} = \Delta_{G_{S2},M} = \Delta_{G_{S2},H} = 0$, then there is no population that has an effective dose.

- If $H_0^{(9)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$ or $\Delta_{G_{S1},L} > 0$ or $\Delta_{G_{S1},M} > 0$ or $\Delta_{G_{S1},H} > 0$ or $\Delta_{G_{S2},L} > 0$ or $\Delta_{G_{S2},M} > 0$ or $\Delta_{G_{S2},H} > 0$, and if $H_0^{(8)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} = \Delta_{G_2,M} = 0$, then we conclude $\Delta_{G_2,H} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation G_2 and the MED for this population is the high dose. We denote this choice by $\{G_2$, High$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$ or $\Delta_{G_{S1},L} > 0$ or $\Delta_{G_{S1},M} > 0$ or $\Delta_{G_{S1},H} > 0$ or $\Delta_{G_{S2},L} > 0$ or $\Delta_{G_{S2},M} > 0$, and if $H_0^{(7)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} = 0$, then we conclude $\Delta_{G_{S2},M} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation G_{S2} and the MED for this population is the medium dose. We denote this choice by $\{G_{S2}$, Medium$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, and $H_0^{(7)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$ or $\Delta_{G_{S1},L} > 0$ or $\Delta_{G_{S1},M} > 0$ or $\Delta_{G_{S1},H} > 0$ or $\Delta_{G_{S2},L} > 0$, and if $H_0^{(6)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = \Delta_{G_{S1},H} = 0$, then we conclude $\Delta_{G_{S2},L} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation G_{S2} and the MED for this population is the low dose. We denote this choice by $\{G_{S2}$, Low$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, $H_0^{(7)}$ is false, and $H_0^{(6)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$ or $\Delta_{G_{S1},L} > 0$ or $\Delta_{G_{S1},M} > 0$ or $\Delta_{G_{S1},H} > 0$, and if $H_0^{(5)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},M} = 0$, then we conclude $\Delta_{G_{S1},H} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$ and/or $\Delta_{G_{S2},L} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation G_{S1} and the MED for this population is the high dose. We denote this choice by $\{G_{S1}$, High$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, $H_0^{(7)}$ is false, $H_0^{(6)}$ is false, and $H_0^{(5)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$ or $\Delta_{G_{S1},L} > 0$ or $\Delta_{G_{S1},M} > 0$, and if $H_0^{(4)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = 0$, then we conclude $\Delta_{G_{S1},M} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$ and/or $\Delta_{G_{S2},L} > 0$ and/or $\Delta_{G_{S1},H} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation G_{S1} and the MED for this population is the medium dose. We

denote this choice by $\{G_{S1}$, Medium$\}$.

- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, $H_0^{(7)}$ is false, $H_0^{(6)}$ is false, $H_0^{(5)}$ is false, and $H_0^{(4)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$ or $\Delta_{G_{S1},L} > 0$, and if $H_0^{(3)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = \Delta_{G_A,H} = 0$, then we conclude $\Delta_{G_{S1},L} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$ and/or $\Delta_{G_{S2},L} > 0$ and/or $\Delta_{G_{S1},H} > 0$ and/or $\Delta_{G_{S1},M} > 0$. We thereby establish the largest population that has a positive dose effect is the subpopulation G_{S1} and the MED for this population is the low dose. We denote this choice by $\{G_{S1}$, Low$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, $H_0^{(7)}$ is false, $H_0^{(6)}$ is false, $H_0^{(5)}$ is false, $H_0^{(4)}$ is false, and $H_0^{(3)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$ or $\Delta_{G_A,H} > 0$, and if $H_0^{(2)}$ is true, i.e., $\Delta_{G_A,L} = \Delta_{G_A,M} = 0$, then we conclude $\Delta_{G_A,H} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$ and/or $\Delta_{G_{S2},L} > 0$ and/or $\Delta_{G_{S1},H} > 0$ and/or $\Delta_{G_{S1},M} > 0$ and/or $\Delta_{G_{S1},L} > 0$. We thereby establish the largest population that has a positive dose effect is the overall population G_A and the MED for this population is the high dose. We denote this choice by $\{G_A$, High$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, $H_0^{(7)}$ is false, $H_0^{(6)}$ is false, $H_0^{(5)}$ is false, $H_0^{(4)}$ is false, $H_0^{(3)}$ is false, and $H_0^{(2)}$ is false, i.e., $\Delta_{G_A,L} > 0$ or $\Delta_{G_A,M} > 0$, and if $H_0^{(1)}$ is true, i.e., $\Delta_{G_A,L} = 0$, then we conclude $\Delta_{G_A,M} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$ and/or $\Delta_{G_{S2},L} > 0$ and/or $\Delta_{G_{S1},H} > 0$ and/or $\Delta_{G_{S1},M} > 0$ and/or $\Delta_{G_{S1},L} > 0$ and/or $\Delta_{G_A,H} > 0$. We thereby establish the largest population that has a positive dose effect is the overall population G_A and the MED for this population is the medium dose. We denote this choice by $\{G_A$, Medium$\}$.
- If $H_0^{(9)}$ is false, $H_0^{(8)}$ is false, $H_0^{(7)}$ is false, $H_0^{(6)}$ is false, $H_0^{(5)}$ is false, $H_0^{(4)}$ is false, $H_0^{(3)}$ is false, $H_0^{(2)}$ is false, and $H_0^{(1)}$ is false, i.e., $\Delta_{G_A,L} > 0$, then we conclude $\Delta_{G_A,L} > 0$, where it is possible that also $\Delta_{G_{S2},H} > 0$ and/or $\Delta_{G_{S2},M} > 0$ and/or $\Delta_{G_{S2},L} > 0$ and/or $\Delta_{G_{S1},H} > 0$ and/or $\Delta_{G_{S1},M} > 0$ and/or $\Delta_{G_{S1},L} > 0$ and/or $\Delta_{G_A,H} > 0$ and/or $\Delta_{G_A,M} > 0$. We thereby establish the largest population that has a positive dose effect is the overall population G_A and the MED for this population is the low dose. We denote this choice by $\{G_A$, Low$\}$.

For this illustration, Table 11 summarizes all the possible outcomes and the corresponding conclusions.

Similarly as for the 2 population and 2 dose cases, we can obtain unbiased estimators of the drug effects, as well as the mean and covariance structure of the test statistics (see Appendix A.1 for details). To test each individual hypothesis, we can again directly apply Follmann's procedure. Based on the test results and Table 11, we can make conclusions of the study.

Table 11: Outcomes of the Procedure and Decision Rule for Section 3.5

$H_0^{(9)}$	$H_0^{(8)}$	$H_0^{(7)}$	$H_0^{(6)}$	$H_0^{(5)}$	$H_0^{(4)}$	$H_0^{(3)}$	$H_0^{(2)}$	$H_0^{(1)}$	$\Delta_{G_A,L}$	$\Delta_{G_A,M}$	$\Delta_{G_A,H}$	$\Delta_{G_{S1},L}$	$\Delta_{G_{S1},M}$	$\Delta_{G_{S1},H}$	$\Delta_{G_{S2},L}$	$\Delta_{G_{S2},M}$	$\Delta_{G_{S2},H}$	Conclusion
ACC	NT	NT	NT	NT	NT	NT	NT	NT	0	0	0	0	0	0	0	0	0	Empty
REJ	ACC	NT	NT	NT	NT	NT	NT	NT	0	0	0	0	0	0	0	0	> 0	$\{G_{S2}$, High$\}$
REJ	REJ	ACC	NT	NT	NT	NT	NT	NT	0	0	0	0	0	0	0	> 0	UKN	$\{G_{S2}$, Medium$\}$
REJ	REJ	REJ	ACC	NT	NT	NT	NT	NT	0	0	0	0	0	0	> 0	UKN	UKN	$\{G_{S2}$, Low$\}$
REJ	REJ	REJ	REJ	ACC	NT	NT	NT	NT	0	0	0	0	0	> 0	UKN	UKN	UKN	$\{G_{S1}$, High$\}$
REJ	REJ	REJ	REJ	REJ	ACC	NT	NT	NT	0	0	0	0	> 0	UKN	UKN	UKN	UKN	$\{G_{S1}$, Medium$\}$
REJ	REJ	REJ	REJ	REJ	REJ	ACC	NT	NT	0	0	0	> 0	UKN	UKN	UKN	UKN	UKN	$\{G_{S1}$, Low$\}$
REJ	REJ	REJ	REJ	REJ	REJ	REJ	ACC	NT	0	0	> 0	UKN	UKN	UKN	UKN	UKN	UKN	$\{G_A$, High$\}$
REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	ACC	0	> 0	UKN	UKN	UKN	UKN	UKN	UKN	UKN	$\{G_A$, Medium$\}$
REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	> 0	UKN	UKN	UKN	UKN	UKN	UKN	UKN	UKN	$\{G_A$, Low$\}$

Note: ACC=Accept, REJ=Reject, NT=Not Tested, UKN=Unknown.

4.0 ADAPTIVE DESIGNS

In this chapter, we propose two two-stage adaptive designs extending the non-adaptive designs in Chapter 3. In the first adaptive design, only the total sample size of the second stage depends on the first stage data. In the second adaptive design, we may further enrich the subpopulation, as well as increase the sample size, for the second stage.

For illustrative purpose, we again use use the 2 population and 2 dose case.

4.1 ADAPTIVE DESIGN WITH SECOND STAGE SAMPLE SIZE ADJUSTMENT

4.1.1 The Adaptive Design

Before the study, the testing scheme and the hypotheses test ordering should be determined based on the study goal and prior beliefs of the drug effects, as in Section 3.4. We pre-specify the first stage and second stage sampling proportions of subjects from the subpopulation. In the first stage, we are going to sample N^{I}_{Total} subjects, where N^{I}_{Total} is pre-determined. After Stage I, we examine the first stage data, and the test statistics of the first stage data are calculated. The statistics of stage I are denoted by $\mathbf{Z}^{I}$, where $\mathbf{Z}^{I} = [Z^{I}_{G_A,L}, Z^{I}_{G_A,H}, Z^{I}_{G_S,L}, Z^{I}_{G_S,H}]'$ and they are specified in (3.5) and (3.6). We choose the second stage sample size based on $\mathbf{Z}^{I}$. In the second stage, we are going to sample $N^{II}_{Total}(\mathbf{Z}^{I})$ subjects, where $N^{II}_{Total}(\mathbf{Z}^{I})$ is obtained by the pre-specified adaptation rule given at the beginning of the study. How to obtain $N^{II}_{Total}(\mathbf{Z}^{I})$ is discussed in Section 4.1.4. The statistics of stage II are denoted by $\mathbf{Z}^{II}$, where $\mathbf{Z}^{II} = [Z^{II}_{G_A,L}, Z^{II}_{G_A,H}, Z^{II}_{G_S,L}, Z^{II}_{G_S,H}]'$. Also we pre-specify a weight, denoted by w, where $0 < w < 1$, which is used to combine the information from the two stages.

Similar to the non-adaptive design proposed in Chapter 3, we make the following assumptions.

1. The true proportion of the subpopulation to the overall population is known, denoted as f.
2. We assume that the responses $X_{G_l,m,i}$ are normally distributed with mean $\mu_{G_l,m}$ and a common variance σ^2. Assume that the variance is known.
3. We assume that the drug has nonnegative effect for both doses.
4. For simplicity, we assume that subjects within a population are randomly assigned to high dose, low dose, or placebo with equal numbers in each.

Denote the sampling proportion of subjects from the subpopulation in Stage I by g^I, and let $N^I = N^I_{Total}/3$. Again both g^I and N^I_{Total} are determined before the trial starts. The design of the first stage is the same as the non-adaptive design in Section 3.1.1. Stratified sampling is used with $g^I N^I$ subjects from the subpopulation G_S being assigned for each treatment (low dose, high dose, and control), and $(1-g^I)N^I$ subjects form the complimentary population G_{S^-} being assigned for each treatment (Table 12).

Table 12: First Stage Adaptive Design Sample Sizes: 2 by 2 Case

	Doses			
	Low	High	Control	Total
Subpopulation G_S	$g^I N^I$	$g^I N^I$	$g^I N^I$	gN^I_{Total}
Complimentary Pop. G_{S^-}	$(1-g^I)N^I$	$(1-g^I)N^I$	$(1-g^I)N^I$	$(1-g)N^I_{Total}$
Overall Pop. G_A	N^I	N^I	N^I	N^I_{Total}

Note: g^I is the first stage sampling proportion of subjects from the subpopulation, N^I_{Total} is the Stage I total sample size, and N^I is the Stage I sample size for each dose.

After completion of Stage I experiment, we examine the stage I data, and based on the information we obtained from stage I, we decide the sample size for stage II: $N^{II}_{Total}(\mathbf{Z}^I)$, according to the pre-specified adaptation rule. Let $N^{II}(\mathbf{Z}^I) = N^{II}_{Total}(\mathbf{Z}^I)/3$. Denote the sampling proportion of subjects selected from the subpopulation in Stage II by g^{II}, where

g^{II} is also pre-specified prior to the start of the study. Usually, we want to make g^{II} equal to g^{I}. However, we want to use different notation in this dissertation in order to point out that the second stage sampling proportion could be different from the first stage sampling proportion. Stratified sampling is used again, and the design of the second stage is shown in Table 13.

Table 13: Second Stage Adaptive Design Sample Sizes: 2 by 2 Case

	Doses			
	Low	High	Control	Total
Subpopulation G_S	$g^{II}N^{II}(\mathbf{Z}^I)$	$g^{II}N^{II}(\mathbf{Z}^I)$	$g^{II}N^{II}(\mathbf{Z}^I)$	$gN^{II}_{Total}(\mathbf{Z}^I)$
Complimentary Pop. G_{S^-}	$(1-g^{II})N^{II}(\mathbf{Z}^I)$	$(1-g^{II})N^{II}(\mathbf{Z}^I)$	$(1-g^{II})N^{II}(\mathbf{Z}^I)$	$(1-g)N^{II}_{Total}(\mathbf{Z}^I)$
Overall Pop. G_A	$N^{II}(\mathbf{Z}^I)$	$N^{II}(\mathbf{Z}^I)$	$N^{II}(\mathbf{Z}^I)$	$N^{II}_{Total}(\mathbf{Z}^I)$

Note: g^{II} is the second stage sampling proportion of subjects from the subpopulation, $N^{II}_{Total}(\mathbf{Z}^I)$ is the total Stage II sample size, and $N^{II}(\mathbf{Z}^I)$ is the Stage II sample size for each dose.

At the end of each stage, we collect the response data for Stage I and Stage II respectively, and we compute $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$, i.e., $\mathbf{Z}^I = [Z^I_{G_A,L}, Z^I_{G_A,H}, Z^I_{G_S,L}, Z^I_{G_S,H}]'$ and $\mathbf{Z}^{II} = [Z^{II}_{G_A,L}, Z^{II}_{G_A,H}, Z^{II}_{G_S,L}, Z^{II}_{G_S,H}]'$, using equations (3.5) and (3.6) for Stage I and Stage II data. We then use these data to obtain the statistics for the final analysis.

4.1.2 The Final Statistics

It follows from (3.5) and (3.6) $\mathbf{Z}^I$ has a multivariate normal distribution:

$$\mathbf{Z}^I \sim \mathcal{MVN}(\mu_{\mathbf{Z}^I}, \begin{pmatrix} 1 & \frac{1}{2} & D^I & \frac{D^I}{2} \\ \frac{1}{2} & 1 & \frac{D^I}{2} & D^I \\ D^I & \frac{D^I}{2} & 1 & \frac{1}{2} \\ \frac{D^I}{2} & D^I & \frac{1}{2} & 1 \end{pmatrix}), \tag{4.1}$$

where

$$\mu_{\mathbf{Z}^I} = \begin{pmatrix} \frac{\Delta_{G_A,L}}{\sqrt{\frac{2\sigma^2}{N^I}(\frac{f^2}{g^I}+\frac{(1-f)^2}{1-g^I})}} \\ \frac{\Delta_{G_A,H}}{\sqrt{\frac{2\sigma^2}{N^I}(\frac{f^2}{g^I}+\frac{(1-f)^2}{1-g^I})}} \\ \frac{\Delta_{G_S,L}}{\sqrt{\frac{2\sigma^2}{g^I N^I}}} \\ \frac{\Delta_{G_S,H}}{\sqrt{\frac{2\sigma^2}{g^I N^I}}} \end{pmatrix}, \tag{4.2}$$

$$D^I = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^I}{f^2(1-g^I)}}}. \tag{4.3}$$

The distribution of the second stage test statistics $\mathbf{Z}^{II}$ depends on the second stage sample size $N^{II}_{Total}(\mathbf{Z}^I)$ and the second stage sample size depends on the first stage data through $\mathbf{Z}^I$. Consequently, the distribution of $\mathbf{Z}^{II}$ depends on the first stage data through $\mathbf{Z}^I$. We calculate the distribution of $\mathbf{Z}^{II}|\mathbf{Z}^I$, as in equations (3.5), (3.6) and (3.7-3.12). Conditional on $\mathbf{Z}^I$, the responses are normally distributed with known covariance, and the sample size for the second stage is known. Therefore, $\mathbf{Z}^{II}|\mathbf{Z}^I$ have a multivariate normal distribution:

$$\mathbf{Z}^{II}|\mathbf{Z}^I \sim \mathcal{MVN}(\mu_{\mathbf{Z}^{II}|\mathbf{Z}^I}, \begin{pmatrix} 1 & \frac{1}{2} & D^{II} & \frac{D^{II}}{2} \\ \frac{1}{2} & 1 & \frac{D^{II}}{2} & D^{II} \\ D^{II} & \frac{D^{II}}{2} & 1 & \frac{1}{2} \\ \frac{D^{II}}{2} & D^{II} & \frac{1}{2} & 1 \end{pmatrix}), \tag{4.4}$$

where

$$\mu_{\mathbf{Z}^{II}|\mathbf{Z}^I} = \begin{pmatrix} \frac{\Delta_{G_A,L}}{\sqrt{\frac{2\sigma^2}{N^{II}(\mathbf{Z}^I)}(\frac{f^2}{g^{II}}+\frac{(1-f)^2}{1-g^{II}})}} \\ \frac{\Delta_{G_A,H}}{\sqrt{\frac{2\sigma^2}{N^{II}(\mathbf{Z}^I)}(\frac{f^2}{g^{II}}+\frac{(1-f)^2}{1-g^{II}})}} \\ \frac{\Delta_{G_S,L}}{\sqrt{\frac{2\sigma^2}{g^{II}N^{II}(\mathbf{Z}^I)}}} \\ \frac{\Delta_{G_S,H}}{\sqrt{\frac{2\sigma^2}{g^{II}N^{II}(\mathbf{Z}^I)}}} \end{pmatrix}, \tag{4.5}$$

$$D^{II} = \sqrt{\frac{1}{1 + \frac{(1-f)^2 g^{II}}{f^2(1-g^{II})}}}. \tag{4.6}$$

So far, we have obtained the distribution of $\mathbf{Z}^I$ and the distribution of $\mathbf{Z}^{II}$ conditional on $\mathbf{Z}^I$. It is important to note that $\mu_{\mathbf{Z}^{II}|\mathbf{Z}^I}$ depends on $\mathbf{Z}^I$ and $cov(\mathbf{Z}^{II}|\mathbf{Z}^I)$ does not depend on $\mathbf{Z}^I$ since D^{II} in (4.6) does not depend on $\mathbf{Z}^I$. Next we consider combining the two stages of data to obtain the final statistics.

The final statistics $\mathbf{Z}^{Final}$, where $\mathbf{Z}^{Final} = [Z^{Final}_{G_A,L}, Z^{Final}_{G_A,H}, Z^{Final}_{G_S,L}, Z^{Final}_{G_S,H}]'$, are calculated by combining the two stage data using this fixed weight w:

$$Z^{Final}_{G_l,m} = \sqrt{w} Z^I_{G_l,m} + \sqrt{(1-w)} Z^{II}_{G_l,m}, \tag{4.7}$$

where $l = A, S$, $m = L, H$, and where $0 < w < 1$ is the pre-specified weight.

We show in the next subsection that under the null hypotheses, $\mathbf{Z}^{Final}$ and all subsets of $\mathbf{Z}^{Final}$ are multivariate normally distributed. This property makes the critical values for Follmann's test easy to compute. However, under the alternatives, neither the distribution of $\mathbf{Z}^{Final}$ nor the distribution of any subset of $\mathbf{Z}^{Final}$ is multivariate normal. Instead, they are mixtures of multivariate normal distributions.

4.1.3 Final Analysis and Strong Control of Type I Error

For the non-adaptive designs, a testing scheme is given in Section 3.2, which is closed under intersection. Since that testing scheme follows the closed testing procedure, the test strongly controls Type I error rate, if each individual hypothesis is tested at α level. Furthermore, Follmann's test, which is a one-sided multivariate test, can be used to test each individual hypothesis, which is discussed in Section 3.3.

In our adaptive design, we similarly pre-specify the testing scheme and the ordering before the study begins, as in the non-adaptive design. Assuming we demonstrate again we are using a closed testing scheme, then if we could test each individual hypothesis at level α following the testing order at the end of the study, the Type I error rate would be strongly protected.

As we have discussed in the Section 3.2, each testing scheme includes four hypotheses tests. Since the hypothesis $H_0^{(4)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$ is included in all testing schemes, let's first take a look at the distribution of $\mathbf{Z}^{Final}$ under $H_0^{(4)}$.

Under $H_0^{(4)}$, the distributions of $\mathbf{Z}^I$ and $\mathbf{Z}^{II}|\mathbf{Z}^I$ become:

$$\mathbf{Z}^I \sim \mathcal{MVN}(\mathbf{0}, \begin{pmatrix} 1 & \frac{1}{2} & D^I & \frac{D^I}{2} \\ \frac{1}{2} & 1 & \frac{D^I}{2} & D^I \\ D^I & \frac{D^I}{2} & 1 & \frac{1}{2} \\ \frac{D^I}{2} & D^I & \frac{1}{2} & 1 \end{pmatrix}),$$

$$\mathbf{Z}^{II}|\mathbf{Z}^I \sim \mathcal{MVN}(\mathbf{0}, \begin{pmatrix} 1 & \frac{1}{2} & D^{II} & \frac{D^{II}}{2} \\ \frac{1}{2} & 1 & \frac{D^{II}}{2} & D^{II} \\ D^{II} & \frac{D^{II}}{2} & 1 & \frac{1}{2} \\ \frac{D^{II}}{2} & D^{II} & \frac{1}{2} & 1 \end{pmatrix}),$$

where $D^I = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^I}{f^2(1-g^I)}}}$ and $D^{II} = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^{II}}{f^2(1-g^{II})}}}$, as defined in (4.3) and (4.6).

It is obvious that under the null hypotheses $H_0^{(4)}$, $\mathbf{Z}^{II}$ is independent of $\mathbf{Z}^I$, so that $\mathbf{Z}^{II}$ is multivariate normally distributed unconditionally. Since $\mathbf{Z}^{Final}$ is a linear combination of $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$, which are independent and both multivariate normally distributed under $H_0^{(4)}$, and the weight w is fixed, $\mathbf{Z}^{Final}$ also has a multivariate normal distribution under $H_0^{(4)}$:

$$\mathbf{Z}^{Final} \sim \mathcal{MVN}(\mathbf{0}, cov(\mathbf{Z}^{Final})),$$

where

$$\begin{aligned} cov(\mathbf{Z}^{Final}) &= cov(\sqrt{w}\mathbf{Z}^I + \sqrt{1-w}\mathbf{Z}^{II}) = w\, cov(\mathbf{Z}^I) + (1-w) cov(\mathbf{Z}^{II}) \\ &= w \begin{pmatrix} 1 & \frac{1}{2} & D^I & \frac{D^I}{2} \\ \frac{1}{2} & 1 & \frac{D^I}{2} & D^I \\ D^I & \frac{D^I}{2} & 1 & \frac{1}{2} \\ \frac{D^I}{2} & D^I & \frac{1}{2} & 1 \end{pmatrix} + (1-w) \begin{pmatrix} 1 & \frac{1}{2} & D^{II} & \frac{D^{II}}{2} \\ \frac{1}{2} & 1 & \frac{D^{II}}{2} & D^{II} \\ D^{II} & \frac{D^{II}}{2} & 1 & \frac{1}{2} \\ \frac{D^{II}}{2} & D^{II} & \frac{1}{2} & 1 \end{pmatrix}, \end{aligned}$$

where $D^I = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^I}{f^2(1-g^I)}}}$ and $D^{II} = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^{II}}{f^2(1-g^{II})}}}$, as defined in (4.3) and (4.6).

The remaining hypotheses in the testing scheme test if subsets of the four drug effects are all zero. Therefore, for the other hypothesis, the test statistics are a subset of $\mathbf{Z}^{Final}$, and the covariance matrix among the test statistics are the corresponding subset of the covariance matrix $cov(\mathbf{Z}^{Final})$. We can similarly show that the final test statistics under any possible individual null hypothesis are multivariate normal.

Since the final statistics $\mathbf{Z}^{Final}$ follow a multivariate normal distribution under $H_0^{(4)}$, Follmann's procedure, which protects the local α level, can be performed on $\mathbf{Z}^{Final}$. Similarly, all the other three hypotheses in the testing scheme can be tested using Follmann's procedure at the α level. Therefore, using the closed testing schemes (see Table 3) proposed in Section 3.2, the strong control of Type I error rate is guaranteed for the adaptive procedures.

4.1.4 The Adaptation Rule

A variety of methods can be used to determine the second stage sample size. One method is based on the predictive power [59] [49]. Another method uses the conditional power (CP) [48]. Conditional power is a popular method that is widely used in adaptive designs. Wang et al. [68] proposed an adaptive population enrichment design, which built its adaptation rule based on the conditional power. Wang et al. [68] suggested computing the conditional power for each population/subpopulation at the end of stage I using the originally planned second stage sample size. If the conditional power of concluding the overall population is small, then the subpopulations should be enriched. Wang et al. [68] also suggested increasing the second stage sample size to a maximum value that the authors had pre-specified.

We can use various methods to develop adaptation rules, and we use the conditional power in this dissertation for illustrative purposes. Here we give an example of adaptation rule using the conditional power. After stage I, we conduct simulation studies and choose the second stage sample size $N_{Total}^{II}(\mathbf{Z}^I)$ so that the conditional power is big (for example, greater than 0.8) .

As discussed in Section 2.1, the design setting can be applied to achieve various objectives.

Thus, power is a complex notion in these designs, differing among the study objectives. In Section 3.4, we described power as the probability of concluding the population and/or the dose that are desired. The reasonable notions of power corresponding to various trial objectives are:

1. To find any population where there is at least one dose that is effective,

$$Power = Prob(Conclude\ \{G_A, L\}\ or\ \{G_A, H\}\ or\ \{G_S, L\}\ or\ \{G_S, H\}).$$

2. To find the largest population where at least one dose is effective,

$$Power = Prob(Conclude\ \{G_A, L\}\ or\ \{G_A, H\}).$$

3. To find the lowest dose where the dose is effective for at least one population,

$$Power = Prob(Conclude\ \{G_A, L\}\ or\ \{G_S, L\}).$$

The conditional power after Stage I is the probability of concluding the desired population and dose combination/combinations given the first stage data, the postulated effect sizes, which are denoted as $\mathbf{\Delta}_0$, where $\mathbf{\Delta}_0 = [\Delta_{G_A,L,0}, \Delta_{G_A,H,0}, \Delta_{G_S,L,0}, \Delta_{G_S,H,0}]'$, and the second stage sample size. We can also use the estimated effect sizes from the first stage data to obtain the conditional power using the estimated effect sizes instead of the postulated effect sizes, $\hat{\mathbf{\Delta}}^I = [\hat{\Delta}^I_{G_A,L}, \hat{\Delta}^I_{G_A,H}, \hat{\Delta}^I_{G_S,L}, \hat{\Delta}^I_{G_S,H}]'$, from the first stage. However, Bauer and Koenig [5] pointed out the instability of conditional power when the effect sizes were evaluated based on the interim observed data. Thus, in our adaptive rule, the conditional power is calculated conditional on the postulated effect sizes to avoid this instability, i.e.,

$$CP = \text{Probability(concluding the desired population and dose}|\mathbf{\Delta}_0, N^{II}_{Total}, \mathbf{Z}^I). \qquad (4.10)$$

It is very complex to compute the conditional power because the distribution of the final test statistics is a mixture of normal distributions under the alternative hypothesis. We use simulation to obtain the conditional power based on the postulated drug effect sizes, $\mathbf{\Delta}_0$, varying second stage sample size, and the first stage data, $\mathbf{Z}^I$. Then we pick the second stage sample size in order to obtain good conditional power, i.e., $N_{Total}^{II}(\mathbf{Z}^I) = min(N_{Total}^{II} : CP > 0.8)$. The adaptation rule, especially the simulation steps and rules, are pre-specified. Simulation studies on how to obtain conditional power and how to obtain $N_{Total}^{II}(\mathbf{Z}^I)$ are shown in Section 4.3.1.

4.2 THE ADVANCED ADAPTIVE DESIGN

In the second adaptive design, we allow further enriching the subpopulation, as well as increasing the sample size for Stage II. In comparison to the first adaptive design, this design is more flexible in allowing partial enrichment of the subpopulation. This leads us to call it the "advanced adaptive design".

4.2.1 The Design and the Final Statistics

In Section 4.1, we introduced the adaptive design where the second stage total sample size, $N_{Total}^{II}(\mathbf{Z}^I)$, depends on the first stage data. However, the sampling proportion of the subpopulation in the second stage, g^{II}, is pre-fixed. In the advanced adaptive design, both N_{Total}^{II} and g^{II} are determined based on the first stage data, i.e., $N_{Total}^{II}(\mathbf{Z}^I)$ and $g^{II}(\mathbf{Z}^I)$.

In the advanced adaptive design, the Stage I study design is the same as the Stage I study design in our first adaptive design, summarized in Table 12. At the end of Stage I, the total sample size and the sampling proportion of the subpopulation for the second stage are determined based on the first stage data through $\mathbf{Z}^I$ using a pre-specified adaptation rule. Based on $N_{Total}^{II}(\mathbf{Z}^I)$ and $g^{II}(\mathbf{Z}^I)$, stratified sampling is used, and our second stage study design is summarized in Table 14.

Again, there are various adaptation rules that can be used. Here we introduce a simple adaptation rule as an example, where the second stage total sample size N_{Total}^{II} and the second stage sampling proportion of the subpopulation g^{II} are originally planned to be $N_{Total,0}^{II}$ and g_0^{II}, respectively. However, the second stage total sample size and second stage sampling proportion may be increased to $N_{Total,max}^{II}$ and g_{max}^{II}. The adaptation rule says that after examining the first stage data, we keep using $N^{II}_{Total,0}$ and g_0^{II} if the conditional power is large (for example, greater than 0.8); otherwise, $N^{II}_{Total,max}$ and g^{II}_{max} are used as the second stage total sample size and sampling proportion of the subpopulation, respectively. All $N^{II}_{Total,0}$, g_0^{II}, $N^{II}_{Total,max}$, and g^{II}_{max} are pre-specified, so that their values do not depend on the observed value of $\mathbf{Z}^I$. Due to cost and safety reasons, $N^{II}_{Total,max}$ and g^{II}_{max} are the allowed maximum values of sample size and sampling proportion for stage II. How to obtain $N^{II}_{Total,0}$ and g_0^{II} will be shown in Section 4.3.2.

Table 14: Second Stage Advanced Adaptive Design Sample Sizes: 2 by 2 Case

	Doses			
	Low	High	Control	Total
G_S	$g^{II}(\mathbf{Z}^I)N^{II}(\mathbf{Z}^I)$	$g^{II}(\mathbf{Z}^I)N^{II}(\mathbf{Z}^I)$	$g^{II}(\mathbf{Z}^I)N^{II}(\mathbf{Z}^I)$	$g^{II}(\mathbf{Z}^I)N^{II}_{Total}(\mathbf{Z}^I)$
G_{S^-}	$[1-g^{II}(\mathbf{Z}^I)]N^{II}(\mathbf{Z}^I)$	$[1-g^{II}(\mathbf{Z}^I)]N^{II}(\mathbf{Z}^I)$	$[1-g^{II}(\mathbf{Z}^I)]N^{II}(\mathbf{Z}^I)$	$[1-g^{II}(\mathbf{Z}^I)]N^{II}_{Total}(\mathbf{Z}^I)$
G_A	$N^{II}(\mathbf{Z}^I)$	$N^{II}(\mathbf{Z}^I)$	$N^{II}(\mathbf{Z}^I)$	$N^{II}_{Total}(\mathbf{Z}^I)$

Note: $g^{II}(\mathbf{Z}^I)$ is the second stage sampling proportion of the subpopulation, $N^{II}_{Total}(\mathbf{Z}^I)$ is the total sample size in Stage II, and $N^{II}(\mathbf{Z}^I)$ is the sample size for each dose in Stage II. G_S is the subpopulation, G_{S^-} is the complimentary of subpopulation, and G_A is the overall population.

At the end of the study, we compute $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$ for stage I and stage II, using equations (3.5) and (3.6), and the final statistics $\mathbf{Z}^{Final}$ are computed by equation (4.7), where w is pre-specified.

Next, we investigate the distributions of these test statistics. Again, the first stage test statistics, $\mathbf{Z}^I$, has a multivariate normal distribution, as described in equations (4.1), (4.2) and (4.3). The second stage test statistics given the first stage data, $\mathbf{Z}^{II}|\mathbf{Z}^I$, also has a normal distribution,

$$\mathbf{Z}^{II}|\mathbf{Z}^I \sim \mathcal{N}(\mu_{\mathbf{Z}^{II}|\mathbf{Z}^I}, \begin{pmatrix} 1 & \frac{1}{2} & D_*^{II}(\mathbf{Z}^I) & \frac{D_*^{II}(\mathbf{Z}^I)}{2} \\ \frac{1}{2} & 1 & \frac{D_*^{II}(\mathbf{Z}^I)}{2} & D_*^{II}(\mathbf{Z}^I) \\ D_*^{II}(\mathbf{Z}^I) & \frac{D_*^{II}(\mathbf{Z}^I)}{2} & 1 & \frac{1}{2} \\ \frac{D_*^{II}(\mathbf{Z}^I)}{2} & D_*^{II}(\mathbf{Z}^I) & \frac{1}{2} & 1 \end{pmatrix}), \qquad (4.11)$$

where

$$\mu_{\mathbf{Z}^{II}|\mathbf{Z}^{I}} = \begin{pmatrix} \frac{\Delta_{G_A,L}}{\sqrt{\frac{2\sigma^2}{N^{II}(\mathbf{Z}^I)}(\frac{f^2}{g^{II}(\mathbf{Z}^I)}+\frac{(1-f)^2}{1-g^{II}(\mathbf{Z}^I)})}} \\ \frac{\Delta_{G_A,H}}{\sqrt{\frac{2\sigma^2}{N^{II}(\mathbf{Z}^I)}(\frac{f^2}{g^{II}(\mathbf{Z}^I)}+\frac{(1-f)^2}{1-g^{II}(\mathbf{Z}^I)})}} \\ \frac{\Delta_{G_S,L}}{\sqrt{\frac{2\sigma^2}{g^{II}(\mathbf{Z}^I)N^{II}(\mathbf{Z}^I)}}} \\ \frac{\Delta_{G_S,H}}{\sqrt{\frac{2\sigma^2}{g^{II}(\mathbf{Z}^I)N^{II}(\mathbf{Z}^I)}}} \end{pmatrix}, \tag{4.12}$$

$$\text{and } D_*^{II}(\mathbf{Z}^I) = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^{II}(\mathbf{Z}^I)}{f^2[1-g^{II}(\mathbf{Z}^I)]}}}. \tag{4.13}$$

Note that in comparison to (4.5), both $\mu_{\mathbf{Z}^{II}|\mathbf{Z}^{I}}$ and $cov(\mathbf{Z}^{II}|\mathbf{Z}^{I})$ in (4.12), (4.13) depend on $\mathbf{Z}^I$. Since $\mathbf{Z}^{II}$ are not independent of $\mathbf{Z}^I$, the distribution of $\mathbf{Z}^{Final}$ is a mixture of multivariate normals. We will show in next subsection that even under the null hypotheses, the distribution of $\mathbf{Z}^{Final}$ is not multivariate normal, which makes the critical values of Follmann's procedure difficult to compute.

4.2.2 Final Analysis

Again, the testing scheme and the test ordering are determined before the study based on the study goal and the prior beliefs about the drug effects. Because the testing scheme is closed under intersection, we require that each of the individual hypothesis in the testing scheme be tested at α level in order to control the type I error in the strong sense.

To illustrate an approach to do individual level-α testing, we consider how to test the null hypothesis $H_0^{(4)}$, i.e., $\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$, which is included in every testing scheme, by investigating the distribution of $\mathbf{Z}^{Final}$ under $H_0^{(4)}$. Under $H_0^{(4)}$, it can be seen from (4.2) and (4.12), $\mathbf{Z}^I$ and $\mathbf{Z}^{II}|\mathbf{Z}^I$ both have means $\mathbf{0}$, that is,

$$E(\mathbf{Z}^I) = \mathbf{0}, \tag{4.14}$$

$$E(\mathbf{Z}^{II}|\mathbf{Z}^I) = \mathbf{0}. \tag{4.15}$$

The distributions of $\mathbf{Z}^I$ and the conditional distribution of $\mathbf{Z}^{II}|\mathbf{Z}^I$ are multivariate normal,

$$\mathbf{Z}^I \sim \mathcal{N}(\mathbf{0}, \begin{pmatrix} 1 & \frac{1}{2} & D^I & \frac{D^I}{2} \\ \frac{1}{2} & 1 & \frac{D^I}{2} & D^I \\ D^I & \frac{D^I}{2} & 1 & \frac{1}{2} \\ \frac{D^I}{2} & D^I & \frac{1}{2} & 1 \end{pmatrix}), \tag{4.16}$$

$$\mathbf{Z}^{II}|\mathbf{Z}^I \sim \mathcal{N}(\mathbf{0}, \begin{pmatrix} 1 & \frac{1}{2} & D_*^{II}(\mathbf{Z}^I) & \frac{D_*^{II}(\mathbf{Z}^I)}{2} \\ \frac{1}{2} & 1 & \frac{D_*^{II}(\mathbf{Z}^I)}{2} & D_*^{II}(\mathbf{Z}^I) \\ D_*^{II}(\mathbf{Z}^I) & \frac{D_*^{II}(\mathbf{Z}^I)}{2} & 1 & \frac{1}{2} \\ \frac{D_*^{II}(\mathbf{Z}^I)}{2} & D_*^{II}(\mathbf{Z}^I) & \frac{1}{2} & 1 \end{pmatrix}), \tag{4.17}$$

where $D^I = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^I}{f^2(1-g^I)}}}$ and $D_*^{II}(\mathbf{Z}^I) = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^{II}(\mathbf{Z}^I)}{f^2[1-g^{II}(\mathbf{Z}^I)]}}}$ are defined in (4.3) and (4.13).

As we indicated, it is obvious that $\mathbf{Z}^{II}$ is not independent of $\mathbf{Z}^I$. Thus, $\mathbf{Z}^{Final}$ as computed by equation (4.7) doesn't have a multivariate normal distribution. Therefore, we cannot directly use Follmann's procedure for each individual hypothesis, since application of Follmann's procedure requires that $\mathbf{Z}^{Final}$ have a multivariate normal distribution. In Sections 4.2.2.1 - 4.2.2.3, we propose three methods that may be useful for each individual test.

Note that we will only demonstrate how to test the null hypothesis $H_0^{(4)}$. There are three more null hypotheses to be tested in each of our proposed testing schemes. Since the other three hypotheses are testing whether a subset of the drug effect tested in $H_0^{(4)}$ are zero, we can use the same methods to test $H_0^{(3)}$, $H_0^{(2)}$, and $H_0^{(1)}$ as we propose for testing $H_0^{(4)}$.

4.2.2.1 Numerical Integration

Due to the simple adaptation rule, in which g^{II} is chosen from two values, g_0^{II} and g_{max}^{II}, it is possible to identify the distribution of $\mathbf{Z}^{Final}$. The cumulative distribution function of

$\mathbf{Z}^{Final}$ is,

$$\begin{aligned}
&P(\mathbf{Z}^{Final} \leq \mathbf{c}) \\
=&\int\int\int\int P(\mathbf{Z}^{Final} \leq \mathbf{c}|\mathbf{Z}^I)\phi(\mathbf{Z}^I) d_{Z^I_{G_A,L}} d_{Z^I_{G_A,H}} d_{Z^I_{G_S,L}} d_{Z^I_{G_S,H}} \\
=&\int\int\int\int P(w\mathbf{Z}^I + (1-w)\mathbf{Z}^{II} \leq \mathbf{c}|\mathbf{Z}^I)\phi(\mathbf{Z}^I) d_{Z^I_{G_A,L}} d_{Z^I_{G_A,H}} d_{Z^I_{G_S,L}} d_{Z^I_{G_S,H}} \\
=&\int\int\int\int P(\mathbf{Z}^{II} \leq \frac{\mathbf{c} - w\mathbf{Z}^I}{1-w}|\mathbf{Z}^I)\phi(\mathbf{Z}^I) d_{Z^I_{G_A,L}} d_{Z^I_{G_A,H}} d_{Z^I_{G_S,L}} d_{Z^I_{G_S,H}} \\
=&\int\int\int\int \Phi(\frac{\mathbf{c} - w\mathbf{Z}^I}{1-w})\phi(\mathbf{Z}^I) d_{Z^I_{G_A,L}} d_{Z^I_{G_A,H}} d_{Z^I_{G_S,L}} d_{Z^I_{G_S,H}} \\
=&\int\int\int\int [I_{\{\mathbf{Z}^I : g^{II}(\mathbf{Z}^I) = g^{II}_0\}} + I_{\{\mathbf{Z}^I : g^{II}(\mathbf{Z}^I) = g^{II}_{max}\}}]\Phi(\frac{\mathbf{c} - w\mathbf{Z}^I}{1-w})\phi(\mathbf{Z}^I) d_{Z^I_{G_A,L}} d_{Z^I_{G_A,H}} d_{Z^I_{G_S,L}} d_{Z^I_{G_S,H}} \\
=&\int\int\int\int_{\{\mathbf{Z}^I : g^{II}(\mathbf{Z}^I) = g^{II}_0\}} \Phi(\frac{\mathbf{c} - w\mathbf{Z}^I}{1-w})\phi(\mathbf{Z}^I) d_{Z^I_{A,L}} d_{Z^I_{A,H}} d_{Z^I_{S,L}} d_{Z^I_{S,H}} \\
&+\int\int\int\int_{\{\mathbf{Z}^I : g^{II}(\mathbf{Z}^I) = g^{II}_{max}\}} \Phi(\frac{\mathbf{c} - w\mathbf{Z}^I}{1-w})\phi(\mathbf{Z}^I) d_{Z^I_{A,L}} d_{Z^I_{A,H}} d_{Z^I_{S,L}} d_{Z^I_{S,H}}.
\end{aligned}$$

To accurately evaluate this cumulative density function (cdf) of $\mathbf{Z}^{Final}$, one needs to obtain a suitable expression for the multivariate normal cdf and then use four-fold numerical integration. There is no closed form expression for the actual multivariate normal cdf. However, the multivariate normal cdf values may be accurately approximated by a variety of methods, such as Taylor series, asymptotic series and continued fractions [40] [9] [70]. Moreover, numerical integration over four folds can be complex. We are not going to provide further details of the integration in this dissertation. We only point out that numerical integration may be a possible solution to this problem with small number of doses and populations.

4.2.2.2 Approximation to Follmann's Test

Although the exact distribution of $\mathbf{Z}^{Final}$ is hard to compute, we can obtain its mean and covariance matrix under the null.

Under the null hypotheses, using (4.7) and (4.15), we easily obtain the conditional mean of $\mathbf{Z}^{Final}$ given $\mathbf{Z}^I$,

$$E(\mathbf{Z}^{Final}|\mathbf{Z}^I) = E[(\sqrt{w}\mathbf{Z}^I + \sqrt{1-w}\mathbf{Z}^{II})|\mathbf{Z}^I] = \mathbf{0}. \tag{4.19}$$

Thus, the unconditional mean of $\mathbf{Z}^{Final}$ is:

$$E(\mathbf{Z}^{Final}) = E[E(\mathbf{Z}^{Final}|\mathbf{Z}^I)] = \mathbf{0}. \tag{4.20}$$

Under the null hypothesis, using (4.15), we can obtain the covariance of $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$ as

$$\begin{aligned}
&Cov(\mathbf{Z}^I, \mathbf{Z}^{II}) \\
&= E(\mathbf{Z}^I\mathbf{Z}^{II}) - E(\mathbf{Z}^I)E(\mathbf{Z}^{II}) \\
&= E(\mathbf{Z}^I\mathbf{Z}^{II}) - E(\mathbf{Z^I})E_{\mathbf{Z^I}}[E(\mathbf{Z}^{II}|\mathbf{Z}^I)] \\
&= E_{\mathbf{Z}^I}[E(\mathbf{Z}^I\mathbf{Z}^{II}|\mathbf{Z}^I)] \\
&= E_{\mathbf{Z}^I}[\mathbf{Z}^I E(\mathbf{Z}^{II}|\mathbf{Z}^I)] \\
&= \mathbf{0}.
\end{aligned} \tag{4.21}$$

Using (4.1), (4.7), (4.11), (4.15), (4.19) and (4.21), we obtain the unconditional covariance of $\mathbf{Z}^{Final}$ under $H_0^{(4)}$ as:

$$\begin{aligned}
&Cov(\mathbf{Z}^{Final}) \\
&= E[Cov(\mathbf{Z}^{Final}|\mathbf{Z}^I)] + Cov[E(\mathbf{Z}^{Final}|\mathbf{Z}^I)] \\
&= E[Cov(\mathbf{Z}^{Final}|\mathbf{Z}^I)] \\
&= E[wCov(\mathbf{Z}^I) + (1-w)Cov(\mathbf{Z}^{II}|\mathbf{Z}^I)] \\
&= wCov(\mathbf{Z}^I) + (1-w)E[Cov(\mathbf{Z}^{II}|\mathbf{Z}^I)] \\
&= w(\begin{pmatrix} 1 & \frac{1}{2} & D^I & \frac{D^I}{2} \\ \frac{1}{2} & 1 & \frac{D^I}{2} & D^I \\ D^I & \frac{D^I}{2} & 1 & \frac{1}{2} \\ \frac{D^I}{2} & D^I & \frac{1}{2} & 1 \end{pmatrix}) + (1-w)E(\begin{pmatrix} 1 & \frac{1}{2} & D_*^{II}(\mathbf{Z}^I) & \frac{D_*^{II}(\mathbf{Z}^I)}{2} \\ \frac{1}{2} & 1 & \frac{D_*^{II}(\mathbf{Z}^I)}{2} & D_*^{II}(\mathbf{Z}^I) \\ D_*^{II}(\mathbf{Z}^I) & \frac{D_*^{II}(\mathbf{Z}^I)}{2} & 1 & \frac{1}{2} \\ \frac{D_*^{II}(\mathbf{Z}^I)}{2} & D_*^{II}(\mathbf{Z}^I) & \frac{1}{2} & 1 \end{pmatrix}) \\
&= \begin{pmatrix} 1 & \frac{1}{2} & D_* & \frac{D_*}{2} \\ \frac{1}{2} & 1 & \frac{D_*}{2} & D_* \\ D_* & \frac{D_*}{2} & 1 & \frac{1}{2} \\ \frac{D_*}{2} & D_* & \frac{1}{2} & 1 \end{pmatrix},
\end{aligned} \tag{4.22}$$

where,

$$\begin{aligned}
&D^I = \sqrt{\frac{1}{1+\frac{(1-f)^2 g^I}{f^2(1-g^I)}}} \text{as defined in (4.3)},\\
&D_* = wD^I + (1-w)E(D_*^{II}(\mathbf{Z}^I)),\\
&E(D_*^{II}(\mathbf{Z}^I)) = \{P[g^{II}(\mathbf{Z}^I) = g_0^{II}]\}D_*^I + \{P[g^{II}(\mathbf{Z}^I) = g_{max}^{II}]\}D_*^{II},\\
&D_*^I = \sqrt{\frac{1}{1+\frac{(1-f)^2 g_0^{II}}{f^2(1-g_0^{II})}}}, \text{ and } D_*^{II} = \sqrt{\frac{1}{1+\frac{(1-f)^2 g_{max}^{II}}{f^2(1-g_{max}^{II})}}}.
\end{aligned}$$

Since the adaptation rule is pre-specified, and the distribution of $\mathbf{Z}^I$ is known, the probabilities $P[g^{II}(\mathbf{Z}^I) = g_0^{II}]$ and $P[g^{II}(\mathbf{Z}^I) = g_{max}^{II}]$ can be obtained either by mathematical computation or simulation studies.

Recall that Follmann's procedure requires data to be from a multivariate normal distribution, and it rejects $H_0^{(4)}$ if $\{\mathbf{Z}^T\mathbf{\Sigma}^{-1}\mathbf{Z} > \chi^2_{p,2\alpha}\ and\ \Sigma Z_i^{>} 0\}$, where the critical value $\chi^2_{p,2\alpha}$ is the 95^{th} percentile of the χ^2 distribution with 4 degrees of freedom.

As we know the mean of $\mathbf{Z}^{Final}$ in (4.20) and the covariance of $\mathbf{Z}^{Final}$ in (4.22), with sufficiently large sample size, we are willing to assume that $\mathbf{Z}^{Final}$ is approximately multivariate normal and use Follmann's procedure for an approximation. In the following subsection, we propose a method to improve upon this approximation.

4.2.2.3 Obtaining Critical Values by Simulation in Follmann's Procedure

We want to propose an approach that is similar to Follmann's procedure, where we reject the null hypothesis if $\{\mathbf{Z}^{Final^T}\Sigma^{Final^{-1}}\mathbf{Z}^{Final} > c^*$ and $\Sigma Z_i^{Final} > 0\}$. However, instead of using critical values from χ^2 distribution as in Follmann's procedure, we propose to obtain the critical values c^*'s from simulation.

Follmann proved in his paper that if the first part $\mathbf{Z}^{Final^T}\Sigma^{Final^{-1}}\mathbf{Z}^{Final} > c^*$ is a 2α test, the overall procedure is an α level procedure with the addition of the constraint $\Sigma Z_i^{Final} > 0$ (Theorem 1 [16]). Therefore, simulation studies can be conducted to obtain c^*. Simulation studies are demonstrated in Section 4.3.2.

4.3 SIMULATION STUDIES

Monte Carlo simulation studies are conducted in order to examine the operating characteristics of our proposed adaptive designs. Due to the complexity of the proposed trial designs and the data analysis methods, the simulation studies in this section are limited and are included for illustrative purposes.

4.3.1 Simulation Studies for the Adaptive Design with Sample Size Adjustment

In Section 4.1, we proposed an adaptive design where the second stage sample size can be determined based on the first stage data. In this section, we show one feasible method of determining the second stage sample size using simulation studies.

Our goal is to find a second stage sample, $N^{II}_{Total}(\mathbf{Z}^I)$, so that the conditional power is greater than 0.8. We conduct simulation studies to obtain conditional powers for varying possible second stage sample sizes, N^{II}_{Total}'s, based on $\mathbf{\Delta}_0$ and $\mathbf{Z}^I$. Then we pick the smallest sample size such that the conditional power is greater than 0.8, i.e., $N^{II}_{Total}(\mathbf{Z}^I) = min(N^{II}_{Total} : CP_{\mathbf{\Delta}_0, \mathbf{Z}^I, N^{II}_{Total}} > 0.8)$.

Due to safety and cost reasons in clinical trials, there may exist limits for the enrolled number of subjects for the second stage. Denote the minimum second stage sample size by $N^{II}_{Total,min}$. Denote the maximum second stage sample size by $N^{II}_{Total,max}$.

For simplicity, we assume that subjects within a population are randomly assigned to treatment arms with equal numbers in each, i.e., $N^I = N^I_{Total}/3$, $N^{II}(\mathbf{Z}^I) = N^{II}_{Total}(\mathbf{Z}^I)/3$, $N^{II} = N^{II}_{Total}/3$, $N^{II}_{min} = N^{II}_{Total,min}/3$, and $N^{II}_{max} = N^{II}_{Total,max}/3$. Our second stage sample size becomes :

$$N^{II}(\mathbf{Z}^I) = \begin{cases} N^{II}_{min}, & \text{if } min(N^{II} : CP_{\mathbf{\Delta}_0, \mathbf{Z}^I, N^{II}} > 0.8) < N^{II}_{min}; \\ min(N^{II} : CP_{\mathbf{\Delta}_0, \mathbf{Z}^I, N^{II}} > 0.8), & \text{if } N^{II}_{min} \leq min(N^{II} : CP_{\mathbf{\Delta}_0, \mathbf{Z}^I, N^{II}} > 0.8) \leq N^{II}_{max}; \\ N^{II}_{max}, & \text{if } min(N^{II} : CP_{\mathbf{\Delta}_0, \mathbf{Z}^I, N^{II}} > 0.8) > N^{II}_{max}. \end{cases}$$

Simulation studies are conducted and powers are obtained for using N^{II} varying from N^{II}_{min} to N^{II}_{max}. The following are the simulation steps for obtaining power for a specific

N^{II}, for example $N^{II} = N^{II}_{*}$. Obtaining powers for other values of N^{II} follows the same simulation steps.

1. Let $N^{II}(\mathbf{Z}^{I})$=N^{II}_{*}.
2. Randomly generate $\mathbf{Z}^{II}$ from the distribution in equation (4.4), (4.5) and (4.6) using $\mathbf{\Delta}_0$ and $N^{II}(\mathbf{Z}^{I})$ as defined in step 1.
3. Compute $\mathbf{Z}^{Final}$ from $\mathbf{Z}^{I}$, obtained from the stage I data, and $\mathbf{Z}^{II}$, generated in step 2, using equation (4.7).
4. Use Follmann's procedure to test each individual hypothesis following the testing scheme, and make conclusion for the population and dose.
5. Repeat Step 1-Step 4 for N times. Compute the conditional power.

Simulation studies are conducted for all possible N^{II}'s. From the results, we can decide $N^{II}(\mathbf{Z}^{I})$.

Let's look at an example. Consider a clinical trial where the drug effects (low dose and high dose) are evaluated on both the overall population, G_A, and the subpopulation, G_S. Suppose that the primary goal of the trial sponsor is to find the largest population where there is at least one effective dose; and the secondary goal is for this largest population find the minimum effective dose. Suppose that the prior beliefs concerning the drug effect sizes are listed as in Table 15. Suppose $f = g = 0.4$, then $\mathbf{\Delta}_0 = [\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H}]' = [0.18, 0.27, 0.3, 0.45]'$.

Table 15: Drug Effects for Simulation Studies for the Adaptive Designs

	Low	High
G_A	0.18	0.27
G_S	0.3	0.45
G_{S^-}	0.1	0.15

The first stage sample size for each treatment arm is planned to be 150, i.e., $N^I = 150$. The minimum second stage sample size per each treatment is 50 and the maximum second stage sample size per each treatment is 150, i.e., $N^{II}_{min} = 50$ and $N^{II}_{max} = 150$.

Before the trial begins, we need to find an appropriate ordering of the testing scheme. Simulation studies are conducted as for the non-adaptive designs (Section 3.4). Let the sample size for each treatment arm be $N^I + \frac{N_{min}^{II}+N_{max}^{II}}{2}$, which equals to 250 in our example. Each simulation study uses 10,000 iterations. Powers for all 24 test orderings are shown in Table 16. The corresponding test orderings are listed in Table 3. From the simulation results, we decide that ordering 1 is the most appropriate ordering and will be used in all subsequent simulation studies. To be specific, test ordering 1 tests

$$\begin{aligned}
&H_0^{(4)} : \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0,\\
&H_0^{(3)} : \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = 0,\\
&H_0^{(2)} : \Delta_{G_A,L} = \Delta_{G_A,H} = 0,\\
&H_0^{(1)} : \Delta_{G_A,L} = 0;
\end{aligned}$$

versus

$$\begin{aligned}
&H_a^{(4)} : \Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0 \quad or \quad \Delta_{G_S,L} > 0 \quad or \quad \Delta_{G_S,H} > 0,\\
&H_a^{(3)} : \Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0 \quad or \quad \Delta_{G_S,L} > 0,\\
&H_a^{(2)} : \Delta_{G_A,L} > 0 \quad or \quad \Delta_{G_A,H} > 0,\\
&H_a^{(1)} : \Delta_{G_A,L} > 0.
\end{aligned}$$

Suppose that after stage I, we have observed that $\mathbf{Z}^I = [0.62, 1.71, 1.59, 3.26]'$. We perform simulation studies to obtain conditional power based on $\mathbf{Z}^I$ for N^{II} ranging from 50 to 150 by 10, i.e., $N^{II} = 50, 60, 70, ..., 150$. Each simulation study uses 10,000 iterations, i.e., N=10,000. The simulation results are presented in Table 17. From the results, we can decide $N^{II}(\mathbf{Z}^I) = 100$ in order to achieve 80% power of the study.

Table 16: Finding the Best Ordering for $\mathbf{\Delta} = [0.18, 0.27, 0.3, 0.45]'$.

Ordering	G_A	G_S	L	H	G_A, L	G_A, H	G_S, L	G_S, H	None
1	0.73	0.09	0.52	0.30	0.48	0.25	0.04	0.05	0.18
2	0.73	0.09	0.05	0.77	0.01	0.72	0.04	0.05	0.18
3	0.68	0.14	0.51	0.31	0.42	0.26	0.08	0.05	0.18
4	0.31	0.51	0.51	0.31	0.05	0.26	0.45	0.05	0.18
5	0.74	0.08	0.03	0.78	0.01	0.73	0.02	0.05	0.18
6	0.24	0.58	0.53	0.28	0.01	0.23	0.52	0.05	0.18
7	0.74	0.08	0.50	0.32	0.48	0.26	0.01	0.06	0.18
8	0.74	0.08	0.03	0.79	0.01	0.73	0.01	0.06	0.18
9	0.51	0.31	0.50	0.31	0.49	0.02	0.01	0.30	0.18
10	0.03	0.78	0.03	0.79	0.01	0.02	0.01	0.77	0.18
11	0.76	0.05	0.02	0.79	0.01	0.75	0.01	0.04	0.18
12	0.03	0.79	0.02	0.79	0.01	0.02	0.01	0.78	0.18
13	0.46	0.36	0.51	0.31	0.42	0.04	0.08	0.27	0.18
14	0.09	0.73	0.51	0.31	0.05	0.04	0.46	0.27	0.18
15	0.52	0.29	0.49	0.32	0.49	0.04	0.01	0.29	0.18
16	0.05	0.77	0.02	0.80	0.01	0.04	0.01	0.76	0.18
17	0.06	0.76	0.54	0.27	0.02	0.04	0.53	0.24	0.18
18	0.06	0.76	0.03	0.79	0.02	0.04	0.01	0.75	0.18
19	0.76	0.06	0.04	0.78	0.01	0.74	0.03	0.03	0.18
20	0.25	0.56	0.54	0.28	0.01	0.24	0.53	0.03	0.18
21	0.76	0.05	0.02	0.79	0.01	0.75	0.01	0.04	0.18
22	0.03	0.79	0.02	0.79	0.01	0.02	0.01	0.78	0.18
23	0.04	0.77	0.54	0.28	0.01	0.03	0.53	0.24	0.18
24	0.04	0.77	0.02	0.80	0.01	0.03	0.01	0.76	0.18

Table 17: Determining the Second Stage Sample Size

N^{II}	$CP_{\boldsymbol{\Delta}_0,\mathbf{Z}^I,N^{II}}$
60	0.72
70	0.75
80	0.77
90	0.78
100	0.80
110	0.81
120	0.83
130	0.84
140	0.85
150	0.86

4.3.2 Simulation Studies for the Advanced Adaptive Design

In Section 4.2, we proposed a more advanced adaptive design where both the second stage sample size and sampling proportion of the subpopulation can be determined based on the first stage data. One possible method to find $N^{II}_{Total}(\mathbf{Z}^I)$ and $g^{II}(\mathbf{Z}^I)$ is to again use the idea introduced in Section 4.3.1 in order to achieve 0.8 conditional power. However, we suggest another feasible method in this section to show an alternate way of determining the second stage sample size and sampling proportion of the subpopulation.

We consider a simple adaptation rule: if the conditional power after stage I is large, $N^{II}_{Total,0}$ and g^{II}_0 are used as the second stage sample size and the second stage sampling proportion of the subpopulation; if the conditional power is small, $N^{II}_{Total,max}$ and g^{II}_{max} are used as the second stage sample size and the second stage sampling proportion of the subpopulation, where $N^{II}_{Total,0}$, $N^{II}_{Total,max}$, g^{II}_0 and g^{II}_{max} are pre-specified. Due to cost and safety reasons, $N^{II}_{Total,max}$ and g^{II}_{max} are the maximum values allowed by the clinical trials. In this section, we introduce how to pre-specify $N^{II}_{Total,0}$ and g^{II}_0 using simulation studies.

To determine the values for g_0^{II} and N_0^{II}, we consider a non-adaptive design with two stages to approximate the adaptive design. The first stage sample size and sampling proportion are N^I and g^I, respectively, as in the adaptive design. The second stage sample size is N^{II} and sampling proportion is g^{II}. Note that N^{II} and g^{II} do not depend on the first stage data. We perform simulation studies using varying values of N^{II} and g^{II} to obtain powers of the two-stage non-adaptive design. The final statistics and their distribution under the null hypothesis are computed from (4.7), (4.20) and (4.22). Then N_0^{II} and g_0^{II} will be assigned by the values of N^{II} and g^{II} that provide appropriate power of the trial, i.e., the probability of concluding any effective dose on any population is greater than 0.8.

However, the final test statistics do not follow multivariate normal distributions, as indicated in Section 4.2.2. Therefore, before finding N_0^{II} and g_0^{II}, simulation studies need to be first performed to find the critical values, c^*'s, for Follmann's procedure as discussed in Section 4.2.2.3.

Let N^{II} and g^{II} vary in the range allowed by the trial. We will find a set of critical values $\{c_4^*, c_3^*, c_2^*, c_1^*\}$ for every possible combination of N^{II} and g^{II}. In this dissertation, we only focus on describing how to obtain the critical value c_4^* for $H_0^{(4)}$, i.e., $\Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_S,L} = \Delta_{G_S,H} = 0$, for one possible set of values of N^{II} (for example, N_{**}^{II}) and g^{II} (for example, g_{**}^{II}). Methods of obtaining the critical values for $H_0^{(3)}$, $H_0^{(2)}$ and $H_0^{(1)}$ and for other combinations of N^{II} and g^{II} are similar. In the simulation study, we randomly generate N values (for example, $N = 10,000$) of $\mathbf{Z}^{Final^T} cov(\mathbf{Z}^{Final})^{-1}\mathbf{Z}^{Final}$, and then find the $100(1-2\alpha)^{th}$ percentile of all these values as the critical value c_4^*.

The simulation steps to obtain c_4^* for $H_0^{(4)}$ for N_{**}^{II} and g_{**}^{II} are described as follows:

1. Randomly generate $\mathbf{Z}^I$ from the distribution in equation (4.16).
2. Randomly generate $\mathbf{Z}^{II}$ from the distribution in equation (4.17) using $g^{II}(\mathbf{Z}^I) = g_{**}^{II}$.
3. Compute $\mathbf{Z}^{Final}$ from $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$ using equation (4.7).
4. Repeat Step 1 - Step 3 for N times. Let c^* be the $100(1-2\alpha)^{th}$ largest number of all the $\mathbf{Z}^{Final^T} cov(\mathbf{Z}^{Final})^{-1}\mathbf{Z}^{Final}$ values.

Then simulation studies are conducted to determine N_0^{II} and g_0^{II}. The simulation steps for N_{**}^{II} and g_{**}^{II} are as the following. Obtaining powers for other combinations of N^{II} and

g^{II} follows the same simulation steps.

1. Randomly generate $\mathbf{Z}^I$ from the distribution in (4.1), (4.2) and (4.3).
2. Randomly generate $\mathbf{Z}^{II}$ from the distribution in (4.11) and (4.13) using $N^{II}(\mathbf{Z}^I) = N^{II}_{**}$ and $g^{II}(\mathbf{Z}^I) = g^{II}_{**}$.
3. Compute $\mathbf{Z}^{Final}$ from $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$ from (4.7).
4. Use Follmann's procedure, but use the critical values generated by the previous simulation, to test each individual hypothesis following the testing scheme, and make conclusion for the population and dose.
5. Repeat Step 1 - Step 4 for N times. Compute the power.

Based on the simulation results, we can decide N^{II}_0 and g^{II}_0 in order to reach 0.8 power of the two-stage non-adaptive design. Note that we could have multiple combinations of N^{II}_0 and g^{II}_0 values. Also note that we only use the non-adaptive design to approximate the adaptive design. Therefore, the unconditional power of the adaptive design is not guaranteed to reach 0.8.

Let look at an example. Still consider the previous clinical trial. The setting of the trial is exactly the same as described in Section 4.3.1. The primary goal of the trial sponsor is to find the largest population where there is at least one effective dose; and the secondary goal is for this largest population find the minimum effective dose. Suppose $f = g = 0.4$. The prior belief of the drug effect sizes are listed in Table 15, where $\boldsymbol{\Delta}_0 = [\Delta_{G_A,L}, \Delta_{G_A,H}, \Delta_{G_S,L}, \Delta_{G_S,H}]' = [0.18, 0.27, 0.3, 0.45]'$. As in Section 4.3.1, ordering 1 will again be used for all simulation studies in this section. Again, for simplicity, we assume that subjects within a population are randomly assigned to treatment arms with equal numbers in each, i.e., $N^I = N^I_{Total}/3$, $N^{II}(\mathbf{Z}^I) = N^{II}_{Total}(\mathbf{Z}^I)/3$, $N^{II}_{min} = N^{II}_{Total,min}/3$, and $N^{II}_{max} = N^{II}_{Total,max}/3$. The first stage sample size per each treatment is 150, i.e., $N^I = 150$.

Let N^{II} vary from 50 to 120 by 10 and let $g^{II} = 0.4, 0.6, 0.8$. Simulation studies are conducted for every combination of indicated values of N^{II} and g^{II}. Each simulation study uses 10,000 iterations, i.e., N=10,000. The simulation results of critical values are listed in Table 18.

Simulation studies are then conducted for all indicated sample sizes and sampling pro-

Table 18: Finding Critical Values: $c^{\star}$'s

	$g^{II} = 0.4$				$g^{II} = 0.6$				$g^{II} = 0.8$			
N^{II}	c_1^*	c_2^*	c_3^*	c_4^*	c_1^*	c_2^*	c_3^*	c_4^*	c_1^*	c_2^*	c_3^*	c_4^*
50	3.808461	5.936774	7.730863	9.425242	3.780432	5.883105	7.760212	9.440906	3.867756	5.950001	7.817996	9.557915
60	3.951102	6.042804	7.758814	9.537202	3.757424	5.925588	7.784478	9.428409	3.955842	6.018503	7.889129	9.468446
70	3.849488	6.106879	7.824094	9.586355	3.765772	5.947064	7.758804	9.373043	3.792991	5.910098	7.728014	9.358025
80	3.823605	5.837915	7.708902	9.385033	3.742856	5.900719	7.636444	9.410998	3.774041	5.942636	7.895747	9.483035
90	3.830666	6.036252	7.905278	9.440555	3.81436	6.055347	8.041597	9.579794	3.739802	5.983838	7.704245	9.511243
100	3.851146	6.098712	7.781723	9.536403	3.881481	6.018855	7.897952	9.53479	3.89356	6.0954	7.788135	9.498599
110	3.840209	6.034233	7.955338	9.573312	3.77169	5.945029	7.757748	9.356418	3.903191	5.975523	7.819939	9.485663
120	3.819175	5.987406	7.828519	9.417969	3.84695	5.994449	7.924097	9.552963	3.698731	5.983403	7.820055	9.49176

portions to find the powers. Each simulation study uses 10,000 iterations, i.e., N=10,000. The results are listed in Table 19. From the results, we can decide $N_0^{II} = 100, g_0^{II} = 0.4$, or $N_0^{II} = 70, g_0^{II} = 0.6$, or $N_0^{II} = 64, g_0^{II} = 0.8$ (result not shown in Table 19) in order to achieve 80% power of concluding any effective dose on any population.

Therefore, depending on how much the researchers want to enrich the population, we can decide the corresponding N_0^{II} accordingly. For example, if the sampling proportion of the subpopulation can be enriched to 0.6 in the second stage, our adaptation rule can be: if the conditional power of concluding any population and dose is greater than 0.8, use $N_0^{II} = 70, g_0^{II} = 0.4$; otherwise, use $N_{max}^{II}, g_{max}^{II}$.

In the clinical trial, after collecting the first stage data, we obtain the conditional power based on $\mathbf{Z}^I$, N_0^{II}, g_0^{II} and $\mathbf{\Delta}_0$ by simulation. The following are the simulation steps to obtain conditional power after stage I:

1. Randomly generate $\mathbf{Z}^{II}$ from the distribution in equations (4.11) and (4.13) using N_0^{II} and g_0^{II}.
2. Compute $\mathbf{Z}^{Final}$ from $\mathbf{Z}^I$ and $\mathbf{Z}^{II}$ using equation (4.7).
3. Use Follmann's procedure, but use the critical values in Table 18, to test each individual hypothesis following the testing scheme, and make conclusion for the population and dose.
4. Repeat Step 1 - Step 3 for N times. Compute the conditional power.

In this section, we have explicitly explained how to make the adaptation rule before the study using an example. For other scenarios, the simulation studies will be similar and will not be shown in this dissertation.

Table 19: Finding N_0^{II} and g_0^{II}

N^{II}	G_A	G_S	L	H	G_A, L	G_A, H	G_S, L	G_S, H	none	Power
	$g^{II} = 0.4$									
50	0.61	0.11	0.42	0.30	0.37	0.23	0.04	0.06	0.29	0.71
60	0.62	0.10	0.42	0.30	0.38	0.24	0.04	0.06	0.28	0.72
70	0.64	0.10	0.45	0.29	0.41	0.23	0.04	0.06	0.26	0.74
80	0.67	0.09	0.47	0.30	0.43	0.24	0.04	0.06	0.23	0.77
90	0.69	0.10	0.49	0.30	0.46	0.23	0.04	0.06	0.21	0.79
100	0.71	0.09	0.50	0.30	0.46	0.25	0.04	0.05	0.20	0.80
110	0.72	0.09	0.51	0.30	0.47	0.25	0.04	0.06	0.19	0.81
120	0.74	0.09	0.53	0.30	0.50	0.24	0.04	0.05	0.17	0.83
	$g^{II} = 0.6$									
50	0.61	0.14	0.43	0.32	0.37	0.24	0.06	0.08	0.26	0.74
60	0.63	0.14	0.46	0.31	0.40	0.23	0.06	0.08	0.23	0.77
70	0.66	0.14	0.47	0.33	0.41	0.25	0.06	0.08	0.20	0.80
80	0.67	0.14	0.49	0.32	0.43	0.25	0.06	0.07	0.19	0.81
90	0.67	0.14	0.50	0.32	0.44	0.24	0.06	0.08	0.18	0.82
100	0.71	0.14	0.51	0.33	0.45	0.26	0.06	0.07	0.16	0.84
110	0.73	0.09	0.53	0.29	0.49	0.24	0.03	0.05	0.18	0.82
120	0.74	0.09	0.53	0.30	0.49	0.25	0.03	0.06	0.17	0.83
	$g^{II} = 0.8$									
50	0.56	0.20	0.42	0.34	0.33	0.23	0.09	0.11	0.24	0.76
60	0.58	0.21	0.43	0.36	0.34	0.24	0.10	0.12	0.21	0.79
70	0.61	0.21	0.47	0.35	0.37	0.24	0.10	0.11	0.18	0.82
80	0.61	0.21	0.48	0.35	0.37	0.24	0.10	0.11	0.18	0.82
90	0.64	0.21	0.50	0.35	0.39	0.24	0.11	0.10	0.15	0.85
100	0.64	0.22	0.50	0.36	0.38	0.26	0.12	0.11	0.14	0.86
110	0.71	0.10	0.50	0.31	0.46	0.25	0.04	0.06	0.19	0.81
120	0.74	0.09	0.55	0.28	0.51	0.23	0.04	0.05	0.17	0.83

5.0 CLUSTER VALIDATION

5.1 INTRODUCTION

In this chapter, we present auxiliary material concerning studying the population heterogeneity among schizophrenia subjects based on post-mortem studies in schizophrenia. Our previous work identified a subtype of schizophrenia. Here we focus on externally validating this subtype finding in independent studies.

A review of the previous study is given in Section 5.1 and a description of the motivating data set is given in Section 5.2. Then two approaches are used to externally validating the previous LGM cluster finding. The first approach, as discussed in Section 5.3, extends Kapp and Tibshirani's [27] classification idea and modifies their approach for cluster validation to handle our situation. In doing so, we discuss why their ideas cannot be directly applied to our motivating data set, thereby suggesting a new method for cluster validation. The second approach again applies the clustering analysis used for the defining data set on the validating data set and the combination of the defining and validating data sets. The methods and results are shown in Section 5.4. In Section 5.5, a summary of findings is provided.

5.2 MOTIVATING DATA

In a previous study, Volk et al. [66] identified a subset of schizophrenia subjects that consistently showed the most severe deficits in GAD67, parvalbumin, somatostatin and Lhx6 mRNA transcript levels. Lhx6 plays a critical role in the specification, migration, and maturation of neurons that express parvalbumin or somatostatin. GAD67 is the principal enzyme in the Gamma-AminoButyric Acid (GABA) synthesis system. The identification of this subset of schizophrenia subjects suggested that Lhx6 deficits may contribute to a failure of some parvalbumin and somatostatin neurons to successfully migrate or develop a detectable GABA-ergic phenotype (normal GAD67 expression level).

The defining data from the previous study consists of a sample of 42 pairs of schizophrenia and control subjects, where the 42 schizophrenia subjects were matched individually to one healthy control subject by gender, as closely as possible for age and post mortem interval. Samples from subjects in a pair were processed together throughout all stages of the study in order to control the experimental variation. To account for significant effects of covariates, somatostatin mRNA levels for each subject were adjusted to the average age at death of all subjects (47.6 year old) and Lhx6 mRNA levels were adjusted to the average tissue storage time (121 month) of all subjects. To account for varying scales among the four mRNAs, standardized mRNA levels of GAD67, parvalbumin, age-adjusted somatostatin, and tissue storage time-adjusted Lhx6 values were computed for all subjects by subtracting the overall mean and then dividing by the overall standard deviation. Cluster analysis was conducted using the standardized GAD67, parvalbumin, age-adjusted somatostatin, and tissue storage time-adjusted Lhx6 expression levels from each of the schizophrenia and control subjects (N=84) to determine whether a subset of subjects express these four transcripts in a distinct pattern. This cluster analysis ignored paring or diagnosis to examine if any cluster exists for the 84 subjects in the study. The average linkage method was used to cluster all 84 schizophrenia and control subjects (PROC CLUSTER in SAS 9.2; SAS Institute, Cary, N.C.). To check the robustness of the clustering found by the average linkage method, the expected maximum likelihood method was also used, and the same clusters were identified.

Two clusters were identified using the defining data. Figure 3 displays the cluster tree result of the cluster analysis using average linkage method. In one cluster composed of 61 subjects, 22 schizophrenia subjects and 39 control subjects were generally intermixed. However, the other cluster of 23 subjects was composed mostly of schizophrenia subjects (N=20 subjects), and only a few control subjects (N=3 subjects). This cluster contained 48% of all schizophrenia subjects (20/42) and only 7% of the entire control subjects (3/42). The 20 schizophrenia subjects in this cluster had lower levels of the four mRNA transcripts relative to other the 22 schizophrenia subjects in the intermixed cluster and to the 42 control subjects (see Volk et al. [66]). This cluster with 20 schizophrenia subjects expressing low levels of GABA markers was termed the Low-GABA-Marker (LGM) cluster, and the intermixed cluster was consequently termed the non-LGM cluster. The cluster analysis result is summarized in Table 20.

Figure 3: Two Cluster Identified in Volk et al. (2012)

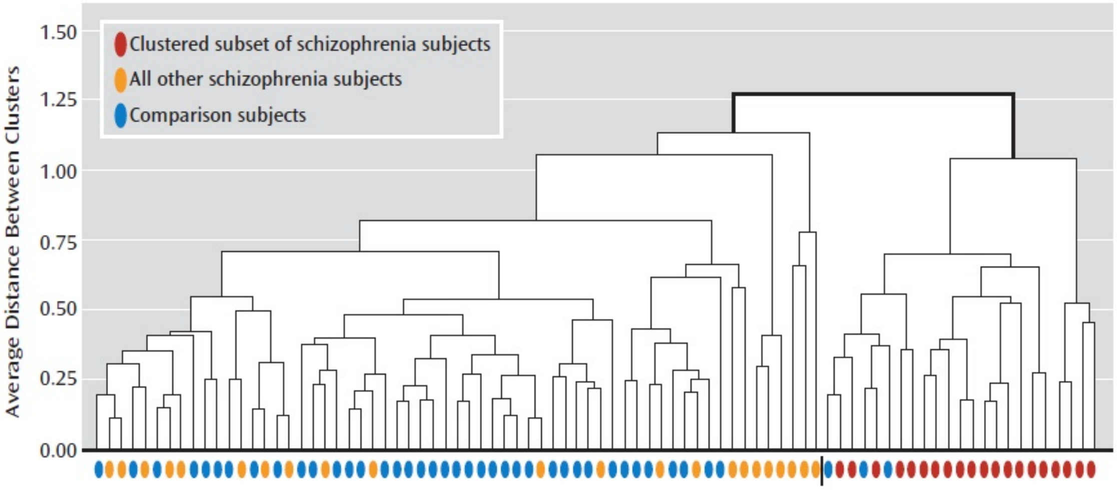


Table 20: Summary of Clustering Result on the Defining Data (N=84)

	Non-LGM Cluster	LGM Cluster	Total
Schizophrenia Subjects	22	20	42
Comparison Subjects	39	3	42
Total	61	23	84

To validate this LGM cluster finding, a new study was undertaken recently. The validating data set from the new study consists of the expression levels of the same 4 mRNA transcripts measured on another 20 matched pairs of schizophrenia and healthy control subjects. We are interested in whether the clusters previously identified in the defining data set, especially the LGM cluster composed mostly of schizophrenia subjects, are still present in the independent new validating data set. Importantly, we want to propose the methodology we will use for our assessment in advance of seeing the data to avoid any possibility of subjective biases.

5.3 CLUSTER VALIDATION USING KAPP AND TIBSHIRANI'S CLASSIFICATION APPROACH

5.3.1 Literature Review

There are many approaches proposed for validating the clustering analysis methods or the number of clusters in a single data set. However, the literature is relatively sparse concerning whether or not a cluster defined in previous data set (or defining data set) is still present in an independent new data set (validating data set). If the cluster is still present in the validating data set, this cluster is called "reproducible", and this means that this cluster may be biologically significant. Kapp and Tibshirani [27] proposed a general approach for individually validating clusters. In this section, we focus on reviewing Kapp and Tibshirani's approach.

While some of Kapp and Tibshirani's ideas are applicable, their entire cluster validation approach cannot be used to solve our problem. We discuss this in detail in Section 5.3.2. Modifications of their approach motivate our methods introduced in Section 5.3.2.

The procedure Kapp and Tibshirani [27] proposed tests in the validating data set H_0: There is no cluster structure; versus H_a: The previously defined cluster is valid. The procedure first classifies each of the subjects in the validating data set into one of the previously defined clusters. Then a cluster quality measure is obtained for each cluster. A suitable null distribution for the cluster quality measure is generated by simulation, and the p value is calculated based on this distribution.

We introduce each step in detail.

5.3.1.1 Classification When the validating data set and the defining data set are both measuring the same variables, it is suggested that the most appropriate approach for cluster validation analyses is to use a classifier made from the defining data [13] [14] [63]. Kapp and Tibshirani [27] proposed a classification rule, where the classifier is based on the averages of all variables over subjects within each cluster in the defining data set, called “centroids”. Based on the classification rule, each subject from the new data set can be classified into one of the previously defined clusters, whose centroids are most similar to the subject, with the possibility of not being able to classify a new subject into any of the previously identified cluster.

Following the notations used in Kapp and Tibshirani [27], we denote the defining data set by A, which is an $m \times n$ matrix, where m is the number of variables that are measured and n is the number of subjects in the defining data. Suppose that p clusters were found in the defining data. Let C be an $m \times p$ matrix, and the u^{th} column of C is the averages of the variables over the defining data in cluster u, where $u = 1, 2, \cdots, p$. The matrix C is called the centroids matrix. The validating data set is denoted by X, which is an $m \times q$ matrix, where q is the number of subjects in the validating data. A function $d(X[,j], C[,u])$ measuring the similarity between the j^{th} subject in the validating data set and the u^{th} centroids of the defining data is defined and a cutoff value, c, is selected. Kapp and Tibshirani [27] in their work let the similarity function be Pearson’s correlation. The correlations between a new subject and all centroids are computed. If the correlations are all smaller than c, then the subject is classified to a ”below-cutoff” group; otherwise, the subject is classified into the cluster whose correlation is the largest.

5.3.1.2 Cluster Quality Measure A number of cluster quality measures have been used within one data set to determine which clustering analysis procedure to use and the number of clusters present [13] [14] [63]. For independent validations, Kapp and Tibshirani [27] proposed a new cluster quality measure called In-Group Proportion (IGP), which is similar to cluster quality measures proposed by Tibshirani and Walther [63] and by Bailey and Dubes [2]. IGP is defined as “the proportion of (new) observations classified to a cluster whose nearest neighbor is also classified to the same cluster” [27]. Kapp and Tibshirani term the nearest neighbor for subject j to be the subject in the new data set who is most similar, e.g., most correlated if the “distance function” is Pearson’s correlation, to subject j. The nearest neighbor of subject j is denoted by j^N.

Other cluster quality measures were proposed by Chen et al. [7]. Kapp and Tishirani [27] compared IGP with four other cluster quality measures, homogeneity score, separation score, sihouette width, and weighted average discrepant pairs, proposed in Chen et al. [7], and they showed that IGP was the best measure for the purposes of validating previously identified clusters in a new data set.

5.3.1.3 Generating the Null Distribution The cluster validation procedure uses the IGP as the test statistics for the hypotheses H_0: there is no cluster structure; versus H_a: the previously defined cluster is valid. It is important to generate a null distribution of IGPs and compare the actual IGP obtained from the new data set with the null distribution.

Kapp and Tibshirani [27] proposed four versions of how to generate a null distribution of IGPs. The basic issue is to conceptually identify a ”least favorable” null distribution for the composite H_0. All of the four methods are based on repeatedly generating new centroid matrix C^* that corresponds to clusters that are placed randomly in the data. Therefore, the clusters defined by C^* are not likely to be high-quality clusters. There is one version that they found to have good performance and is widely applicable, and the null distribution generation takes the following steps: 1) Decompose C by singular value decomposition. $C = UDV^T$; 2) Define $C_1 = CV$; 3) Permute the columns of C_1 to obtain C_1^*; 4) Let $C^* = C_1^* V^T$; 5) Replace Z for C; 6) Calculate an IGP based on the newly generated C^* using the same classification; 7) Repeat step 1) to 6) for N times, e.g., $N = 100,000$, and the N IGPs build the null distribution.

5.3.1.4 Conclusion and Interpret the Results A high-quality cluster will have IGP close to 1, when the the subject and its neighbor are classified into the same cluster. Therefore, the p-value is defined as the proportion of null distribution IGPs that are greater than the actual IGP. The cluster is concluded valid if p-value is significant.

5.3.2 Methods

The question arises to whether the cluster validation procedure proposed by Kapp and Tibshirani is an appropriate procedure to be applied to our motivating data set. There is a difficulty because Kapp and Tibshirani's approach proposes to permute the columns of C when generating the null distribution, as discussed in Section 5.3.1.3, where m is the number of mRNA's being studied. In the microarray study, which motivates Kapp and Tibshirani, there are a larger number of mRNA's being studied at the same time. Consequently, m is large and the null distribution will be constructed from a large number, i.e., $m!$, of possible values of IGPs, which leads to a good null distribution. However, in our setting, there are only 4 mRNA's under consideration; thus, $m = 4$ and there will be only $m! = 4! = 24$ possible values for IGP. This is definitely not very useful when constructing a null distribution. In addition, in our setting (which Kapp and Tibshirani [27] don't consider), we need to show that the diagnostic results for the defined clusters in the validating data set are statistically similar to the previous Volk et al. [66] diagnostic results.

We extend Kapp and Tibshirani's idea of defining the clusters in the validation set based on the clusters from the defining data set. We propose to use the Mahalanobis Distance as the distance function for classification, instead of Pearson's Correlation used by Kapp and Tibshirani [27]. This is because our motivating data has many fewer dimensions than the microarray data considered by Kapp and Tibshirani, and we believe that Mahalanobis Distance is a more straight forward measurement of the distance. Also, we define the cut-off value c to be zero. During classification, the Mahalanobis Distance between a new subject and each centroid are computed, and the subject is classified into the cluster which is closest (smallest Mahalanobis Distance) to the subject.

To classify subjects in the validating data set using Kapp and Tibshirani's idea, we find the mean of defining data set's, i.e., from the initial 84 subjects, standardized values of GAD67, PV, age-adjusted SST and storage time-adjusted Lhx6, within each cluster. To be consistent, the mRNA levels in the validating data were also adjusted based on the adjusting slopes obtained from the defining data set and then standardized based on the new 40 subjects. We then compute the Mahalanobis distance of the standardized mRNA levels to each of the two clusters' mean standardized values for each subject in the validating data and classify that subject to the cluster with the shortest distance.

Our key issue is to show that the new clusters obtained from the validating data set have the same diagnostics distributions as the defining LGM cluster.

Assume that the validating data set is a random sample of schizophrenia and control subject pairs from the same population as the defning data set. To simplify our approach, we additionally assume that the previously defined two clusters are true LGM and non-LGM clusters with true proportions of schizophrenia subjects in each cluster. Since the previous study didn't consider pairing effect, and instead treated the 84 schizophrenia and control subjects as independent but did adjust for covariates, we also assume that there is no pairing effect in our validation study. Hence, we ignore subject parings and treat the 40 subjects in the validating data as independent.

The previously identified LGM cluster is composed mostly of schizophrenia subjects, which suggests LGM is a biological subtype of schizophrenia. Because we are assuming that the validating data is a random sample of schizophrenia and control subjects from the same population as the defining data, there should be reasonable number of new schizophrenia subjects classified into the LGM cluster and into the non-LGM cluster. In the validating data set, we want to test whether LGM is a biological subtype of schizophrenia, i.e., LGM cluster is composed mostly of schizophrenia subjects, rather than composed of generally intermixed schizophrenia and control subjects. The hypothesis testing considers H_0: The disease status (schizophrenia/control) and the found cluster type (LGM/non-LGM) are independent; versus H_a: There is a positive relationship between the disease status of schizophrenia and the cluster type LGM.

Suppose the validating study consists of N pairs of schizophrenia and control subjects. The $2N$ subjects are classified into the LGM cluster or the non-LGM cluster using Kapp and Tibshirani's idea only with the Mahalanobis distance, thus yielding our "found" LGM cluster. Let N_{LGM} subjects be classified into the LGM cluster and N_{nonLGM} subjects be classified into the non-LGM cluster, where $N_{LGM} + N_{nonLGM} = 2N$. Denote the number of schizophrenia subjects that are classified into the LGM cluster by X. The results can be summarized in the following 2×2 contingency table (Table 21).

Table 21: Classification Results for Random Samples

	Non-LGM Cluster	LGM Cluster	Total
Schizophrenia Subjects	$N - X$	X	N
Control Subjects	$2N - (N - X) - X - (N_{nonLGM} - X)$	$N_{nonLGM} - X$	N
Total	N_{nonLGM}	N_{LGM}	$2N$

Visual inspection of this table can provide strong support to the LGM finding based on the validating data set. However, we also want to formally test this. To understand the distribution of X under H_0: there is no relationship between the disease status and the cluster type, we consider random samples containing the same $2N$ subjects' data as the validating data set with the disease status of these subjects being randomly assigned, i.e., N subjects out of $2N$ are randomly chosen and assumed schizophrenia subjects while the other N subjects are assumed as control subjects. Then the same classification rule is applied based on the randomly assigned samples, and the null distribution is based upon the number of schizophrenia subjects that are classified into the LGM cluster, X, from these random samples.

For a specific validating data set with N pairs of schizophrenia and control subjects, we are using the 4 mRNA transcripts of the same $2N$ subjects and the same classification rule. Hence, there will always be a fixed number of subjects classified into the non-LGM cluster and a fixed number of subjects classified into the LGM cluster, i.e., N_{nonLGM} and N_{LGM} are fixed for a specific validating data set. Thus, the column totals in Table 21 are fixed. In addition, since we are randomly assigning N schizophrenia subjects and N control subjects, the row totals in Table 21 are fixed as well. Therefore, the constraints that the row and column marginal totals are fixed should be imposed. Fisher's exact test, often used to test equality of two binomial probabilities with these constraints [35], is consequently an appropriate test here. Under H_0, X follows a hypergeometric distribution,

$$P(X = x) = \frac{\binom{N}{x}\binom{N}{N_{LGM}-x}}{\binom{2N}{N_{LGM}}}. \tag{5.1}$$

The one-sided p-value of this Fisher's exact test is the probability that the number of schizophrenia subjects from the validating data set that are classified into the LGM cluster is greater than or equal to the observed number x. By (5.1), we have

$$p - value = P(X \geq x) = \sum_{X=x}^{N_{LGM}} \frac{\binom{N}{X}\binom{N}{N_{LGM}-X}}{\binom{2N}{N_{LGM}}}. \tag{5.2}$$

If p-value is smaller than the significance level $\alpha/2$, we conclude that the LGM cluster is mostly composed of the schizophrenia subjects, and hence we validate the biological significance of the LGM cluster.

5.3.3 Classification and Cluster Validation Result

Prior to our seeing the new data set we planned to use the general Kapp and Tibshirani approach [27] using our distance measures (Mahalanobis) and still being aware we would not be able to use their approach to obtain p-values given the difference between the type of our data and the micro-array considered by Kapp and Tibshirani. We classified the 40 new subjects into either the LGM or the non-LGM cluster. The result is that 12 subjects were classified into the LGM cluster, and 28 subjects were classified into the non-LGM cluster. There were 11 out of 20 schizophrenia subjects and 1 out of 20 control subjects in the validation data set being classified by our first method into the LGM cluster.The result is described in Table 22

Fisher's exact test with a one-sided alternative was performed to analyze whether the LGM cluster identified in the validating data set was again composed mostly of schizophrenia subjects, i.e., test for H_0: The disease status (schizophrenia/control) and the cluster type (LGM/non-LGM) are independent; versus H_a: There is a positive relationship between the disease status of schizophrenia and the cluster type LGM. The resulting p-value is

Table 22: Classification Results for the New Data Set

	Non-LGM Cluster	LGM Cluster	Total
Schizophrenia Subjects	9	11	20
Comparison Subjects	19	1	20
Total	28	12	40

$$p - value = P(X \geq 11) = \sum_{X=11}^{12} \frac{\binom{20}{X}\binom{20}{12-X}}{\binom{40}{12}} = 0.0006.$$

Since this p-value is smaller than 0.025, the defined LGM cluster in the validation data set was composed mostly of schizophrenia subjects (one-sided p=0.0006). We have statistically validated the biological significance of the LGM cluster in the new data set.

In addition, following the approach of Volk et al, mRNA levels were compared among the clusters arising from the validating data set using ANCOVA models. The 11 schizophrenia subjects classified to the LGM cluster had significantly lower levels of the four mRNA transcripts relative to the other 9 schizophrenia subjects in the intermixed cluster and to the 20 control subjects. The result will be included in an upcoming collaborative paper.

5.4 USING CLUSTER ANALYSIS TO VALIDATE THE PREVIOUS FINDINGS

The second approach directly applies clustering methodology to the validating data set or the combination of the defining and validating data sets.

5.4.1 On the Validating Data Set N=40

First, we directly apply clustering methodology to the validating data set ($N = 40$). To be consistent, the mRNA levels in the validating data were adjusted based on the adjusting slopes obtained from the defining data set and then standardized based on the new 40 subjects. The average linkage method is applied based on the standardized values of GAD67, PV, age-adjusted SST and storage time-adjusted Lhx6 on the schizophrenia and control subjects in the validating data set.

The clustering analysis on the validation data set led to 8 schizophrenia and no control subjects in the LGM cluster (see Figure 4, and Table 23). Fisher's exact test with a one-sided alternative was performed to analyze whether the LGM cluster identified in the validating data set was again composed mostly of schizophrenia subjects. The defined LGM cluster in the validation data set was composed mostly of schizophrenia subjects (one-sided p=0.0016). In addition, mRNA levels were compared among the clusters arising from the validating data set using ANCOVA models. The defined LGM cluster of schizophrenia subjects (N=8) did express significantly lower mRNA levels relative to other schizophrenia and comparison subjects. Furthermore, all the 8 schizophrenia subjects identified by this approach were also classified into the LGM cluster in Section 5.3.3.

A cluster analysis on the schizophrenia subjects of the validating data set and on the pairwise differences [(S-C)/C] of the validating data set led to similar results.

5.4.2 On the Combined Data Set N=124

We also conducted the same cluster analysis, i.e., using average linkage method, based on the overall 124 subjects to see if the two clusters are still present.

Since some covariates were previously found significantly related to mRNA expression levels, the covariate effects need to be examined in order to determine whether the 4 mRNA levels' adjustments were needed. Two approaches were used to examine the covariates - age, gender, PMI, PH, RIN, TST, effects and to adjust the mRNA expressions. In the primary approach, we adjusted the mRNA levels using the slopes estimated from all the 124 subjects. In the secondary approach, we adjusted the mRNA levels using the slopes previously estimated from the first study with 84 subjects. The two approaches led to similar results. In this dissertation, we only focus on the methods and results from the primary approach.

Figure 4: Clustering Analysis Result on the Validating Data (N=40)

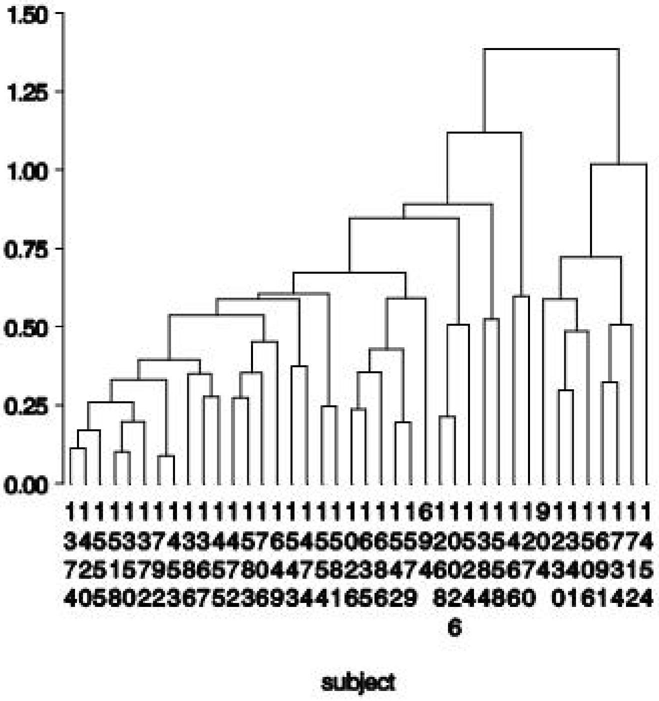


Table 23: Clustering Analysis Result on the Validating Data (N=40)

	Non-LGM Cluster	LGM Cluster	Total
Schizophrenia Subjects	12	8	20
Comparison Subjects	20	0	20
Total	32	8	40

Since the two studies were conducted separately, there was a concern whether the covariate effects were the same in two separate runs. Significant covariate*run interaction effects indicated that the covariate effects need to be adjusted differently within each run. Run effect also needs to be adjusted for the mRNA values.

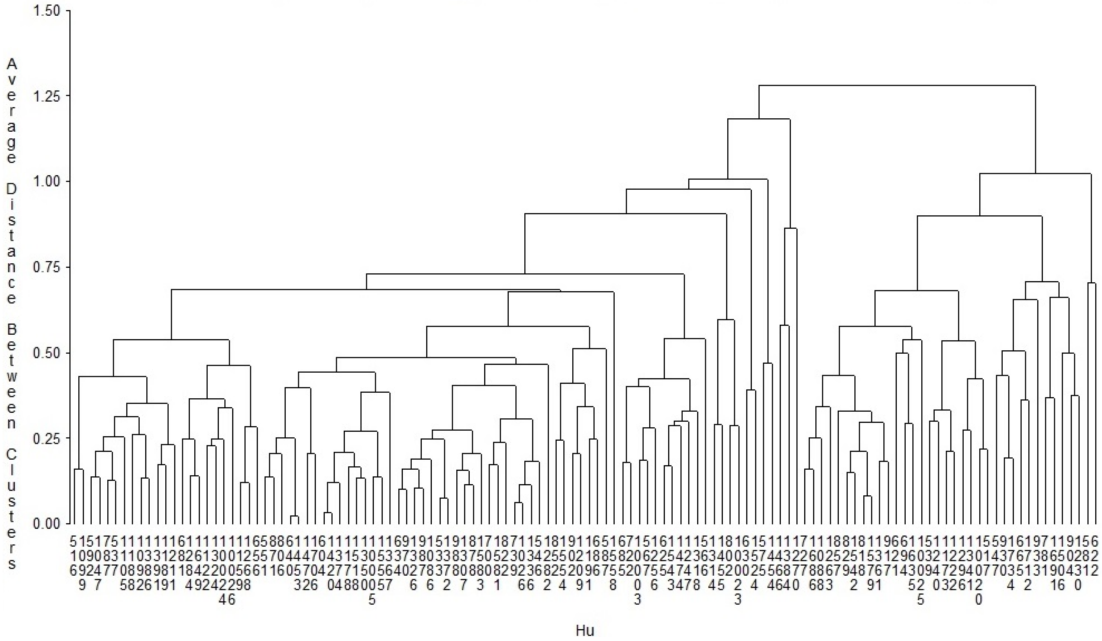
To account for varying scales among the four mRNAs, standardized values of the adjusted mRNA levels were computed over all 124 subjects by subtracting the overall mean and then dividing by the overall standard deviation. Standardized mRNA levels were used to cluster all 124 subjects using the average linkage method (SAS PROC CLUSTER, SAS, Cary, NC).

The clustering result is presented in the Figure 5 tree dendrogram, where we can see two distinct clusters from the dendrogram. Cluster 1 (on the left) consisted of 88 subjects (31 S + 57 C), which was a general mix of schizophrenia and control subjects, while cluster 2 (on the right) consisted of 36 subjects (31 S + 5 C), most of which were schizophrenia subjects.

A detailed summary of findings is given in Table 24. For example, the table indicates that among the 84 “original” subjects, 21 subjects in the new cluster 1 were in the previously defined LGM cluster and 3 were in the previously defined non-LGM cluster; whereas 58 subjects in the new cluster 2 were in the previously defined non-LGM cluster and 2 were in the previously defined LGM cluster. Thus, 79 of the 84 subjects in the “original” data were placed in the same cluster and 5 of 84 in a different cluster by the new clustering of the entire 124 subjects.

The cluster analysis (average linkage method) showed that there existed an ”LGM” cluster consisted mostly of schizophrenia subjects (31/36). In addition, mRNA levels were

Figure 5: Clustering Analysis Result on the Validating Data (N=124)



compared among the clusters arising from the validating data set using ANCOVA models. The defined LGM cluster of schizophrenia subjects (N=31) had significantly lower mRNA expression levels relative to other schizophrenia and comparison subjects.

5.5 SUMMARY OF FINDINGS

In summary, we used two approaches and each validated the LGM cluster finding in the previous study.

The first approach extended Kapp and Tibshirani's classification approach. Each subject from the validating data set was classified into either of the previously identified clusters. Fisher's exact test was then performed to show that the LGM cluster from the validating data set was again composed mostly of the schizophrenia subjects.

The second approach applied the average linkage clustering method for the validating data set (N=40) and the combined data set (N=124). As might be expected, the sec-

Table 24: Summary of Findings (N=124)

	In 84 subjects						In 40 subjects (KT classification)						Total		
	LGM			non-LGM			LGM			non-LGM					
	S	C	Total	S	C	Total	S	C	Total	S	C	Total	S	C	Total
Cluster 1	19	2	21	2	1	3	10	1	11	0	1	1	31	5	36
Cluster 2	1	1	2	20	38	58	1	0	1	9	18	27	31	57	88
Total	20	3	23	22	39	61	11	1	12	9	19	28	62	62	124

ond method again produced two clusters with one cluster clearly containing a subset of schizophrenia subjects with the same types of severe deficits in GAD67, parvalbumin, somatostatin and Lhx6 mRNA transcript levels as seen in the defining data set.

Therefore, the LGM subtype finding in schizophrenia population is seen to be a valid finding in the defining data set and the validating data set.

6.0 CONCLUSIONS AND FUTURE RESEARCH

6.1 CONCLUSIONS

In this dissertation, we first develop non-adaptive designs and corresponding statistical methodologies to evaluate the drug effects of two doses on two populations, and to simultaneously choose a dose and a population where the drug is beneficial. We propose designs that partially enrich the subpopulation, and we provid unbiased estimators of the drug effects on all populations.

When there are multiple doses and multiple populations, the biggest concern is how to conduct multiple comparisons while controlling the FWER in the strong sense. We propose testing schemes which are constructed under the principles of closed testing. We show that all our proposed testing procedures strongly protect FWER.

To accomplish various study goals based on differing prior beliefs of the drug effects, we establish testing schemes with flexible testing orders. There is no unique testing order that is appropriate for all study goals and all prior beliefs of the drug effects. We can decide which ordering to use before the study begins based on the study goal and our prior beliefs of the drug effects using simulation studies. This flexibility allows us to pick the ordering that will lead to the largest power, where power is a more generalized concept.

We also propose adaptive designs to add more flexibility to our design. Following Wang et al. [68], we propose possible adaptation rules in the dissertation based on evaluating the conditional power after Stage I. Limited simulation studies are conducted to show how to find appropriate adaption rules.

For each individual hypothesis in the testing scheme, we suggest using Follmann's testing procedure if the distribution of test statistics is multivariate normal under the null hypothesis, and we also derive alternations of Follmann's testing procedure when the distribution of test statistics under the null hypothesis is not multivariate normal.

In Chapter 5, we use two approaches where each validates a Low GABA Marker (LGM) schizophrenia subset finding that was identified in the previous study in which we were involved. In our first approach, we extend Kapp and Tibshirani's idea of defining the clusters in the validating data set based on the clusters from the defining data set. The second approach applies the clustering analysis as originally used in the previous study for the validating data set and the combination of the validating and the defining data sets. Both the two approaches produce two clusters with one cluster clearly containing a subset of schizophrenia subjects with the same types of severe deficits in the four mRNA transcripts as seen in the defining data set. Therefore the LGM subtype finding in schizophrenia population is seen to be a valid finding in the defining data set and the validating data set.

6.2 FUTURE RESEARCH

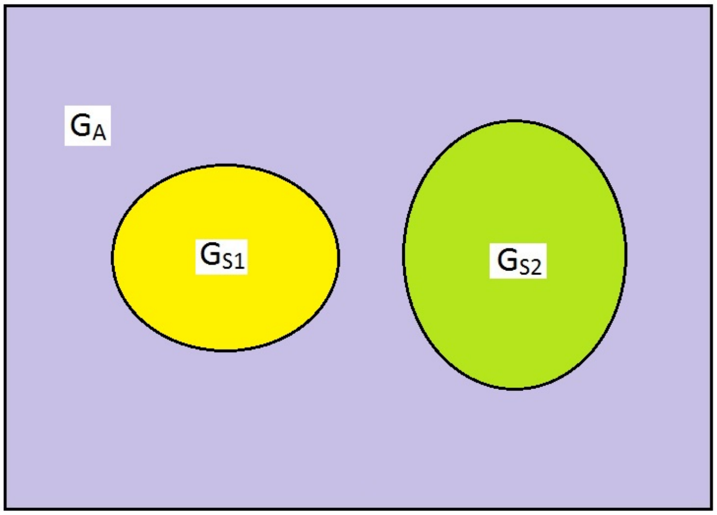
In Chapter 3 and Chapter 4, we propose methodologies for simultaneously selecting the desired population and dose for multiple doses and multiple nested subpopulations. In the future, we plan to consider parallel subpopulations.

For example, let's consider 2 doses: low and high; and 3 populations: the overall population (G_A), subpopulation 1 (G_{S1}), and subpopulation 2 (G_{S2}). The two subpopulations are nested in the overall population, and they are parallel to each other. The two subpopulations can be overlapped or discrete. In this dissertation, we only introduce discrete parallel subpopulations (See Figure 6).

We are interested in testing if the there is any positive drug effects, denoted by Δ as in previous chapters, on $\{G_A, L\}$, $\{G_A, H\}$, $\{G_{S1}, L\}$, $\{G_{S1}, H\}$, $\{G_{S2}, L\}$, $\{G_{S2}, H\}$, which are the low dose of the overall population, the high dose of the overall population, the low dose of the subpopulation 1, the high dose of the subpopulation 1, the low dose on the subpopulation 2, and the high dose of the subpopulation 2, respectively.

Our previously proposed step down testing scheme will not work for the parallel subpop-

Figure 6: Discrete Parallel Subopulations



ulations setting since often there is no preference between the two subpopulations. A possible solution is to use a ”tree” testing scheme constructed under the closed testing scheme. Suppose there is no severe side effects for the high dose on any of these populations and there is no preference of concluding one of the two subpopulations rather than the other. The goal of the sponsor is to find the largest population, and then to find the lowest dose for this largest population. This means, if either the low dose or the high dose of the overall population is effective, we want to conclude the overall population and then find the MED for the overall population. If neither dose is effective for the overall population, we want to find if low dose is effective for either of the two subpopulations. If yes, then we want to conclude low dose for the subpopulation/subpopulations; otherwise, we want to find out whether the high dose is effective for any of the two subpopulations.

There can be many possible testing schemes. We only list one here in order to show the possibility of building a tree testing scheme under the closed testing scheme. The null hypotheses are

$$
\begin{aligned}
&H_0^{(6)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} = \Delta_{G_{S2},H} = 0,\\
&H_0^{(5a)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S1},H} = \Delta_{G_{S2},L} = 0,\\
&H_0^{(5b)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S2},L} = \Delta_{G_{S2},H} = 0,\\
&H_0^{(4)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = \Delta_{G_{S2},L} = 0,\\
&H_0^{(3a)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_{S1},L} = 0,\\
&H_0^{(3b)}: \Delta_{G_A,L} = \Delta_{G_A,H} = \Delta_{G_{S2},L} = 0,\\
&H_0^{(2)}: \Delta_{G_A,L} = \Delta_{G_A,H} = 0,\\
&H_0^{(1)}: \Delta_{G_A,L} = 0. \qquad (6.1)
\end{aligned}
$$

Each corresponding alternative hypothesis states that there is at least one positive drug effect among the populations and doses considered in the null hypothesis.

There are 6 levels of the null hypotheses, indicated as (6) to (1) in (6.1), where there is a single hypothesis at level $(6), (4), (2)$ and (1), and two null hypotheses at level (5) and level (3). We test the family of hypotheses in a step down manner from the highest level until we accept at least one null hypothesis at one level. We are able to prove this testing scheme is closed under intersection. There is a one-to-one correspondence decision rule, which is shown in Table 6.2.

There is a possible ambiguity for the testing schemes. For example, when $H^{(6)}$ is rejected, we move one level down to test $H^{(5a)}$ and $H^{(5b)}$. There is chance that both $H^{(5a)}$ and $H^{(5b)}$ could be both accepted, which provides contradictory results to rejecting $H^{(6)}$. This suggests that in the future we need to modify our testing scheme or make relative assumptions to deal with this contradiction. This is only a very brief idea of building testing schemes for the parallel subpopulation settings and clearly much more work remains. In the future, we will develop methodologies in detail for the study design and data analysis for both non-adaptive and adaptive designs in order to simultaneously select the desired population and dose for multiple doses and multiple parallel subpopulations.

Table 25: Outcomes of the Procedure and Decision Rule for Parallel Subpopulations.

$H_0^{(6)}$	$H_0^{(5a)}$	$H_0^{(5b)}$	$H_0^{(4)}$	$H_0^{(3a)}$	$H_0^{(3b)}$	$H_0^{(2)}$	$H_0^{(1)}$	$\Delta_{G_A,L}$	$\Delta_{G_A,H}$	$\Delta_{G_{S1},L}$	$\Delta_{G_{S1},H}$	$\Delta_{G_{S2},L}$	$\Delta_{G_{S2},H}$	Conclusion
ACC	NT	NT	NT	NT	NT	NT	NT	0	0	0	0	0	0	Empty
REJ	REJ	ACC	NT	NT	NT	NT	NT	0	0	0	>0	0	0	$\{G_{S1}$, H$\}$
REJ	ACC	REJ	NT	NT	NT	NT	NT	0	0	0	0	0	>0	$\{G_{S2}$, H$\}$
REJ	REJ	REJ	ACC	NT	NT	NT	NT	0	0	0	>0	0	>0	$\{G_{S1}$, H$\}$ and $\{G_{S2}$, H$\}$
REJ	REJ	REJ	REJ	REJ	ACC	NT	NT	0	0	>0	UKN	0	UKN	$\{G_{S1}$, L$\}$
REJ	REJ	REJ	REJ	ACC	REJ	NT	NT	0	0	0	UKN	>0	UKN	$\{G_{S2}$, L$\}$
REJ	REJ	REJ	REJ	REJ	REJ	ACC	NT	0	0	>0	UKN	>0	UKN	$\{G_{S1}$, L$\}$ and $\{G_{S2}$, L$\}$
REJ	REJ	REJ	REJ	REJ	REJ	REJ	ACC	0	>0	UKN	UKN	UKN	UKN	$\{G_A$,H$\}$
REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	>0	UKN	UKN	UKN	UKN	UKN	$\{G_A$, L$\}$

Note: ACC=Accept, REJ=Reject, NT=Not Tested, UKN=Unknown.

APPENDIX

NON-ADAPTIVE DESIGN FOR MULTIPLE DOSES AND TWO NESTED POPULATIONS

A.1 3 POPULATION AND 3 DOSE CASE

A.1.0.1 Unbiased Estimators of Drug Effects

It is natural to estimate the mean responses at each dose in population G_{S1^-}, $G_{S2^{1-}}$ and G_{S2} by the corresponding sample means.

$$\hat{\mu}_{G_l,m} = \bar{X}_{G_l,m}, \quad \text{for} \quad l = S1^-, S2^{1-}, S2; m = L, M, H, c.$$

Since f_1, the proportion of population G_{S1} to the overall population G_A, and f_2, the proportion of population G_{S2} to population G_{S1}, are assumed known, we estimate the mean responses in population G_{S1} and in the overall population G_A by the weighted averages.

For $m = L, M, H, c$, the estimated mean drug response in population G_{S1} are

$$\hat{\mu}_{G_{S1},m} = f_2\bar{X}_{G_{S2},m} + (1-f_2)\bar{X}_{G_{S2^{1-}},m};$$

For $m = L, M, H, c$, the estimated mean drug response in population G_A are

$$\begin{aligned}\hat{\mu}_{G_A,m} =& f_1\hat{\mu}_{G_{S1},m} + (1-f_1)\hat{\mu}_{G_{S1^-},m}\\ =& f_1[f_2\bar{X}_{G_{S2},m} + (1-f_2)\bar{X}_{G_{S2^{1-}},m}] + (1-f_1)\bar{X}_{G_{S1^-},m}\\ =& f_1f_2\bar{X}_{G_{S2},m} + f_1(1-f_2)\bar{X}_{G_{S2^{1-}},m} + (1-f_1)\bar{X}_{G_{S1^-},m}.\end{aligned}$$

Then for $m = L, M, H$, the drug effects $\Delta_{G_{S2},m}$ for population G_{S2} are estimated by

$$\hat{\Delta}_{G_{S2},m} = \hat{\mu}_{G_{S2},m} - \hat{\mu}_{G_{S2},c} = \bar{X}_{G_{S2},m} - \bar{X}_{G_{S2},c};$$

For $m = L, M, H$, the drug effects $\Delta_{G_{S1},m}$ for population G_{S1} are estimated by

$$\begin{aligned}\hat{\Delta}_{G_{S1},m} =& \hat{\mu}_{G_{S1},m} - \hat{\mu}_{G_{S1},c} \\ =& f_2\bar{X}_{G_{S2},m} + (1-f_2)\bar{X}_{G_{S2^{1-}},m} - f_2\bar{X}_{G_{S2},c} - (1-f_2)\bar{X}_{G_{S2^{1-}},c};\end{aligned}$$

For $m = L, M, H$, the drug effects $\Delta_{G_A,m}$ for population G_A are estimated by

$$\begin{aligned}\hat{\Delta}_{G_A,m} =& \hat{\mu}_{G_A,m} - \hat{\mu}_{G_A,c} \\ =& f_1f_2\bar{X}_{G_{S2},m} + f_1(1-f_2)\bar{X}_{G_{S2^{1-}},m} + (1-f_1)\bar{X}_{G_{S1^-},m} \\ & - f_1f_2\bar{X}_{G_{S2},c} - f_1(1-f_2)\bar{X}_{G_{S2^{1-}},c} - (1-f_1)\bar{X}_{G_{S1^-},c}.\end{aligned}$$

Since $\bar{X}_{G_l,m}$ is the unbiased estimator of $\mu_{G_l,m}$ on populations $G_{S1^-}, G_{S2^{1-}}, G_{S2}$, i.e.,

$$E(\hat{\mu}_{G_l,m}) = E(\bar{X}_{G_l,m}) = \mu_{G_l,m}, \text{ for } l = S1^-, S2^{1-}, S2; m = L, M, H, c,$$

we have,

$$E(\hat{\mu}_{G_{S2},m}) = E(\bar{X}_{G_{S2},m}) = \mu_{G_{S2},m},$$

$$\begin{aligned}E(\hat{\mu}_{G_{S1},m}) =& E[f_2\bar{X}_{G_{S2},m} + (1-f_2)\bar{X}_{G_{S2^{1-}},m}] \\ =& f_2\mu_{G_{S2},m} + (1-f_2)\mu_{G_{S2^{1-}},m} \\ =& \mu_{G_{S1},m},\end{aligned}$$

and

$$\begin{aligned}E(\hat{\mu}_{G_A,m}) =& E[f_1f_2\bar{X}_{G_{S2},m} + f_1(1-f_2)\bar{X}_{G_{S2^{1-}},m} + (1-f_1)\bar{X}_{G_{S1^-},m}] \\ =& f_1f_2\mu_{G_{S2},m} + f_1(1-f_2)\mu_{G_{S2^{1-}},m} + (1-f_1)\mu_{G_{S1^-},m} \\ =& \mu_{G_A,m},\end{aligned}$$

are also unbiased estimators for $m = L, M, H, c$.

Therefore, $\hat{\Delta}_{G_l,m}$ are unbiased estimators for $\Delta_{G_l,m}$, in that,

$$E(\hat{\Delta}_{G_l,m}) = E(\hat{\mu}_{G_l,m} - \hat{\mu}_{G_l,c}) = \mu_{G_l,m} - \mu_{G_l,c} = \Delta_{G_A,m},$$

for $l = A, S1, S2; m = L, M, H$.

A.1.1 Test Statistics and The Covariance

Next, we calculate suitable test statistics for the drug effects $\boldsymbol{\Delta}$, where

$$\boldsymbol{\Delta} = [\Delta_{G_A,L}, \Delta_{G_A,M}, \Delta_{G_A,H}, \Delta_{G_{S1},L}, \Delta_{G_{S1},M}, \Delta_{G_{S1},H}, \Delta_{G_{S2},L}, \Delta_{G_{S2},M}, \Delta_{G_{S2},H}]'.$$

Since the responses are normally distributed, and the variance is known, $\mathbf{Z}$ test statistics are used. The test statistics of the drug effects at dose for population G_{S2} are computed as, for $m = L, M, H$,

$$Z_{G_{S2},m} = \frac{\hat{\Delta}_{G_{S2},m}}{\sqrt{\sigma^2_{(\hat{\Delta}_{G_{S2},m})}}} = \frac{\bar{X}_{G_{S2},m} - \bar{X}_{G_{S2},c}}{\sqrt{\sigma^2_{(\bar{X}_{G_{S2},m} - \bar{X}_{G_{S2},c})}}} = \frac{\bar{X}_{G_{S2},m} - \bar{X}_{G_{S2},c}}{\sqrt{\sigma^2(\frac{1}{g_1g_2N} + \frac{1}{g_1g_2N})}} = \frac{\bar{X}_{G_{S2},m} - \bar{X}_{G_{S2},c}}{\sqrt{\frac{2\sigma^2}{g_1g_2N}}}.$$

The test statistics of the drug effects for population G_{S1} are computed as, for $m = L, M, H$,

$$\begin{aligned}
Z_{G_{S1},m} &= \frac{\hat{\Delta}_{G_{S1},m}}{\sqrt{\sigma^2_{(\hat{\Delta}_{G_{S1},m})}}} \\
&= \frac{f_2\bar{X}_{G_2,m} + (1-f_2)\bar{X}_{G_{_{21-}},m} - f_2\bar{X}_{G_2,c} - (1-f_2)\bar{X}_{G_{_{21-}},c}}{\sqrt{\sigma^2_{f_2\bar{X}_{G_2,m}+(1-f_2)\bar{X}_{G_{_{21-}},m}-f_2\bar{X}_{G_2,c}-(1-f_2)\bar{X}_{G_{_{21-}},c}}}} \\
&= \frac{f_2\bar{X}_{G_{S2},m} + (1-f_2)\bar{X}_{G_{_{S21-}},m} - f_2\bar{X}_{G_{S2},c} - (1-f_2)\bar{X}_{G_{_{S21-}},c}}{\sqrt{\sigma^2[\frac{^2_2}{g_1g_2N} + \frac{(1-f_2)^2}{g_1(1-g_2)N} + \frac{f_2^2}{g_1g_2N} + \frac{(1-f_2)^2}{g_1(1-g_2)N}]}} \\
&= \frac{f_2\bar{X}_{G_{S2},m} + (1-f_2)\bar{X}_{G_{_{S21-}},m} - f_2\bar{X}_{G_{S2},c} - (1-f_2)\bar{X}_{G_{_{S21-}},c}}{\sqrt{2\sigma^2[\frac{f_2^2}{g_1g_2N} + \frac{(1-f_2)^2}{g_1(1-g_2)N}]}}.
\end{aligned}$$

The test statistics of the drug effects for the overall population are computed as, for $m = L, M, H$,

$$\begin{aligned}
Z_{G_A,m} &= \frac{\hat{\Delta}_{G_A,m}}{\sqrt{\sigma^2_{(\hat{\Delta}_{G_A,m})}}} \\
&= \frac{f_1\hat{\mu}_{G_{S1},m} + (1-f_1)\bar{X}_{G_{_{S1-}},m} - f_1\hat{\mu}_{G_{S1},m} - (1-f_1)\bar{X}_{G_{_{S1-}},c}}{\sqrt{\sigma_{f_1f_2\bar{X}_{G_{S2},m}+f_1(1-f_2)\bar{X}_{G_{_{S21-}},m}+(1-f_1)\bar{X}_{G_{_{S1-}},m}-f_1f_2\bar{X}_{G_{S2},c}-f_1(1-f_2)\bar{X}_{G_{_{S21-}},c}-(1-f_1)\bar{X}_{G_{_{S1-}},c}}}} \\
&= \frac{f_1\hat{\mu}_{G_{S1},m} + (1-f_1)\bar{X}_{G_{_{S1-}},m} - f_1\hat{\mu}_{G_{S1},m} - (1-f_1)\bar{X}_{G_{_{S1-}},c}}{\sqrt{2\sigma^2[\frac{(f_1f_2)^2}{g_1g_2N} + \frac{[f_1(1-f_2)]^2}{g_1(1-g_2)N} + \frac{(1-f_1)^2}{(1-g_1)N}]}}.
\end{aligned}$$

Since the sample means $\bar{X}_{G_{S1-},L}$, $\bar{X}_{G_{S1-},M}$, $\bar{X}_{G_{S1-},H}$, $\bar{X}_{G_{S2^{1-}},L}$, $\bar{X}_{G_{S2^{1-}},M}$, $\bar{X}_{G_{S2^{1-}},H}$, $\bar{X}_{G_{S2},L}$, $\bar{X}_{G_{S2},M}$, and $\bar{X}_{G_{S2},H}$ are mutually independent, it follows that

$$cov(\bar{X}_{G_l,m}, \bar{X}_{G_{l'},m'}) = 0,$$

for $l, l' = S1^-, S2^{1-}, S2$; $m, m' = L, M, H, c$; $\{l, m\} \neq \{l', m'\}$.

Since $Z_{G_l,m}$'s are standardized scores, the variance of each $Z_{G_l,m}$ is 1, and the covariance are computed in the appendix.

Let $\mathbf{Z} = [Z_{G_A,L}, Z_{G_A,M}, Z_{G_A,H}, Z_{G_{S1},L}, Z_{G_{S1},M}, Z_{G_{S1},H}, Z_{G_{S2},L}, Z_{G_{S2},M}, Z_{G_{S2},H}]'$. Normality assumption yields that

$$\mathbf{Z} \sim \mathcal{MVN}(\mu_{\mathbf{Z}}, \mathbf{\Sigma}),$$

where

$$\mu_{\mathbf{Z}} = \begin{pmatrix} \frac{\Delta_{G_A,L}}{\sqrt{2\sigma^2[\frac{(f_1f_2)^2}{g_1g_2N} + \frac{[f_1(1-f_2)]^2}{g_1(1-g_2)N} + \frac{(1-f_1)^2}{(1-g_1)N}]}} \\ \frac{\Delta_{G_A,M}}{\sqrt{2\sigma^2[\frac{(f_1f_2)^2}{g_1g_2N} + \frac{[f_1(1-f_2)]^2}{g_1(1-g_2)N} + \frac{(1-f_1)^2}{(1-g_1)N}]}} \\ \frac{\Delta_{G_A,H}}{\sqrt{2\sigma^2[\frac{(f_1f_2)^2}{g_1g_2N} + \frac{[f_1(1-f_2)]^2}{g_1(1-g_2)N} + \frac{(1-f_1)^2}{(1-g_1)N}]}} \\ \frac{\Delta_{G_{S1},L}}{\sqrt{2\sigma^2[\frac{f_2^2}{g_1g_2N} + \frac{(1-f_2)^2}{g_1(1-g_2)N}]}} \\ \frac{\Delta_{G_{S1},M}}{\sqrt{2\sigma^2[\frac{f_2^2}{g_1g_2N} + \frac{(1-f_2)^2}{g_1(1-g_2)N}]}} \\ \frac{\Delta_{G_{S1},H}}{\sqrt{2\sigma^2[\frac{f_2^2}{g_1g_2N} + \frac{(1-f_2)^2}{g_1(1-g_2)N}]}} \\ \frac{\Delta_{G_{S2},L}}{\sqrt{\frac{2\sigma^2}{g_1g_2N}}} \\ \frac{\Delta_{G_{S2},M}}{\sqrt{\frac{2\sigma^2}{g_1g_2N}}} \\ \frac{\Delta_{G_{S2},H}}{\sqrt{\frac{2\sigma^2}{g_1g_2N}}} \end{pmatrix},$$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{D_1} & \mathbf{D_{12}} & \mathbf{D_{13}} \\ \mathbf{D_{21}} & \mathbf{D_2} & \mathbf{D_{23}} \\ \mathbf{D_{31}} & \mathbf{D_{23}} & \mathbf{D_3} \end{pmatrix},$$

where

$$\mathbf{D_1} = \mathbf{D_2} = \mathbf{D_3} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix},$$

$$\mathbf{D_{12}} = \mathbf{D_{21}} = \begin{pmatrix} d_{12} & \frac{1}{2d_{12}} & \frac{1}{2d_{12}} \\ \frac{1}{2d_{12}} & d_{12} & \frac{1}{2d_{12}} \\ \frac{1}{2d_{12}} & \frac{1}{2d_{12}} & d_{12} \end{pmatrix},$$

$$\mathbf{D_{13}} = \mathbf{D_{31}} = \begin{pmatrix} d_{13} & \frac{1}{2d_{13}} & \frac{1}{2d_{13}} \\ \frac{1}{2d_{13}} & d_{13} & \frac{1}{2d_{13}} \\ \frac{1}{2d_{13}} & \frac{1}{2d_{13}} & d_{13} \end{pmatrix},$$

$$\mathbf{D_{23}} = \mathbf{D_{32}} = \begin{pmatrix} d_{23} & \frac{1}{2d_{23}} & \frac{1}{2d_{23}} \\ \frac{1}{2d_{23}} & d_{23} & \frac{1}{2d_{23}} \\ \frac{1}{2d_{23}} & \frac{1}{2d_{23}} & d_{23} \end{pmatrix},$$

$$d_{12} = \sqrt{\frac{1}{1 + \frac{(1-f_1)^2}{f_1^2} \frac{1}{f_2^2 \frac{1-g_1}{g_1 g_2} + (1-f_2)^2 \frac{1-g_1}{g_1(1-g_2)}}}},$$

$$d_{13} = \sqrt{\frac{1}{1 + \frac{(1-f_2)^2}{f_2^2} \frac{g_2}{1-g_2} + \frac{(1-f_1)^2}{(f_1 f_2)^2} \frac{g_1 g_2}{1-g_1}}},$$

$$d_{23} = \sqrt{\frac{1}{1 + \frac{(1-f_2)^2}{f_2^2} \frac{g_2}{1-g_2}}}.$$

Under the null hypothesis,

$$\mathbf{Z} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}).$$

BIBLIOGRAPHY

[1] Amos, A. F., McCarty, D. J., and Zimmet, P. (1997). The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic Medicine 14*, S7–S85.

[2] Bailey Jr, T. A., and Dubes, R. (1982). Cluster validity profiles. *Pattern Recognition 15*, 61–83.

[3] Baselga, J. (2001). Clinical trials of herceptin (trastuzumab). *European Journal of Cancer 37*, 18–24.

[4] Baselga, J. (2001). Herceptin alone or in combination with chemotherapy in the treatment of her2-positive metastatic breast cancer: pivotal trials. *Oncology 61*, 14–21.

[5] Bauer, P., and Koenig, F. (2005). The reassessment of trial perspectives from interim data: a critical view. *Statistics in Medicine 25*, 23–36.

[6] Bauer, P., and Röhmel, J. (2007). An adaptive method for establishing a dose-response relationship. *Statistics in Medicine 14*, 1595–1607.

[7] Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S., and Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica 12*, 241–262.

[8] Chow, S.-C., and Chang, M. (2008). Adaptive design methods in clinical trials–a review. *Orphanet Journal of Rare Disease 3*, 11.

[9] Cody, W. J. (1969). Rational chebyshev approximations for the error function. *Mathematics of Computation 23*, 631–637.

[10] Cohen, J. (2003). Vaccine results lose significance under scrutiny. *Science 299*, 1495–1495.

[11] De Onis, M., and Blössner, M. (2003). The World Health Organization global database on child growth and malnutrition: methodology and applications. *International Journal of Epidemiology 32*, 518–526.

[12] Di Fiore, F., Blanchard, F., Charbonnier, F., Le Pessot, F., Lamy, A., Galais, M., Bastit, L., Killian, A., Sesboüé, R., Tuech, J., et al. (2007). Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by Cetuximab plus chemotherapy. *British Journal of Cancer 96*, 1166–1169.

[13] Dudoit, S., and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology 3*, research0036.

[14] Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association 97*, 77–87.

[15] Dunnett, C. W., and Tamhane, A. C. (1998). Some new multiple-test procedures for dose finding. *Journal of Biopharmaceutical Statistics 8*, 353–366.

[16] Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association 91*, 854–861.

[17] Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., and Fawcett, J.(2010). Antidepressant drug effects and depression severity. *JAMA: The Journal of the American Medical Association 303*, 47–53.

[18] Freidlin, B., and Simon, R. (2005). 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.

[19] Giaccone, G., Herbst, R. S., Manegold, C., Scagliotti, G., Rosell, R., Miller, V., Natale, R. B., Schiller, J. H., von Pawel, J., Pluzanska, A., et al. (2004). Gefitinib in combination with gemcitabine and cisplatin in advanced non–small-cell lung cancer: A phase III trial: INTACT 1. *Journal of Clinical Oncology 22*, 777–784.

[20] Graybill, F. A. (1976). *Theory and applications of the linear model*. Duxbury.

[21] Grouin, J.-M., Coste, M., and Lewis, J. (2005). Subgroup analyses in randomized clinical trials: statistical and regulatory issues. *Journal of Biopharmaceutical Statistics 15*, 869–882.

[22] Herbst, R. S., Giaccone, G., Schiller, J. H., Natale, R. B., Miller, V., Manegold, C., Scagliotti, G., Rosell, R., Oliff, I., Reeves, J. A., et al. (2004). Gefitinib in combination with paclitaxel and carboplatin in advanced non–small-cell lung cancer: A phase III trial: INTACT 2. *Journal of Clinical Oncology 22*, 785–794.

[23] Hochberg, Y., and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, Inc.

[24] Hung, H. J., Cui, L., Wang, S.-J., and Lawrence, J. (2005). Adaptive statistical analysis following sample size modification based on interim review of effect size. *Journal of Biopharmaceutical Statistics 15*, 693–706.

[25] Hylek, E. M., Evans-Molina, C., Shea, C., Henault, L. E., and Regan, S. (2007). Major hemorrhage and tolerability of warfarin in the first year of therapy among elderly patients with atrial fibrillation. *Circulation 115*, 2689–2696.

[26] Jiang, W., Freidlin, B., and Simon, R. (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute 99*, 1036–1043.

[27] Kapp, A. V., and Tibshirani, R. (2007). Are clusters found in one dataset present in another dataset? *Biostatistics 8*, 9–31.

[28] Kearney, P. M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P. K., He, J., et al. (2005). Global burden of hypertension: analysis of worldwide data. *Lancet 365*, 217–223.

[29] Kelly, T., Yang, W., Chen, C., Reynolds, K., and He, J. (2008). Global burden of obesity in 2005 and projections to 2030. *International journal of obesity 32*, 1431–1437.

[30] Khambata-Ford, S., Garrett, C. R., Meropol, N. J., Basik, M., Harbison, C. T., Wu, S., Wong, T. W., Huang, X., Takimoto, C. H., Godwin, A. K., et al. (2007). Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *Journal of Clinical Oncology 25*, 3230–3237.

[31] Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., and Johnson, B. T. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. *PLoS Medicine 5*, e45.

[32] Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika 50*, 403–418.

[33] Lapham, S., Montgomery, K., and Hoy, W. (1990). HMO databases: Fertile ground for epidemiological research. *Comput Healthc 11*, 18–20.

[34] Laska, E. M., and Meisner, M. J. (1989). Testing whether an identified treatment is best. *Biometrics 45*, 1139–1151.

[35] Lehmann, E. L., and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.

[36] Lievre, A., Bachet, J.-B., Le Corre, D., Boige, V., Landi, B., Emile, J.-F., Côté, J.-F., Tomasic, G., Penna, C., Ducreux, M., et al. (2006). KRAS mutation status is predictive of response to Cetuximab therapy in colorectal cancer. *Cancer Research 66*, 3992–3995.

[37] Liu, Q., and Chi, G. Y. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics 57*, 172–177.

[38] Malani, A., Bembom, O., and Van der Laan, M. (2008). Reforming subgroup analysis. *Available at SSRN 1119970*.

[39] Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

[40] Marsaglia, G. (2004). Evaluating the normal distribution. *Journal of Statistical Software 11*, 1–7.

[41] Mok, T. S., Wu, Y.-L., Thongprasert, S., Yang, C.-H., Chu, D.-T., Saijo, N., Sunpaweravong, P., Han, B., Margono, B., Ichinose, Y., et al. (2009). Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine 361*, 947–957.

[42] Morrison, D. (1976). *Multivariate Statistical Methods*. McGraw-Hill.

[43] Muss, H. B., Thor, A. D., Berry, D. A., Kute, T., Liu, E. T., Koerner, F., Cirrincione, C. T., Budman, D. R., Wood, W. C., Barcos, M., et al. (1994). c-erbb-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *New England Journal of Medicine 330*, 1260–1266.

[44] Ogino, S., Kawasaki, T., Brahmandam, M., Yan, L., Cantor, M., Namgyal, C., Mino-Kenudson, M., Lauwers, G. Y., Loda, M., and Fuchs, C. S. (2005). Sensitive sequencing method for KRAS mutation detection by pyrosequencing. *The Journal of Molecular Diagnostics 7*, 413–421.

[45] Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics 40*, 549–567.

[46] Petersen, P. E., Bourgeois, D., Ogawa, H., Estupinan-Day, S., and Ndiaye, C. (2005). The global burden of oral diseases and risks to oral health. *Bulletin of the World Health Organization 83*, 661–669.

[47] Powers, J. H., Ross, D. B., Lin, D., and Soreth, J. (2004). Linezolid and vancomycin for methicillin-resistant staphylococcus aureus nosocomial pneumonialinezolid and vancomycin for methicillin-resistant staphylococcus aureus nosocomial pneumonia: the subtleties of subgroup analyses. *CHEST Journal 126*, 314–316.

[48] Proschan, M. A., and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics 51*, 1315–1324.

[49] Proschan, M. A., Lan, K. G., and Wittes, J. T. (2007). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer.

[50] Rosenblum, M., and van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika 98*, 845–860.

[51] Ruberg, S. J. (1989). Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association 84*, 816–822.

[52] Ruberg, S. J. (1995). Dose response studies I: some design considerations. *Journal of Biopharmaceutical Statisties 5*, 1–14.

[53] Ruberg, S. J. (1995). Dose response studies II: analysis and interpretation. *Journal of Biopharmaceutical Statisties 5*, 15–42.

[54] Russek-Cohen, E., and Simon, R. M. (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine 16*, 455–464.

[55] Sampson, A. R., and Sill, M. W. (2005). Drop-the-losers design: Normal case. *Biometrical Journal 47*, 257–268.

[56] Sequist, L. V., Martins, R. G., Spigel, D., Grunberg, S. M., Spira, A., Jänne, P. A., Joshi, V. A., McCollum, D., Evans, T. L., Muzikansky, A., et al. (2008). First-line gefitinib in patients with advanced non–small-cell lung cancer harboring somatic egfr mutations. *Journal of Clinical Oncology 26*, 2442–2449.

[57] Simon, R., and Wang, S. (2006). Use of genomic signatures in therapeutics development in oncology and other diseases. *The Pharmacogenomics Journal 6*, 166–173.

[58] Song, Y., and Chi, G. Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine 26*, 3535–3549.

[59] Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials 7*, 8–17.

[60] Tang, D.-I., Gnecco, C., and Geller, N. L. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika 76*, 577–583.

[61] Temple, R. J. (2005). Enrichment designs: efficiency in development of cancer treatments. *Journal of Clinical Oncology 23*, 4838–4839.

[62] Thall, P. F., and Cook, J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. *Biometrics 60*, 684–693.

[63] Tibshirani, R., and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics 14*, 511–528.

[64] Ting, N. (2006). *Dose Finding in Drug Development.* Springer.

[65] Vastag, B. (2006). New clinical trials policy at FDA. *Nature Biotechnology 24*, 1043–1043.

[66] Volk, D. W., Matsubara, T., Li, S., Sengupta, E. J., Georgiev, D., Minabe, Y., Sampson, A., Hashimoto, T., and Lewis, D. A. (2012). Deficits in transcriptional regulators of cortical parvalbumin neurons in schizophrenia. *American Journal of Psychiatry 169*, 1082–1091.

[67] Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine: reporting of subgroup analyses in clinical trials. *New England Journal of Medicine 357*, 2189–2194.

[68] Wang, S., James Hung, H., and O'Neill, R. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal 51*, 358–374.

[69] Wang, S.-J., O'Neill, R. T., and Hung, H. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics 6*, 227–244.

[70] West, G. (2005). Better approximations to cumulative normal functions. *Wilmott Magazine 9*, 70–76.

[71] Zhang, W., Sargent, D. J., and Mandrekar, S. (2005). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine 25*, 2365–2383.

[72] Zhao, Y. D., Dmitrienko, A., and Tamura, R. (2010). Design and analysis considerations in clinical trials with a sensitive subpopulation. *Statistics in Biopharmaceutical Research 2*, 72–83.