# Searching for Predictive Subgroups

David K. Prince

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Susanne May, Chair

Scott Emerson

Michael LeBlanc

Program Authorized to Offer Degree:
Public Health - Biostatistics

University of Washington

**Abstract**

Searching for Predictive Subgroups

David K. Prince

Chair of the Supervisory Committee:
Associate Professor Susanne May
Biostatistics

Our increased understanding of genomics and related fields has led to: 1) the identification of subtypes of cancer based on various biomarkers and 2) the development of drugs to target specific subtypes. We then use clinical trials to test whether these novel treatments are effective for a population in question. However, deciding which patients to enroll in a confirmatory trial is typically based on limited empirical evidence. In this dissertation we propose two methods using phase 2 trial results to identify the group that benefits from a treatment, where this group could be all, some, or none of the enrolled population. This contrasts with the traditional approach where the primary analysis of a trial focuses on deciding between all or none. The primary novel aspect of our methods is finer control in the definition of a predictive subgroup. Our first method is an application of logic regression, a tree-based classification algorithm. The method identifies a subgroup of patients that have differential treatment benefit by finding a Boolean statement of binary baseline covariates most strongly associated with an outcome. The second method we call SHAPES and it restricts candidate subgroups to connected and convex or co-convex collections of points in the Boolean space. For both methods we develop methods for continuous and binary outcomes, decision rules and utility functions. We develop and report several metrics to measure performance in this setting including correct group identification rates, the power to detect a subgroup at or above a specified effect, and any rejection rates. In simulation studies, we evaluate our methods using these metrics under various scenarios, including a range of subgroup and full population effect sizes, the prevalence of subgroups, the presence of effects in the absence of treatment (prognostic effects), the number of covariates considered, sample sizes and tuning parameter values. In the presence of large subgroup effect sizes, simple subgroup definitions, few covariates (2-4), moderate sample sizes (200) and correct tuning, our method identifies the true subgroup well. With complex subgroup definitions or incorrect

tuning, performance deteriorates quickly relative to simpler definitions. We apply the methods to a clinical trial in acute myeloid leukemia and explore the relationship between identified subgroup and tuning parameter settings.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

The author thanks his family, friends and the faculty of the Department of Biostatistics at the University of Washington for their support.

# Chapter 1

# INTRODUCTION

This chapter provides an overview of several conceptual and statistical topics relevant to the identification of subgroups with differential treatment effects. We begin by motivating our research with a brief summary of the drug discovery process and clinical trials and then discuss some topics from oncology, including the likelihood of heterogeneous treatment effects and the potential benefits of searching for such effects. We then differentiate between prognostic and predictive effects and discuss methods proposed in the recent literature for either the verification of heterogeneous effects or the identification of predictive subgroup effects. Next we provide an overview of some mathematical and statistical concepts relevant for our research including Boolean expressions, logic regression, the network representation of Boolean space, and various methods for controlling type I error in the context of exploratory analyses.

## *1.1 Clinical Trials and Subgroups*

Before a drug can enter clinical practice it must go through a discovery process to determine whether it is a safe and effective treatment for some group of patients, called the target population. This process is based on a series of clinical trials in humans with increasing size and cost where modifications to the study design occur in between trials or phases. The first step is a phase I trial, which provides initial safety information and evidence for an appropriate dose. Next, phase II trials provide greater safety information and the first evaluation of efficacy. Phase III trials, typically randomized, aim to determine whether a treatment provides some clinically meaningful benefit. This benefit is considered on average for the population enrolled. These trials also provide more information regarding safety. Trials in this phase commonly have a single endpoint for the primary analysis with a well-defined minimum clinically important difference (MCID). Statisticians often use the MCID to calculate a sample size for a fixed type I error rate and power, where power is the probability that if the MCID is true we have a high probability of observing a statistically significant result and the type I error rate or $\alpha$ is the probability of finding a statistically significant result when in truth there is no difference. Throughout the drug discovery process, investigators may stop researching a drug if there is evidence of unacceptable safety risks or lack of a meaningful benefit.

Traditionally, during this process subgroup analyses are only performed at the end of a phase 2 or 3 trial and consider only a small number of variables to define groups. The analyses may or may not be a priori specified. True large treatment effects may not be detected due to low power. Contrastingly, the investigators may not report or be aware of the number of subgroups explored. With more exploration the chance increases that statistical significance is due to multiple comparisons and noise rather than a true subgroup treatment effect. In partial response to this issue of multiplicity, researchers have developed several checklists of criteria to evaluate the evidence for true heterogeneous effects (for an example see [34]). These checklist approaches are applied retrospectively and assume such effects are relatively unexpected, which may no longer hold for new drugs in oncology.

## *1.2 Oncology*

### *1.2.1 Heterogeneity of Cancer and Targeted Therapies*

Cancer is a heterogeneous, multifaceted, and complex disease that researchers have made incredible progress in understanding and treating. Within any cancer patient, there is extensive genotypic and phenotypic heterogeneity both within and among tumors [6]. Researchers use this heterogeneity to develop therapies that target specific subtypes of cancer, where the subtypes are defined by biomarkers. The use of biomarkers to classify subtypes of cancer has already resulted in several important treatment discoveries as researchers have developed drugs that target specific pathways relevant to a specific subtype of cancer. The quintessential example is the drug trastuzumab, which targets human epidermal growth factor receptor 2 (HER2) positive breast cancer and is highly effective for this subtype [2]. Another example is vemurafenib, which is a targeted therapy approved by the FDA for melanoma patients with a certain mutation of the BRAF gene. We might expect this drug to be effective for all cancer patients with this mutation. However, the drug was not effective for BRAF colon cancer patients indicating that selecting which patients benefit is complicated. The difference may be explained by an additional factor, epidermal growth factor receptor or EGFR, which is absent in BRAF melanoma but is commonly present in BRAF colon cancer. Vemurafenib may be effective for colon cancer patients who are BRAF-positive but EGFR-negative [29].

The vemurafenib example suggests that despite progress in measuring, classifying and treating subtypes of cancer, the human body and biological processes are complex and for many cancers there may not be a single biomarker whose value guarantees that a treatment will be effective. Implicit in cancer heterogeneity within an individual and within a tumor is that a single measurement does not capture the true variability of cancer for an individual patient and thus multiple measurements over

| Generic Name | Cancer Type | Subtype | Examples of Additional Subgroups Studied or Suggested Recently with Differential Effects |
|---|---|---|---|
| Cetuximab | Colorectal | KRAS wild type, EGFR+ | Cancer with unresectable liver metastases may respond differently [16] |
| Crizotinib | Non-small cell lung cancer | ALK+ | Different molecular mechanisms associated with ALK may confer resistance to drug [35] |
| Dabrafenib | Melanoma | BRAF V600+ | Greater effect in metastases, particularly brain [17] |
| Lapatinib | Breast | HER2+ | High H2T/p95 protein ratio confers benefit [27] |
| Panitumumab | Colorectal | KRAS wild type EGFR+ | Other members of the HER family and predictive or prognostic factor [20] |
| Rituximab | Lymphoma | B-cell types | Potential toxicities for hepatitis C patients [1] |
| Trametinib | Melanoma | BRAF V600+ | HGF positive may cause resistance to drug [24] |
| Trastuzumab | Breast | HER2+ | By cancer proliferation status [30] |
| Vemurafenib | Melanoma | BRAF V60+ | Different treatment effects based on V600 subtypes [25] |

Table 1.1: Approved therapies within oncology with evidence of two levels of subgroup effects.

time and multiple biomarkers provide important information in determining and selecting the most effective treatment. Additionally, there may be other factors that confer treatment benefit. For many cancers, research has yet to show which biomarkers are the most predictive of a clinically meaningful effect for a given treatment, which in statistical terms is a treatment and subgroup interaction. To also consider other patient characteristics such as age or overall health is more complicated. Even with the clear example of HER2-positive breast cancer, there is evidence of an increased benefit for the sub-subgroup of high cancer proliferation and HER2-positive breast cancer [30]. In treatments shown to be effective for a particular subtype of cancer, there could be differential treatment effects explained by other factors, as shown in the vemurafenib example. As investigators research drugs across multiple types of cancer, understanding the important biomarkers needed for a treatment effect could inform and guide future trials. More examples of drugs with subgroup effects are contained in Table 1.1. The right hand column contains examples of subgroups within the targeted subtype that researchers investigated or proposed investigating recently.

### *1.2.2 Personalized Medicine*

The challenge is to identify as precisely as possible the group that benefits the most from one treatment (or possibly treatment regimen) compared to some other treatment. This is particularly important for cancer treatments, which may have burdensome side effects and where disease progression can occur rapidly. Our increasing ability to describe cancer heterogeneity for an individual both temporally and physically provides a vast source of information to identify such groups. In breast

cancer, perhaps it is only women with any HER2-positive metastases that respond to trastuzumab; or if the patient has any HER2-positive cells; or, perhaps, a treatment is effective only before metastases proliferate. Or the treatment may not be effective. Data is critical to discriminate between these possibilities. The first time data is available to inform these hypotheses is during or after a phase 2 trial.

A doctor treating a patient must select from one or more approved drugs and possibly those without regulatory approval. According to the National Cancer Institute (NCI) there are over 100 drugs approved for the prevention or treatment of cancer, including 67 for the treatment of breast cancer alone. This multitude of options, along with the variability in cancer, motivate the need for personalized medicine defined 'as the tailoring of medical treatment to the individual characteristics, needs and preferences of a patient during all stages of care, including prevention, diagnosis, treatment and follow-up' [10]. Clinical trials can inform personalized medicine. Analyses from trials can aid in the selection of the optimal treatment for a patient by not only providing results by patient characteristics, but also by actively modifying the targeted population of a trial to hone in on those patients who benefit.

### *1.2.3 Characteristics of Clinical Trials in Oncology*

A common characteristic of phase 2 trials in oncology is the use of single arm trials that compare the patients enrolled, all given the same drug, to historical controls. In 2008, El-Maraghi and Eisenhauer found in a review of 89 phase 2 trials for 19 targeted drugs that only 3 or 3.4% used a two arm randomization strategy [9]. Investigators may favor a single arm trial because these designs have a smaller sample size and the investigators may be biased in favor of the efficacy of a new therapy. The use of historical controls, however, may give invalid results. When the historical patients are different from the current patients (referred to as selection bias) or concurrent treatments or patient care changed (an example of confounding) we may find significant results that are not due to the study drug. In response to these issues, some investigators have promoted the wider use of randomized phase 2 trials in oncology to reduce the false-positive rate of trials [33]. In a broader context, between 2003 and 2011 of 1,803 oncology drugs studied, 6.7% passed from phase 1 to regulatory approval [18]. There are likely several reasons for this low rate, including that some drugs only work in a subset of the studied population and the inappropriate use of historical controls [33].

## *1.3 Predictive and Prognostic Effects*

Some patient characteristics are related to outcome in the absence or presence of treatment. An example is cancer stage. These are prognostic effects, which for our purposes we define as factors associated with the outcome in the control arm patients. While these effects are of clinical interest, in terms of making treatment decisions or identifying subgroups they are nuisance parameters. We are interested in identifying predictive effects, or more precisely a predictive subgroup, which we define as a group of patients characterized by baseline or pre-randomization factors that is associated with differences in outcome in the presence of treatment. The importance of taking prognostic effects into account is unclear.

## *1.4 Predictive Subgroup Identification or Confirmation*

As noted previously, one common approach to identify a predictive subgroup is to pre-specify a limited number of analyses. Several other approaches have been proposed recently in the literature and we divide them into confirmatory methods, where typically a single subgroup is explored, and identification methods, where many subgroups are considered.

### *1.4.1 Confirmatory Methods*

There are several approaches to clinical trial designs in the potential presence of subgroups. Freidlin, McShane and Korn [13] summarize trial designs for a phase 2 trial where a single, previously identified biomarker may be predictive. They find that under most scenarios, a design that randomizes all patients but also analyzes results by biomarker status provides the most information to assess the role of the biomarker. More recently, statisticians have been developing methods to control the type I error rate while allowing for changes to trial design based on interim data. In the context of looking for subgroups, these designs are called subgroup enrichment designs. They are intended to be confirmatory and thus only allow for a limited number of pre-specified subgroups, typically one or two, to ensure type I error control. These designs include interim analyses at one or more timepoints during a trial to decide between: 1) continuing to enroll the full population, 2) continuing with one or more of the subgroups, or 3) stopping the trial (for examples, see [5, 19, 23, 26, 31]).

### *1.4.2 Identification Methods*

Enrichment designs are generally intended for confirmatory trials and consider few subgroups, but phase 2 trials could provide earlier information about the presence of heterogeneous effects. At this point randomization, if used, allows for an unconfounded comparison between treatment and control

and the sample size allows for some examination of subgroup effects. Yet there are few methods available for subgroup exploration at this stage. In this section we present summaries from the two methods for predictive subgroup identification and confirmation that have been presented in at least two articles. We provide summaries of the parameters used in simulation studies and the performance summary metrics.

The first method is Subgroup Identification based on Differential Effect Search (SIDES), which develops candidate subgroups using a recursive algorithm [22]. The method constructs a predictive subgroup by considering a single variable at a time, essentially a telescoping procedure where only subgroups of a nested form are considered. A modified version called SIDESscreen allows a greater number of subgroup-defining covariates through an initial screening procedure that selects a subset of variables to be evaluated by SIDES [21]. We refer to both versions generally as SIDES.

SIDES begins with $K$ covariates. A splitting criterion selects candidate subgroups by measuring differences in effects between the two mutually exclusive subsets of patients defined by each covariate. The authors describe four possible criteria for the user to implement; for example, selecting the subset with the greatest effect or smallest p-value. Next a candidate subgroup is accepted if the ratio of the p-value in the candidate subgroup to the p-value in the full population is less than or equal to some $\gamma$, termed the relative improvement parameter where $0 < \gamma \leq 1$. Multiple candidate subgroups can be accepted at this first level. The method then repeats this process for the subsets within each accepted candidate subgroup defined by the remaining $(K - 1)$ covariates. After $L$ repeats, the process stops with a collection of accepted candidate subgroups that are defined using from 1 to $L$ covariates, which in Boolean form are 'and' combinations of the covariates. To confirm that an accepted candidate subgroup is significant overall, a resampling-based method with permuted treatment arm assignments determines critical values that maintain the overall type I error rate.

The parameters and settings used in the simulation studies in the two SIDES articles are summarized in Table 1.2. The number of covariates ranged from 5 to 100 with a sample size of 300 or 900. Subgroup definitions were restricted to 'and' statements of 1, 2 or 3 binary covariates; although SIDES allows for continuous covariates. Subgroup prevalences were 15, 20 or 50%. For the original SIDES manuscript the full population effect was calibrated to be 0 while the effect within the subgroup was calibrated to have 60 or 80% power to detect the effect in the predictive subgroup assuming the correct subgroup is known. The SIDEScreen manuscript assumed 95% power with corresponding effect sizes in the correct predictive subgroup of 0.35 or 0.60 for samples sizes of 900 and 300, respectively.

The authors propose various metrics to describe how SIDES performs in the simulation stud-

| | Subgroup Identification and Differential Effect Search | |
|---|---|---|
| | SIDES | SIDEScreen |
| Sample size | n=900 | n=300, 900 |
| Type I Error Control | Resampling-based multiplicity adjustment | Resampling-based multiplicity adjustment, |
| Endpoints | Continuous | Continuous |
| Covariate Structure | Binary, K=5, 10, 20, Correlation of 0 or 0.3 | Binary, K=20, 60, 100 Correlation of 0 or 0.2 |
| Subgroup Definitions | $X_1$, $X_1$ AND $X_2$ AND $X_3$ | $X_1$ AND $X_2$ |
| Subgroup Prevalences | 15 to 20% | 50% |
| Subgroup Effects | Calibrated so the overall effect is zero and to have 60% or 80% power when the subgroup is known | An effect size of 0.35 or 0.6 in the prognostic subgroup to ensure 95% power to detect an effect in the correct subgroup |
| Baseline Rate or Variance | Variance of $\sigma^2$ | Variance of $\sigma^2$ |
| Prognostic Effects | Not modeled | Not modeled |
| Tuning Parameters/Subgroup complexity control | Continuation criterion<br>L: maximum number of covariates to define a subgroup<br>S: minimum subgroup size<br>M: maximum number of covariates considered | Same parameters as SIDES along with a screening process to select some subset of the variables to use with SIDES |
| Subgroup Evaluation | Splitting criterion - several options | Splitting criterion - several options |
| Decision Rule | Selection criterion | Selection criterion |
| Utility | Unclear | Unclear |
| Primary Performance Evaluation | Selection rate (proportion of simulations with a subgroup selected),<br>confirmation rate (proportion of simulations with a confirmed subgroup),<br>treatment effect fraction (proportion of treatment effect captured in the selected group) | Same as SIDES along with proportion of perfect correct subgroup identification, and proportion of selection of subsets and supersets of the correct subgroup |

Table 1.2: Summary of SIDES and simulation studies performed.

ies. *Selection rate* is the proportion of simulations where at least one subgroup is accepted and *confirmation rate* is the proportion of simulations where at least one subgroup is accepted and confirmed. *Treatment effect fraction* is the treatment effect in the confirmed subgroup divided by the treatment effect in the correct subgroup. Also reported in the second manuscript is the frequency of correct subgroup confirmation and the frequency of the selection of subsets or supersets of the correct subgroup.

Another method for subgroup identification, Adaptive Signature Design (ASD), was developed by Freidlin for the phase 3 clinical trial setting as a two-stage process [14]. The method was later refined to incorporate $k$-fold cross validation, termed cross-validated ASD or CVASD [12]. For the method, first, an overall alpha level ($\alpha$) is set with some amount apportioned to a full population analysis ($\alpha_1$) and the subgroup search ($\alpha_2$). They generally recommend setting $\alpha_1$ to 80% of *alpha* and $\alpha_2$ to the remaining 20%. The full population analysis is conducted using a level $\alpha_1$ test. For the subgroup analysis, in the original ASD the study patients are split into a training and test set. The training set is used to develop a classifier for sensitive patients, i.e. a predictive subgroup, and the significance of the treatment effect in the subgroup is confirmed using the test set. With CVASD the data is split into $k$ sets and for each set a classifier is developed using all the other sets and then the classifier is applied to the set in question to identify classifier-positive patients. The union of all classifier-positive patients from each fold in the last step is the estimated predictive subgroup and a statistic, $T$, is computed that measures the treatment effect in this group. Statistical significance is determined by repeating the entire procedure $B$ times with permuted treatment labels and comparing the permuted test statistics to the original.

The classifying algorithm can be selected by the user. The approach the authors provide is to construct $K$ logistic regression models with three terms: an intercept, a main effect for treatment and an interaction between treatment and a single covariate. If the significance of the interaction is above some threshold $\eta$ the covariate is included in the predictive set of covariates. The patients in the test set are classified as sensitive if $G$ or more of the patient-specific estimated treatment effect are greater than some threshold $R$.

The parameters and settings used in the simulation studies in the two ASD articles are summarized in Table 1.3. In simulation studies, ASD was evaluated with sample sizes between 300 and 800, a binary endpoint, and continuous covariates. Both heterogeneous and homogeneous effects were simulated where the baseline response rate in the control group was 25%. Homogeneous effects were a 10% increase in response and maximum heterogeneous effects in the predictive subgroup ranged from a 35 to a 73% increase in response where subgroup prevalences ranged from 5 to 40%.

| | Adaptive Signature Design (ASD) | |
|---|---|---|
| | ASD | Cross-validated ASD (CVASD) |
| Sample size | n=300, 400, 500, 600, 800 | n=400 |
| Type I Error Control | Test and training sets | Permutation, cross-validation, and test and training sets |
| Endpoints | Binary | Binary |
| Covariate Structure | Continuous, K=10,000 | Continuous, K=10,000 |
| Subgroup Definitions | 3, 10, 20 correlated continuous variables | 10 covariates |
| Subgroup Prevalences | 5%, 7%, 10%, 15%, 20%, 25% | 10%, 20%, 30%, 40% |
| Subgroup Effects | Linear on the log-odds scale; Heterogeneous: 71%, 80%, 87%, 95%, or 98% response in sensitive subjects and baseline for non-sensitive subjects Homogeneous: 35% response rate in all treated subjects | Linear on the log-odds scale; Heterogeneous: 60%, 70%, 90% response in sensitive subjects and baseline for non-sensitive subjects Homogeneous: 35% response rate in all treated subjects |
| Baseline Rate or Variance | Rate of 25% | Rate of 25% |
| Prognostic Effects | Not modeled | Not modeled |
| Tuning Parameters/Subgroup complexity control | $\eta$: significance for covariate treatment interaction R: threshold to classify a subject as sensitive G: number of genes that must meet threshold for a sensitive subject | Same as ASD |
| Subgroup Evaluation | Univariate logistic regression models with the interaction of treatment and a covariate | Same as ASD |
| Decision Rule | Alpha split between full population and subgroup of 0.04 and 0.01 | Alpha split between full population and subgroup of 0.04 and 0.01 |
| Utility | Unclear | Unclear |
| Primary Performance Evaluation | Rejection for full population, rejection for any subgroup, overall rejection | Rejection for full population, rejection for any subgroup, overall rejection |

Table 1.3: Summary of ASD and simulation studies performed.

Effects were modeled as additive on a logit scale where coefficients for predictive effects are equal and covariates have a multivariate normal distribution where the mean is 0 except for the 3, 10 or 20 sensitivity genes that relate to a predictive subgroup. Alpha was set to 0.05 overall with $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$. To measure performance the authors calculated rates of rejection for the full population, any subgroup and overall rejection rates.

Overall, the set of parameters investigated for SIDES and ASD is relatively small, which is a barrier to the adoption of these methods by investigators. To our knowledge, these methods have never been prospectively applied to a clinical trial. There are several areas where additional research could help an investigator decide whether to use these methods, including:

1. With SIDES and ASD, the simulations cover few scenarios. An investigator planning a clinical trial will not know the power to detect a subgroup with a particular set of characteristics not included in the simulations.

2. The range of subgroup prevalences is restricted. However, uncertainty about the definition of a subgroup implies uncertainty about the prevalence of a subgroup and results for a broader range of subgroup prevalences is important to understand.

3. SIDES avoids selecting the full population as an option. ASD tests for the full population but there is no formal rule about selecting between the subgroup and the full population when both are significant.

4. Only a few effect sizes are investigated for both methods, with a focus on heterogeneous effects and no or few simulations under homogeneous effects.

5. While the authors provide some suggestions about the selection of tuning parameters to optimize performance of the methods, they do not explore how tuning parameters should be modified to accommodate different prior beliefs of an investigator.

6. The performance metrics are limited and could be further developed to incorporate other considerations.

7. With SIDES subgroup definitions are restricted to logical expressions of the 'and' form. With ASD subgroup definitions are not explicitly determined. An investigator may be interested in other definitions, for example expressions of the 'or' form.

8. Covariates are always assumed to be predictive in the simulations for both methods. However, covariates could be prognostic or predictive and it is unclear how this changes performance.

9. Each method has a set of tuning parameters. As aspects of the simulation study change, for example the sample size or the true subgroup definition, it is not clear whether and how tuning parameters should be changed to optimize performance. This also makes comparing performance of SHAPES or ASD to any other approach difficult because one must first determine an appropriate collection of tuning parameters to use.

10. Neither method has been investigated in sample sizes less than 300.

11. SIDES is only available as a program written in Microsoft Excel that can analyze a single trial, which prevents its use in simulation studies. No software is available for ASD.

These areas suggest opportunities for research in methods for subgroup identification and confirmation.

### *1.5 Target Population*

Implicit in our discussion of methods for identifying or confirming subgroups is that a predictive subgroup must be a subset of some target population. In other words, the inclusion and exclusion criteria of a trial restrict the population to which one can generalize. Investigators select these criteria for many reasons, possibly including the predictive or prognostic role of an inclusion criterion or the lack of such a role for an exclusion criterion. The criteria may be selected to decrease variability in measurements (e.g. only include patients that have stable baseline values), to include patients more likely to experience an event, or to include patients more likely to benefit from treatment (the group we seek to identify) [11]. To be willing to enroll from a broader population while also examining for subgroups, an investigator must believe that either of the following scenarios are plausible: 1) the treatment benefits the full population and 2) there are predictive subgroup(s) and the trial has collected the variables to identify this subgroup. We point out that without additional information or assumptions, subgroups can only be described relative to the population enrolled in the trial.

### *1.6 The Drug Discovery Process and Subgroup Selection*

As described in Section 1.1, when a drug is evaluated for efficacy and safety it is tested in a sequential series of trials, with different aims. The subgroup identification and confirmatory methods that exist isolate the question of subgroup selection to a single trial in a particular phase, generally phase 3. Yet, prior to the phase 3 trial, a phase 2 trial is conducted with specific operating characteristics and a target population. To fully understand the performance of subgroup selection methods we

must also consider the operating characteristics of each phase and overall for the process and how these change depending on the time at which such methods are applied.

## *1.7 Boolean Algebra*

To define a subgroup using covariates we use logic or Boolean notation as it allows for complex interactions between patient characteristics. This section provides an overview of notation, concepts and terminology.

### *1.7.1 Boolean Expressions*

Boolean expressions use the operators of conjunction or 'and', written as '$\wedge$' ; disjunction or 'or' written as '$\vee$'; and negation or complementation, written as a superscripted '$C$'. As an example of an expression, consider two binary variables, $X_1$ and $X_2$. The Boolean expression '$X_1 \vee X_2^C$' evaluates to 0 or *false* when $X_1 = 0$ and $X_2 = 1$ and otherwise is 1 or *true*.

From oncology, we can consider three conditions: the presence of HER2-positive primary cancer, the presence of HER2-positive metastases, and the presence of any metastases. These conditions can be represented by three random variables, $X_1$, $X_2$, and $X_3$, that take on value 1 if the condition is satisfied and 0 if not. We can represent women with HER2-positive primary breast cancer without any metastases by $X_1 \wedge X_3^C$ and women with any HER2-positive cancer by $X_1 \vee X_2$.

An additional example comes from an FDA press release announcing the approval of ibrance dated Feburary 3, 2015:

> Ibrance is intended for postmenopausal women with estrogen receptor (ER)-positive, human epidermal growth factor receptor 2 (HER2)-negative, metastatic breast cancer who have not yet received an endocrine-based therapy.

Several patient characteristics define the indication. Considering true or false versions of each characteristic, we set $X_1$ equal to postmenopausal, $X_2$ to female, $X_3$ to ER-positive, $X_4$ to HER2-positive, $X_5$ to metastases present, and $X_6$ to previously treated with endocrine-based therapy. The indication as a Boolean expression is then $X_1 \wedge X_2 \wedge X_3 \wedge X_4^C \wedge X_5 \wedge X_6^C$, which could represent a collection of both predictive and prognostic effects. Here metastases status could be prognostic because metastases present is associated with a worse outcome in the with or without treatment whereas HER2 status could be predictive if only the HER2-positive patients have treatment benefit.

### *1.7.2 Boolean Space and Boolean Functions*

Expanding on this idea of Boolean expressions, we turn to Boolean functions. First consider $K$ observed baseline binary covariates $X_i$ for $i \in \{1, ..., K\}$, where $X_i = 0$ or $X_i = 1$. We refer to all possible combinations of these binary variables as either the Hamming cube or Boolean space, represented as $\{0,1\}^K$. There are $2^K$ unique points in this space, where a point is a vector of length $K$ with each element equal to 0 or 1. To define a subgroup, we need a function, $f$, that is a mapping from this space to a binary indicator for predictive subgroup membership:

$$f : \{0,1\}^K \rightarrow \{0,1\}.$$

This $f$, or Boolean function, can be represented as a Boolean expression, like the examples in the preceding section. We define a subgroup as the points in the Boolean space that map to 1 or equivalently as the collection of points that evaluate to true.

Different Boolean expressions can represent the same $f$ [28]. To evaluate the equality of two expressions we construct what is named a truth table that contains a list of each unique point in the space (a total of $2^K$) and the output values. When the output values for each expression are the same for every input point the expressions are equivalent. When $K$ is small, we can list all possible mappings. As $K$ increases the number of possible mappings increases rapidly. When $K = 3$, there are 8 points and 256 unique combinations of points and when $K = 8$ this increases to essentially infinity.

## *1.8 Logic Regression*

In determining a Boolean expression to define a predictive subgroup we want to incorporate observed patient outcomes and randomized treatment arm along with baseline covariates. One method that we can modify to incorporate this information is logic regression, which is a stochastic, tree-based algorithm that identifies a Boolean function of the covariates that optimizes some scoring function [32]. It can be used with binary, continuous, and time-to-event outcomes. For continuous outcomes, one can use the sum of the squared residuals as the scoring function and for binary outcomes the deviance. The generalized linear form of logic regression with link function, $g()$, setting $X = X_1, ..., X_K$ and adjusting for covariates or prognostic factors, $m(\mathbf{x})$, is:

$$g(E[Y|X = \mathbf{x}]) = \alpha + \beta f(\mathbf{x}) + \gamma m(\mathbf{x}).$$

Logic regression selects a best fitting $f(\mathbf{x})$ using a simulated annealing or a greedy search algorithm with two tuning parameters that modify performance: *number of leaves* and *temperature*. The former defines the maximum number of variables included in the Boolean expression corresponding to $f(\mathbf{x})$ but this provides limited control over the complexity of the expression as variables can be repeated and expressions do not uniquely define a subgroup. Given a large enough *number of leaves* or $n_{leaves}$ all possible Boolean expressions are considered by logic regression. *Temperature* controls the likelihood of a new move being accepted as the logic tree is created and modified over time.

### *1.9 Connectivity and Graph Theory*

While logic regression considers all possible Boolean functions with large enough $n_{leaves}$, all of these may not be of interest. We instead could place restrictions directly on the set of functions from which we select. One such restriction relates to connectivity. First consider the Hamming distance between two points, $a, b \in \{0,1\}^K$:

$$\delta(a,b) = \#\{i : a_i \neq b_i\}$$

This is the number of entries in the vector that are not equal. If $\delta(a,b) = 0$, then the points are equal. If $\delta(a,b) = 1$, then the points differ by one entry. For $\delta(a,b) = c$, where $c > 1$, $c$ provides a measure of how different the points are. Generally, this distance measures how different two patients are with respect to the baseline covariates.

We can visualize these distances by using a *graph* or network model where the *nodes* represent each point and lines or *edges* connect the nodes with a Hamming distance of one. Figure 1.1 shows an example of a graph of the full target population when $K = 3$. A predictive subgroup is some subset of the nodes of this graph along with the corresponding edges, called a *subgraph*. This subgraph is defined by the points that are mapped to 1 or *true* by a Boolean expression.

The graph representation of the Boolean space preserves the relationships among the points, which is an important feature to highlight if we expect to make similar treatment decisions for patients that do not differ much in terms of baseline covariates. Relatedly, is the concept of connectedness, which with regards to a subgraph means that any two points in the subgraph can be reached via a path, where a path is defined as a sequence of points of Hamming distance 1 such that each of those points are also in the subgraph. We could restrict our candidate predictive subgroups to only consist of *connected* subgraphs. This idea of connectedness can also describe the set of points not in the subgraph. If this set is connected, we term the original subgraph co-connected and if the subgraph is connected and co-connected, it is *strongly connected* [8].

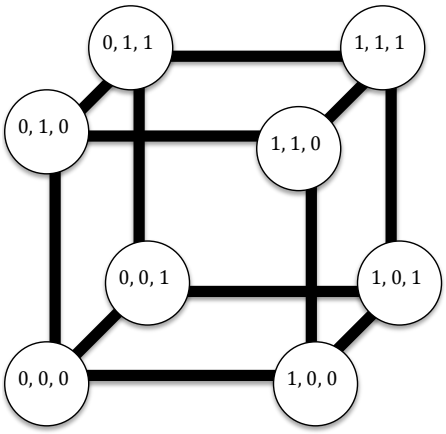



Figure 1.1: Graph for $K = 3$.

A stronger condition than connectivity is convexity. A subgraph is *convexly connected* if and only if all the shortest paths connecting any two points are included in the subgraph. Alternatively, a subgraph is *co-convexly connected* if and only if the set of points not in the subgraph is convexly connected. We make use of these concepts in the next chapter.

### *1.10 Summary and Concluding Remarks*

The development of flexible methods for subgroup identification in the area of oncology is valuable given the heterogeneity of cancer and the development of targeted therapies. No current subgroup identification methods model interactions in covariates directly and most consider a small range of sample sizes, effect sizes, and/or number of covariates. In this dissertation we will develop and evaluate novel methods for the identification and confirmation of subgroups with enhanced treatment benefit in the context of the drug discovery process. Specifically, we have the following three aims:

1. We will assess for heterogeneous treatment effect in the phase 2 oncology clinical trial setting using logic regression. In a variety of scenarios we will evaluate the performance of this approach with various rules to decide between selecting the full population, the identified subgroup, or discontinuing the study.

2. We will develop and evaluate the performance of a new method, called SHAPES, that restricts the possible covariate combinations used to define a subgroup to a connected set or convexly connected set in Boolean space.

3. We will assess the performance of SHAPES in the context of a collection of drugs to be studied where some drugs have a heterogeneous effect, some have a homogeneous effect, and some have no effect.

The basic goal of this research is to provide knowledge about when a search for subgroups provides benefits, what these benefits are and conversely when it does not, and what disadvantages may result.

# Chapter 2

# TWO METHODS FOR SUBGROUP SELECTION: LOGIC REGRESSION AND SHAPES

In this chapter, we begin by specifying the notation for the problem of subgroup selection and then provide a description of two different methods for subgroup selection. The first method is an application of an existing regression technique, logic regression, described in the previous chapter. The second method, called SHAPES, builds on the concepts of connectivity and convexity to restrict how subgroups are defined.

## *2.1 Problem Outline*

### *2.1.1 Data*

The general setting is a randomized, controlled clinical trial where for subjects $i \in \{1, ..., n\}$ we observe the baseline covariates $X_k = x_{ki}$ where $x_{ki} \in \{0, 1\}$, $k \in \{1, ..., K\}$ and $X = \{X_1, ..., X_K\}$. These covariates represent already binary or dichotomized continuous, nominal or ordinal variables. We denote the observed vector of covariates for the $i^{th}$ patient as $\mathbf{x}_i$. We randomly assign each subject to a condition, $T = t_i$, where $T = 1$ indicates active treatment and $T = 0$ indicates control or standard of care, and then observe a continuous or binary outcome, $Y = y_i$. Generally we assume a 1:1 randomization where $Pr(T = 1) = Pr(T = 0) = \frac{1}{2}$, but this is not mandatory.

The $X$ divide our Boolean predictor space into $2^K$ disjoint *subcubes* or *points*. Consider a Boolean function, $D$, that maps from the Boolean space of $\{0, 1\}^K$ into the decision we make under the truth to enroll ($D = 1$) or discontinue study ($D = 0$) in a phase 3 trial for each point, that is:

$$D : \{0, 1\}^K \rightarrow \{0, 1\}.$$

The ideal or true subgroup is then $S = \{\mathbf{x}_i : D(\mathbf{x}_i) = 1\}$. For our purposes, $D$ should map to 1 for the points where the contrast between treatment and control is at or above the MCID and to 0 otherwise. If $D$ always equals 1, we should choose to enroll the same population in the next trial or if $D$ always equals 0, we should choose to discontinue the study. Other combinations of zeros and ones correspond to some proper subgroup.

### *2.1.2 Primary Objective*

Our target is to determine a reasonable estimate of $D$ and $S$ based on the observed data. We can define an estimate of D as:

$$\hat{D}|data : \{0,1\}^K \rightarrow \{0,1\}.$$

The selected subgroup is then $\hat{S} = \{\mathbf{x}_i : \hat{D}(\mathbf{x}_i) = 1\}$. Intuitively, $\hat{D}$ should map to 1 in the points where the contrast between treatment and control looks *interesting* at phase 2 by some objective measure and to 0 where it looks *uninteresting*.

### *2.1.3 Prospective or Retrospective*

We design our methodology with the assumption that it will be prospectively used, in that the search for subgroups is pre-specified in the analysis plan. In practice, a likely application is to retrospectively use this methodology when a trial has not found a statistically significant result and subgroup exploration was not considered until after this result.

## *2.2 Logic Regression*

Our first method uses logic regression to estimate $D$ as a Boolean expression of covariates that minimizes some scoring function. Two reasonable scoring functions are the sums of squared residuals (RSS) with continuous outcomes or the deviance with binary outcomes. The generalized linear form of logic regression is:

$$g(E[Y|X = \mathbf{x}]) = \alpha + \beta d(\mathbf{x}) + \gamma m(\mathbf{x}) \text{ with } d : \{0,1\}^K \rightarrow \{0,1\}.$$

The link function, $g()$, can be the identity link for continuous outcomes or the logit link for binary outcomes. This model also allows adjustment for covariates or prognostic factors via the $m(\mathbf{x})$ term. The Boolean function, $d(\mathbf{x})$, describes the logical combination of $X$ most associated with the outcome as judged by the scoring function. However, as this form of logic regression does not explicitly incorporate a term for treatment we propose a multipstep process to incorporate treatment status. The general process is:

1. Fit a model on the control group data to identify prognostic effects (*the prognostic model*)
2. Select a subgroup by fitting logic regression on the treatment group data accounting for the prognostic effects from step 1 (*the predictive model adjusted for the prognostic model*).
3. Conduct hypothesis testing to select between the full target population, the subgroup identified in step 2 or to discontinue study in any group (*testing*).

In the following subsections, we detail each step including three versions of step 2. Initially we explored an approach that fit a predictive and prognostic expression concurrently in a single step but due to poor performance we omitted this approach.

In the remainder of this dissertation we refer to capitalized Logic Regression or LR to refer to the application of this method to our problem and lowercase logic regression to refer to the method in general.

### *2.2.1 The Prognostic Model*

When using only treatment arm data, we anticipate that a predictive subgroup cannot be accurately identified if some of the $K$ covariates also have prognostic effects. Any detected variation in participant outcome might be due to these effects and not to a heterogeneous treatment response and therefore to accurately identify a predictive subgroup we must incorporate information about these effects into our models. The modeling of prognostic effects is not done with SIDES or ASD. With control group data, we model prognostic effects using some form of regression, such as linear, logic or penalized regressions. From these models we can either: a) identify a subgroup or subgroups with varying outcome or b) estimate prognostic effects. Specifically, a general form of a prognostic model where we regress $Y$ on $X$ for subjects where $T = 0$ is:

$$g(E[Y|X = \mathbf{x}, T = 0]) = \gamma_0 + \sum_{g=1}^{G} \gamma_g p_g(\mathbf{x}).$$

The $\{\gamma_g\}$ are the estimated prognostic effects. Using logic regression each $p_g$, for $G \geq 1$, is a Boolean expression that identifies a prognostic subgroup. Using the $\ell_1$ penalized approach, or LASSO, $G$ is the number of covariates with non-zero coefficient estimates as LASSO will shrink many of these estimates to zero depending on the selected tuning parameter.

### *2.2.2 Predictive Model Adjusted for Prognostic Model, First Version (LR-Adjusted)*

In the first version of step 2 we use the prognostic subgroup(s) identified in step 1 and determine which participants in the treatment group are part of the subgroup, or for patients with T=1, we calculate $p_g$ for $g \in \{1, ..., G\}$. Then, using only the treatment group data, we fit the following logic regression model controlling for the identified prognostic subgroup to identify a predictive subgroup $\hat{D} = d(\mathbf{x})$:

$$g(E[Y|X = \mathbf{x}, T = 1]) = \beta_0 + \beta_1 d(\mathbf{x}) + \sum_{g=1}^{G} \gamma_g^* p_g(\mathbf{x}).$$

Should $\hat{\beta}_1 < 0$ we select $(1 - d(x))$ as the subgroup. We apply this tranformation for the identified predictive subgroup as well. Note the estimates of $\{\gamma_g^*\}$ are not the same as the estimates of $\{\gamma_g\}$ from step 1 but the subgroup definitions $\{p_g\}$ are the same. This method requires no distinction between continuous or binary outcomes other than the use of the corresponding link function. We term this approach *Logic Regression-Adjusted* or *LR-A*.

### *2.2.3 Predictive Model Adjusted for Prognostic Model, Second Version (LR-Residuals)*

For the second version of step 2, we model prognostic effects for the control group and then calculate residuals for the treatment group based on the step 1 model. Specifically, for patients with T=1, we calculate residuals based on the control data: $y_i^* = y_i - \hat{y}_i$; where $\hat{y}_i$ is the fitted value based on using the estimates of $\{\gamma_g\}$ and $\{p_g\}$ from the model in step 1. For continuous outcomes, we fit logic regression on the $\{y_i^*\}$ to identify $\hat{D} = d(\mathbf{x})$:

$$g(E[Y^*|X = \mathbf{x}, T = 1]) = \beta_0 + \beta_1 d(\mathbf{x}).$$

For binary outcomes, $\{y_i^*\}$ is no longer binary and we model this outcome as continuous using the identity link function and the residual sum of squares as the scoring function. For this second version, observe that both the identification and magnitude of prognostic effects are based on the control group data. We term this approach *Logic Regression-Residuals* or *LR-R*.

### *2.2.4 Predictive Model (not) Adjusted for Prognostic Model, Third Version (LR-Unadjusted)*

For the third version of step 2, we assume no prognostic effects and, in effect, skip step 1. We fit the model:

$$g(E[Y|X = \mathbf{x}, T = 1]) = \beta_0 + \beta_1 d(\mathbf{x}).$$

We expect this model to not be able to discern prognostic and predictive effects but may function well in the absence of prognostic effects. We term this approach *Logic Regression-Unadjusted* or *LR-U*.

### *2.2.5 Testing*

From step 2 we have an identified subgroup, $\hat{S} = \{\mathbf{x} : d(\mathbf{x}) = 1\}$, that performs best according to our criterion, but we have not formally tested the significance of the effect in the subgroup using the combined treatment and control data or tested the effect in the full population. In this section we focus on setting up the LR models for hypothesis testing. To test, we consider either interactions

or stratification, by which we mean testing within a candidate predictive subgroup. The approach of allowing for multiple methods of testing is similar to both SIDES and ASD where with either method a user could select a test statistic of interest. To test for an interaction using both treatment and control data with the first version we use the generalized linear regression model with $d(\mathbf{x})$ and $p_g(\mathbf{x})$ as identified in the previous steps:

$$g(E[Y|X = \mathbf{x}, T = t]) = \beta_0' + \beta_1' t + \beta_2'(t * d(\mathbf{x})) + \sum_{g=1}^{G} \gamma_g' p_g(\mathbf{x}). \tag{2.1}$$

We are interested in the significance of $\beta_2'$. However, a positive estimate of $\beta_2'$ does not prove an overall positive treatment effect in the subgroup as the value of $\beta_1'$ must be considered. Hence we are also interested in whether $\beta_1' + \beta_2'$ is greater than some minimum clinically important threshold. For the second version of logic regression we fit a similar model except without prognostic terms in the model since the residuals already account for this:

$$g(E[Y^*|X = \mathbf{x}, T = t]) = \beta_0' + \beta_1' t + \beta_2'(t * d(\mathbf{x})). \tag{2.2}$$

We again are interested in the significance of $\beta_2'$ and whether $\beta_1' + \beta_2'$ is greater than some threshold.

We also implement a stratified approach where we compare the participants identified as the predictive subgroup in the treatment group to the same participants in the control group, i.e. patients where $\hat{D} = 1$ with this function determined in the previous steps. In the first version of logic regression, we fit the model:

$$g(E[Y|\hat{D} = 1, T = t]) = \beta_0^* + \beta_1^* t. \tag{2.3}$$

We are interested in the significance of $\beta_1^*$, as this represents the treatment effect within the subgroup. In the second version of logic regression, we model the residuals:

$$g(E[Y^*|\hat{D} = 1, T = t]) = \beta_0^* + \beta_1^* t. \tag{2.4}$$

We again are interested in inference about $\beta_1^*$. For binary outcomes with this model we use the identity link, i.e. use a continuous approach.

### *2.2.6 Simulated Annealing and the Tuning Parameters*

In fitting logic regression, we have avoided discussing the tuning parameters of number of leaves or $n_{leaves}$ and *temperature* for these models [32]. The first parameter controls the number of variable forms allowed in the Boolean expression that logic regression finds (Note: a variable can take on two forms, for example $X_1$ and $X_1^C$). With SIDES the related tuning parameter is $L$. SIDES only considers expressions with the 'and' operator but LR allows all possible Boolean expressions with up to $n_{leaves}$ covariates. The $n_{leaves}$ parameter limits the number of subgroups we consider; allowing four leaves versus two leaves greatly increases the number of models fit and will inflate type I error if uncorrected. However, if the truth is that a subgroup is defined by a minimum of four leaves and we only allow two leaves we never recover the truth. Additionally, we have limited control over the complexity of the developed expression as variables can be repeated in a expression and expressions are not unique. As an example, some subgroups that can be represented with two leaves can also be represented with four leaves.

The other tuning parameter is *temperature*, which controls the likelihood of a new move being accepted as a logic expression is developed. During simulated annealing a series of random changes are made to the Boolean expression. Changes that improve the model with respect to the scoring function are accepted. Changes that do not improve the model are accepted with some probability that depends on how different the previous and current models are and the current *temperature*. Higher values make it more likely new moves are accepted and lower values make it less likely while the algorithm systematically decreases the *temperature* over time. Provided the start value is reasonably high at the start of the search and cooling occurs gradually this parameter has little impact on the selected subgroup [32]. The use of a stochastic search process to the subgroup search process is novel.

## *2.3 SHAPES*

Logic regression indirectly restricts the number and structure of subgroups through the tuning parameter of $n_{leaves}$. Alternatively, we could directly restrict the types of subgroups we will accept, specifically the types of Boolean expressions that define a subgroup. This benefits the selection process by decreasing the number of possible combinations of points in the Boolean space, which lessens the multiplicity problem, prevents overfitting of the data, and may be more scientifically appropriate if an investigator has an a apriori belief about which subgroups are possibly predictive. With the SHAPES approach we consider two restrictions: the selected subgroup must be connected in Boolean space and the selected subgroup must be either convex or co-convex, which is to say

the complement of the selected subgroup is convex. With this method we refer to an acceptable collection of points as a *shape*.

To illustrate some possible shapes, define $X_1$ to be male, $X_2$ to be old, and $X_3$ to be high body mass index (BMI) with the complements of these variables respectively representing female, young and low BMI. Examples of acceptable shapes include $X_1$ (males), $X_1^C \wedge X_2$ (old females), or $X_1^C \vee X_2^C \vee X_3$ (everyone but old males with low BMI). Examples of unacceptable shapes are $(X_1 \wedge X_2) \vee (X_1^C \wedge X_2^C)$ (old males or young females) and $X_1 \wedge (X_1 \wedge X_2 \wedge X_3)^C$ (males except old males with high BMI. The former is neither convex nor connected and the latter is not convex or co-convex.

Restricting candidate subgroups to be convex or co-convex and connected results in a collection of possible subgroups that is a superset of the SIDES collection. Essentially, we include all subgroups possible with SIDES but also consider the complement of each possible subgroup: Boolean expressions which only contain the 'or' operator. Based on our development of SHAPES, an investigator could also customize which subgroup definitions are and are not acceptable.

### *2.3.1 Acceptable Shapes and the Tuning Parameter*

In addition to the number of covariates under consideration, $K$, we define a tuning parameter $\ell$ as the maximum number of covariates selected for a subgroup definition. This parameter increases or decreases the complexity of the subgroup definition. Generally, we set $\ell \leq K$ and $\ell \leq 3$. With $\ell = 3$, Table 2.1 lists the various subgraphs of points in Boolean space that are connected and convex or co-convex, including the full population and the empty set, for $Z_i \in \{X_1, ..., X_K\} \cup \{X_1^C, ..., X_K^C\}$, and not allowing for duplicate variables in the definition. For each of these possible shapes as we increase $K$ but hold $\ell$ fixed there are more possible shapes. Figure 2.1 shows graph or network model representations of these subgroups. A black node indicates that the point is selected in the subgroup and a gray node indicates it is not. The possible models for when $\ell = 2$ are similar, except models (b) and (f) are excluded and, when $\ell = 1$, only (a), (d) and (g) are considered. We also classify the shapes by the *depth* or the number of covariates in the definition. We may prefer shapes that involve fewer variables or synonymously are of smaller depth.

With $\ell \leq 3$, a stochastic search for the best fitting subgroup is not necessary as all possible models can be readily calculated, an approach we have taken. For larger $\ell$, a stochastic search procedure could be developed that moves in a backward approach by removing points from the Boolean space or in a forward approach by starting with the empty set and adding points that are connected.

| Model | Description | Boolean form | Depth | Connected | | Convex | |
|---|---|---|---|---|---|---|---|
| | | | | Shape | Shape$^C$ | Shape | Shape$^C$ |
| (a) | empty set | $\emptyset$ | 0 | *no* | *yes* | *no* | *yes* |
| (b) | single node | $Z_a \wedge Z_b \wedge Z_c$ | 3 | yes | yes | yes | no |
| (c) | two connected nodes | $Z_a \wedge Z_b$ | 2 | yes | yes | yes | no |
| (d) | square | $Z_a$ | 1 | yes | yes | yes | yes |
| (e) | complement of (c) | $Z_a \vee Z_b$ | 2 | yes | yes | no | yes |
| (f) | complement of (b) | $Z_a \vee Z_b \vee Z_c$ | 3 | yes | yes | no | yes |
| (g) | full population | $A = \emptyset^C$ | 0 | yes | *no* | yes | *no* |

Table 2.1: Acceptable shapes when $\ell = 3$. Note that $a \neq b \neq c$ and for $h \in \{a, b, c\}$ $Z_h \in \{X_h, X_h^C\}$. All shapes are connected.

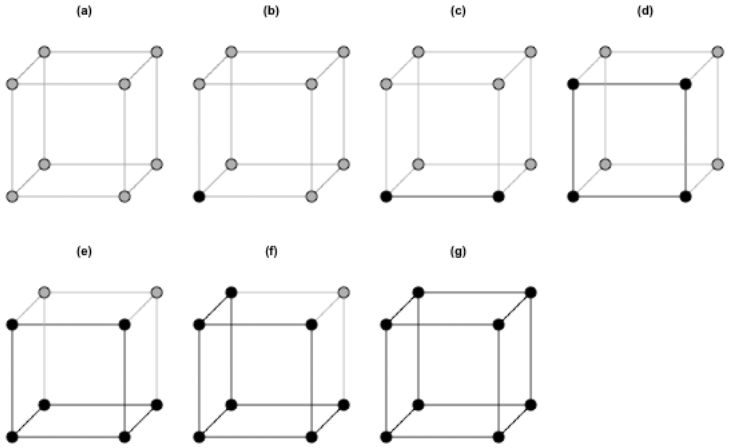



Figure 2.1: For $K = 3$ and $\ell = 3$, the possible shapes if we require convexly connected or co-convexly connected subgroup. Each circle or node represents a unique combination of covariates. Lines or edges connect those nodes that differ by a single covariate value. A black circle indicates a node that is selected as interesting and a gray circle indicates a node that is not interesting.

| $K$ | Points | All Possible | Restricted Shapes | | |
|---|---|---|---|---|---|
| | | | $\ell = 1$ | $\ell = 2$ | $\ell = 3$ |
| 1 | 2 | 4 | 4 | 4 | 4 |
| 2 | 4 | 16 | 6 | 14 | 14 |
| 3 | 8 | 256 | 8 | 32 | 54 |
| 4 | 16 | 65,536 | 10 | 58 | 138 |
| 5 | 32 | $4.29 \times 10^9$ | 12 | 92 | 282 |
| 6 | 64 | $1.85 \times 10^{19}$ | 14 | 134 | 502 |
| 7 | 128 | $3.40 \times 10^{38}$ | 16 | 184 | 814 |
| 8 | 256 | $1.16 \times 10^{77}$ | 18 | 242 | 1,234 |
| 9 | 512 | $\to \infty$ | 20 | 308 | 1,778 |
| 10 | 1024 | $\to \infty$ | 22 | 382 | 2,462 |

Table 2.2: Count of possible subgroups for values of $K$ as acceptable subgroups are restricted.

### *2.3.2 Quantifying SHAPES Restrictions*

The primary purpose of SHAPES is to systematically limit the collection of acceptable subgroups thereby lessening the problem of multiplicity. Table 2.2 contains the number of unique subgroups for various values of $K$. The 'All Possible' column lists the number of subgroups that logic regression can identify with a large value for $n_{leaves}$. This is the number of subgroups if we consider all possible combinations defined by the $K$ covariates. The 'Restricted Shapes' columns display for various values of $\ell$ the number of unique subgroups that are possible. When $K = 4$ there are 65,536 possible unique subgroups but by placing a restriction of $\ell = 3$ limits this to 138 unique subgroups. For $K = 8$, we move from a near infinite number of subgroups to 1,234. The reduction in possible shapes controls for multiplicity and limits inflation of the type I error rate.

### *2.3.3 Testing in SHAPES*

To test the subgroups identified by SHAPES, we can use the same approaches as LR - stratification or interactions. Specifically, we can calculate the models from Equations 2.1 and 2.3 for all candidate subgroups as generated from a value of $\ell$. The challenge becomes accounting for this exploration in our final decision. In the next section we describe possible approaches.

## ***2.4 Accounting for Exploration when Testing and Making a Decision***

Logic Regression and SHAPES identify one and multiple candidate predictive subgroups, respectively. From Equations 2.1 to 2.4, with the generalized linear models framework we can readily test the significance of the parameters associated with a subgroup of interest via Wald tests and linear combinations, where we accept a significant test of a coefficient as sufficient evidence to consider a subgroup interesting. We focus on one-sided tests in favor of the treatment group, but methods could be adapted to two-sided tests. Our approach here is in line with SIDES and ASD in that a user can determine the test of interest. We more explicitly describe our testing procedures and the corresponding null and alternative hypotheses than either of those methods. The primary challenge with testing and these methods is to calibrate the critical values to account for the exploration. In this section we describe possible approaches to determine the critical values and discuss deciding between the full population and the subgroup.

### *2.4.1 Testing in the Full Population*

For the full target population effect, we can test $\phi_1$ using the generalized linear model and a Wald test:

$$g(E[Y|T = t_i]) = \phi_0 + \phi_1 t_i. \tag{2.5}$$

The link function $g()$ is the identity for continuous outcomes, where $\phi_1$ represents the mean difference in outcome comparing treatment to control, and the logit for binary endpoints, where $\phi_1$ corresponds to the log-odds ratio for the same comparison. Under the null hypothesis for the full population, denoted $H_F$, we assume $\phi_1 \leq \delta$, where $\delta$ is some MCID and we denote the p-value from such a test as $p_F$.

### *2.4.2 Testing in the Subgroup*

To test in the subgroup we consider either an interaction or stratification approach, as described in the Logic Regression section. When considering the interaction model we focus on two parameters: $\beta_1$, the main effect for treatment, and $\beta_2$, the interaction coefficient from Equation 2.1. Note that these parameters are calculated differently for our various methods but we use the same framework to test them. We are interested in two hypotheses:

$$\begin{aligned} H_{01} &: \beta_1 + \beta_2 \leq \delta \\ H_{02} &: \beta_2 \leq 0. \end{aligned}$$

The first hypothesis tests whether the effect in the subgroup is less than or equal to some $\delta$. The second hypothesis tests whether the interaction effect is non-zero. We define a composite hypothesis to test for them jointly:

$$H_{03} = H_{01} \cup H_{02}.$$

For the stratified approach, we only consider a model within the subgroup from Equation 2.3 and test the main effect for treatment, $\beta_1^*$ as $H_{04} : \beta_1^* \leq \delta$.

With LR the candidate subgroup is selected prior to this testing stage whereas with SHAPES we consider all candidate subgroups and select the subgroup with the minimum p-value from these testing procedures. In both cases we denote the selected subgroup as $\hat{S}$ to emphasize that it was not pre-specified but was selected from the observed data and hence does not always refer to the exact same subgroup. We refer to the null hypothesis as $H_{\hat{S}}$ regardless of testing approach and the

associated p-value as $p_{\hat{S}}$. The derivation of the test statistics for the continuous case for the full population, interaction and stratified tests are available in Appedix A.

### *2.4.3 Type I Error Control and Decision Rules*

To control type I error we first select an overall or experiment-wise type I error rate of $\alpha$. We then pre-select some portion of this $\alpha$ for a test in the full population, denoted $\alpha_F$ and determine a critical value, $c_F$, on the p-value scale such that:

$$Pr(p_F < c_F | H_F) < \alpha_F.$$

As the Wald test is well-calibrated for reasonable sample sizes we typically select $c_F = \alpha_F$. Then with fixed $c_F$ we use the remaining $\alpha$ to test for a subgroup and determine a $c_S$ that maintains the overall type I error rate:

$$(Pr(p_F < c_F | H_F) \cup Pr(p_{\hat{S}} < c_S | H_{\hat{S}})) < \alpha.$$

These values of $(c_F, c_S)$ form decision rules about when to reject $(H_F, H_{\hat{S}})$. This $\alpha$-allocation can be extended with SHAPES to further partition the remaining $\alpha$ based on the depth of the subgroups by possibly reserving more $\alpha$ for shapes with less depth or sharing the remaining $\alpha$ equally between each depth.

Since both LR and SHAPES use exploratory methods to identify a subgroup, there is no standard way to determine a value of $c_S$ that accounts for the exploration, but four possible approaches are:

- Distributional assumptions: We generate data under some null distributional assumption, for example assuming the effect is null for everyone. We then apply our methods and under replications, we determine a $c_S$ that maintains the selected $\alpha$. For example, if we use a binary outcome with an assumed baseline probability of $Pr(Y = 1|T = 0) = 0.30$ and a covariate distribution with $K = 4$ independent covariates each with prevalence 50%, we generate critical values through simulated replications when assuming no treatment effect. The ability to control type I error with this approach will depend on our assumptions and the precision of $c_S$ based on the number of replications.

- Bonferroni: To perform a Bonferroni correction we set $c_S$ to be the quotient of $(\alpha - \alpha_F)$ divided by the total number of possible subgroup models. Overall, this is a conservative approach and does not use the correlation between the full population and subgroup tests but it does ensure

the type I error is at or below the nominal level. This approach may only be practical for smaller values of $K$, given the increasing number of possible subgroups as $K$ increases.

- Permutation. When using the residuals, we identify unique permutations of each $\mathbf{x_i}$ and $y_i^*$ within the treatment group to generate a null distribution for comparison with the observed data. We then fit the model using our method and compare the observed data with the collection of permuted data to assess the significance of the observed data. The SIDES manuscripts use this approach.

- Test and training set: With a larger sample we can split the data into a training and a test set to first identify a subgroup with the training data and confirm the finding with the test set. As we aim to focus on performance in small samples, this approach is not relevant. This approach was used in the first ASD manuscript.

In later chapters we focus on *distributional assumptions* to determine critical values. We take this approach because our primary focus is binary outcomes, which requires minimal assumptions such as the rate of outcome in the control arm and the distribution of the binary covariates and where violations can be checked by sensitivity analyses.

### *2.4.4 Utility*

The decision rules, derived in part from $\alpha$ and $\alpha_F$, guide when to reject $H_F$ and $H_{\hat{S}}$. We can expect these decisions to be related as any selected subgroup forms part of the full population. As an example, consider if there is a large subgroup effect in a group composed of 75% of the population. We are likely to reject both $H_F$ and $H_S$. Neither SIDES nor ASD consider what final decision to make when multiple groups are selected. SIDES also does not discuss what to do in the situation where multiple subgroups are confirmed. To decide on one group when both tests are significant, we can use a utility function. Three possibilities are:

- Prefer subgroup: Always select the subgroup.

- Minimum of the standardized p-values: select the group that corresponds to the $\min(\frac{p_F}{c_F}, \frac{p_{\hat{S}}}{c_S})$, which is the group that has the minimum p-value divided by the respective critical value.

- Prefer all: Always select the full population.

These utilities span the range of possible tiebreakers between a subgroup and the full population. Each utility may be appropriate for different scenarios. For example, in a drug that has high toxicity

it may be preferable to ensure that we only select patients with strong evidence of benefit. This corresponds to the *prefer subgroup* rule. With a drug that has expected low rates of side effects or which is likely to benefit all patients, a *prefer all* rule might be best. The *minimum p-value* rule may be appropriate for a situation where side effects and benefits are uncertain.

### *2.5 Evaluating Performance*

To determine how LR and SHAPES perform in terms of discriminating between the possibilities of no, subgroup and full effects requires that we develop some metrics that are not needed in a more typical trial paradigm. Traditionally, the focus is to test if a parameter of interest, say $\theta$, is above (or below) some MCID, say $\theta^*$. This $\theta$ captures some contrast in outcomes between the participants assigned to the treatment and the control conditions. In truth there are two possibilities: $\theta > \theta^*$ or $\theta \leq \theta^*$. Investigators plan a trial and analysis to decide which of these hypotheses are most likely, while controlling the likelihood of making one of the two possible incorrect decisions, i.e. holding type I error low and power high (corresponding to a low type II error rate). We refer to this approach as the *full population* analysis.

By testing for subgroup effects, we add the possibility of making an error related to the identified subgroup, for example the identified subgroup could be a mixture of the true subgroup and others where there is no treatment effect. As noted in Chapter 1, SIDES considers several metrics, including selection rate (proportion of simulations with a subgroup selected), confirmation rate (proportion of simulations with a confirmed subgroup), and treatment effect fraction (the true effect in the confirmed subgroup divided by the true effect in the correct subgroup). Results from ASD focus on whether any subgroup was selected and whether the full population was selected. Additional metrics are not reported. We seek to expand on and further develop the metrics from SIDES. How to quantify and measure errors requires some consideration. Assuming we know the truth, one option is to consider the two possible errors for each of the $2^K$ points that divide the covariate space. Specifically, we can identify what percent of patients we incorrectly classify or what percent of patients with $\theta \geq \theta^*$ did we miss. On another level, we can consider the two errors related to the identified subgroup. Specifically, either within the selected subgroup (and in the subgroup not selected) the truth is $\theta > \theta^*$ or $\theta \leq \theta^*$. If the latter is true, we made an error. Both approaches provide metrics that are useful to summarize performance, so we focus on metrics that can be calculated when we know the truth and are applying LR and SHAPES over simulated replications of an experiment.

We introduce statistics calculated for the $q^{th}$ trial out of $Q$ total simulated trials. The statistics require that the truth is known. The decision for each point in Boolean space must be the same,

therefore these metrics are calculated for the $2^K$ points rather than by individual observations. Where $j \in \{1, 2, 3, ..., 2^K\}$ indexes the $2^K$ subcubes, $p_j$ as the true population proportion of the $j^{th}$ group, $S_j$ as an indicator of the true status of this group where $\theta_j$ is the true average effect within the subgroup, $S_j = 1$ indicates $\theta_j > 0$ (or generally $\theta_j > \delta$) and $S_j = 0$ indicates $\theta_j \leq 0$; and $\hat{S}_j$ as an indicator for the subgroup decision.

1. *Misclassification error* is the percent of the study population incorrectly classified in the $q^{th}$ trial, defined as:

$$error_q = \sum_{j=1}^{2^K} p_j (S_j - \hat{S}_j)^2 \in [0, 1].$$

   Lower values on this metric indicate better performance.

2. *Predictive error* is the percent of the predictive subgroup incorrectly classified, i.e. missed, in the $q^{th}$ trial. This is standardized to the size of the predictive subgroup in the population by the denominator.

$$miss_q = \frac{\sum_{j=1}^{2^K} p_j I((S_j - \hat{S}_j) = 1)}{\sum_{j=1}^{2^K} p_j S_j} \in [0, 1]$$

   Lower values on this metric indicate better performance.

3. *Positive predictive error* is similar to the concept of positive predictive value and is the percent of the identified group that actually benefits in the $q^{th}$ trial. This is standardized to the size of the predictive subgroup in the population by the denominator.

$$positive_q = \frac{\sum_{j=1}^{2^K} p_j \hat{S}_j I((S_j - \hat{S}_j) = 0)}{\sum_{j=1}^{2^K} p_j \hat{S}_j} \in [0, 1]$$

   Higher values on this metric indicate better performance.

We also consider average performance of these and other metrics over $Q$ trials.

1. *Any rejection rate* is the percent of trials where any null hypothesis is rejected, either within a subgroup or for the full population. This is similar to confirmation rate from SIDES, except we also consider rejections for the full population.

2. *Empirical power for $\theta^*$ or $\theta^*$-power* reports the frequency of trials where the effect in the identified subgroup is greater than or equal to $> \theta^*$.

3. *Correct group identification* is the percent of trials where the correct group is selected. For $q \in \{1, ..., Q\}$:

$$power_E = \frac{\sum_{q=1}^{Q} I(error_q = 0)}{Q} \in [0, 1].$$

When simulating under only heterogeneous effects we refer to this as *correct subgroup identification.* A score of one would represent perfect group identification all of the time. The SIDESscreen mansucript mentions an equivalent calculation.

4. *Empirical superset power* is the percent of trials where the subgroup selected is a superset of the exact predictive subgroup:

$$power_S = \frac{\sum_{q=1}^{Q} I(miss_q = 0)}{Q} \in [0, 1].$$

5. *Empirical positive predictive value power* captures the frequency of trials where the subgroup identified contains at least $(d \times 100)\%$ of patients that actually benefit:

$$power_{PPV} = \frac{\sum_{q=1}^{Q} I(positive_q > d)}{Q} \in [0, 1].$$

The usefulness of these metrics depend on the investigators' goals. If identifying a group that benefits precisely is important, then use correct group identification. If only treating patients where the treatment works is important, then use empirical positive predictive value power. If not missing any patients who benefit is most important, use empirical superset power. If the investigators plan to conduct a phase 3 trial based on an assumed effect of $\theta^*$, then use $\theta^*$-power to know the likelihood of having nominal power for the next trial.

### *2.6 Summary*

In this chapter we propose two general approaches to identify and test for predictive subgroups. The approaches differ in the number of subgroups considered and the restrictions on possible subgroup definitions. We discuss methods to account for the exploration in our procedures and present some novel metrics to evaluate performance. In the following chapters, we evaluate the performance of these methods using simulated data to describe their operating characteristics.

# Chapter 3

# SIMULATION STUDIES AND PERFORMANCE EVALUATION FOR LOGIC REGRESSION

In this chapter we evaluate the performance of Logic Regression in simulation studies where testing for a subgroup is done through stratification. Our primary focus is the setting of a binary endpoint with $K = 4$ with $n_{leaves} = 3$ because this setting provides a reasonable amount of exploration. We begin by describing the general set-up of the data generation mechanism for the simulation studies and then provide results from systematically manipulating aspects of the mechanism. For the first results we allocate some portion of the overall type I error to the full population analysis with the remainder left for the subgroup search. This split in conjunction with three utilities select between the full population and the best subgroup when both tests are significant. We then explore modifying $K$ and $n_{leaves}$, the subgroup definitions, the proportion of the subgroup, the baseline rate, the presence of prognostic effects, the use of a continuous endpoint and end with a summary of results. In the simulations we include a broader range of scenarios to provide more information to an investigator considering searching for subgroups than compared with ASD and SIDES.

## *3.1 Simulation Set-up*

### *3.1.1 Data Generating Mechanism*

To simulate binary outcomes we sample from a Bernoulli distribution. A general data generation model for the $i^{th}$ patient that incorporates predictive and prognostic effects is:

$$\begin{aligned} y_i|t_i,\vec{x}_i &\sim Bern(p_i) \\ &\text{where } p_i = \beta_0 + \omega\pi(\vec{x_i}) + \eta t_i + \overbrace{\theta_S(\phi(\vec{x_i}) * t_i)}^{\text{predictive terms}} \\ &\text{and } \phi,\pi : \{0,1\}^K \rightarrow \{0,1\}. \end{aligned}$$

The functions $\phi, \pi$ represent Boolean functions that define the predictive and prognostic subgroups, respectively, with the prevalence of the predictive subgroup equal to $\zeta$, where $\zeta = \Pr(\phi(\vec{x}_i) = 1)$. The main effect of treatment is $\eta$ and the predictive effect for a subgroup is $\theta_S$ with the marginal

effect of treatment equal to $\theta = \eta + \zeta\theta_S$. The model also has nuisance parameters for a prognostic effect of $\omega$, effects that were not modeled with SIDES or ASD, and a baseline response rate in the control arm of $\beta_0$. As $p_i$ is on the probability scale, it must hold that $p_i \in (0, 1)$. We also apply this model to the continuous setting, where $p_i$ is the mean of a normal distribution and some error term $\epsilon$ represents the variance.

### *3.1.2 Scenarios*

For a specific sample size we define the MCID as $\theta^*$, which corresponds to the difference that a one-sided test of level $\alpha = 0.10$ has 80% power to detect. Table 3.1 contains the corresponding $\theta^*$ for both binary and continuous endpoints for several $n$ based on formulas contained in Friedman et al. [15].

| $n$ | Binary | Continuous |
|---|---|---|
| 200 | 0.146 | 0.300 |
| 500 | 0.095 | 0.190 |
| 1000 | 0.064 | 0.135 |

Table 3.1: For various sample sizes (n), the $\theta^*$ that correspond to a one-sided test of level $\alpha = 0.10$ and power $\beta$=0.80 when using either continuous outcomes with variance=1 or binary outcomes with a baseline rate of 0.30.

Our simulation studies evaluate how LR discriminates between three general scenarios: an effect in the full population, an effect in a subgroup and no effect where each scenario is generated with the mechanism from the previous section. To aid in interpretation, within these general scenarios we target understanding performance when treatment effects are multiples of $\theta^*$. One specific alternative of interest is testing for an effect of $\theta^*$ in the full population. Testing for subgroup effects requires some assumption of an effect greater than $\theta^*$ to have power at or above the nominal level, so we consider subgroup effects of $\theta_S = \{1.5\theta^*, 2\theta^*\}$. Collectively we focus on the following scenarios (but consider others):

1. **Null**: (null truth) $\eta = 0$, $\theta_S = 0$
2. **Alternative 1**: (full population effect of $\theta^*$) $\eta = \theta^*$, $\theta_S = 0$
3. **Alternative 2**: (subgroup effect of $1.5\theta^*$) $\eta = 0$, $\theta_S = 1.5\theta^*$ where $Pr(\phi(\vec{x}_i) = 1) = \zeta$
4. **Alternative 3**: (subgroup effect of $2\theta^*$) $\eta = 0$, $\theta_S = 2\theta^*$ where $Pr(\phi(\vec{x}_i) = 1) = \zeta$.

Alternatives 1, 2, and 3 are design alternatives that an investigator could pre-specify and develop a design to ensure power across these alternatives. Additionally when $\zeta = 0.5$, alternatives 1 and 3 have the same marginal effect for the population. Describing this problem explicitly in terms of

| | Predictive | $Pr(X_1 = 1)$ | $Pr(X_2 = 1)$ | $Pr(X_3 = 1)$ | $Pr(X_4 = 1)$ |
|---|---|---|---|---|---|
| A | $X_1$ | 0.50 | 0.50 | 0.50 | 0.50 |
| B | $X_1 \vee X_2$ | 0.29 | 0.29 | 0.50 | 0.50 |
| C | $X_1 \wedge X_2$ | 0.71 | 0.71 | 0.50 | 0.50 |
| Z | $(X_1 \wedge X_2) \vee (X_3 \wedge X_4)$ | 0.54 | 0.54 | 0.54 | 0.54 |

Table 3.2: Boolean expressions for subgroup definitions A, B, C and Z. The probabilities for each covariate correspond to an overall subgroup prevalence of 0.5.

multiple alternatives hypotheses differs from SIDES or ASD which generally simulate under one or two alternatives. An investigator is likely to be uncertain about the true state of nature and providing multiple alternatives is informative.

#### *3.1.3 Predictive Subgroup Definitions ($\phi(\vec{x}_i)$)*

In performing simulations, we focus on predictive subgroup definitions that consist of one or two covariates (definitions A, B, C as listed in Table 3.9). These definitions form convexly connected or convexly co-connected subsets in Boolean space. For contrast, we include definition Z that has four covariates and is a disconnected set of points in Boolean space. While LR with $n_{leaves} \geq 2$ can recover definitions A, B and C, only with $n_{leaves} \geq 4$ can it recover Z. Additionally, we can interpret definition A as a single covariate defining a subgroup or it could represent a pre-specified Boolean expression. These definitions provide greater diversity than those explored with SIDES by also considering definitions that are not part of SHAPES and having an 'or' expression.

#### *3.1.4 Comparison Methods*

In reporting results, we compare LR to a full population only analysis. We considered implementing SIDES or ASD as a comparison. However, as these methods were developed assuming a sample size of 300 or more, no software is available to apply to simulations studies, and the selection of tuning parameters is unclear, there is more research needed to make these methods implementable for this setting.

### *3.2 Version of Logic Regression Implemented*

In Section 2.2 we propose three methods to account for prognostic factors with LR: *LR-U*, *LR-A*, and *LR-R*. These methods correspond to not controlling for prognostic effects, controlling by adjustment, or controlling by calculating residuals. We use LR performed on the control group to identify a prognostic subgroup and the corresponding estimated effect (required only for *LR-R*). For

each of the methods we can perform an interaction or stratified test to determine the significance of the selected subgroup, resulting in six total methods. The simulation studies use $n = 200$, $\alpha_F = 0.025$, $\alpha = 0.10$, $K = 4$, $n_{leaves} = 3$, $\omega = 0$, $\beta_0 = 0.30$, $\zeta = 0.50$ and $\theta_S = \{1.5\theta^*, 2\theta^*\}$ with a total of 10,000 replications for each $\theta_S$ and subgroup definition. We derive critical values from the *distributional assumption* approach where the correct covariate distribution and baseline response rate are assumed and 10,000 replications under the null are simulated. We apply the utility of *minimum p-value*.

We aim to determine rates of correct subgroup identification for each of the six methods under the four subgroup definitions. Table 3.3 contains the results. For the entries under the 'Decision' column an 'All' row reports the number of times the full population is selected, a 'Correct sg' row reports how often the correct subgroup is selected, an 'Other sg' row reports how often some other subgroup is selected, a '$\theta_{\hat{S}} \geq \theta^*$' row is a subset of the 'Other sg' row and reports how many of the selected incorrect subgroups had an effect greater than $\theta^*$ and the 'None' row indicates how often no hypothesis was rejected. The sum of these rows, removing the fourth, equals the 10,000 simulations run for each combination of definition, method, and $\theta_S$.

Correct subgroup identification is highest with *LR-U* using stratified testing. There is typically a slight decrease between *LR-A* and *LR-U*. Identification at least doubles comparing $\theta_S = 1.5\theta^*$ to $\theta_S = 2\theta^*$ for most methods. Identification is lowest with *LR-R* using interaction testing, where it is essentially zero, indicating that identifying a prognostic subgroup and estimating the effect on the same data adds systematic noise that makes a predictive subgroup non-recoverable.

We now consider which method to implement in the remainder of the chapter. As we generally expect prognostic effects among the $K$ covariates it follows that *LR-U* is optimistic, and using *LR-A* or *LR-R* more representative of how LR performs for most clinical trial data. The differences between *LR-U* and *LR-A* are small but *LR-A* does better than *LR-R*. We therefore proceed with using *LR-A* with a stratified analysis because of the increase in power relative to interaction tests and the comparable performance to *LR-U*.

### *3.3 Decision Rules and Utility Functions for* $K = 4$ *and* $n_{leaves} = \{1, 3\}$

We next consider the use of different decision rules and utility functions. In our simulations we set $\alpha = 0.10$, use three values of $\alpha_F = \{0.025, 0.05, 0.075\}$ and implement the utilities of *prefer subgroup*, select minimum standardized p-value (*minimum p-value*) and *prefer all* as described in Chapter 2. We determine critical values on the p-value scale assuming a correct baseline rate of 0.30 with no treatment or prognostic effects. Other parameters values are as described in the previous

| | | Interaction Testing | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LR-U | | | | LR-A | | | | LR-R | | | |
| | Decision | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta_S = 1.5\theta^*$ | All | 2453 | 2747 | 2756 | 2990 | 2073 | 2100 | 2294 | 2371 | 2497 | 2472 | 2693 | 2651 |
| | Correct sg | 201 | 436 | 588 | 0 | 175 | 320 | 434 | 0 | 6 | 5 | 7 | 0 |
| | Other sg | 3715 | 3176 | 2710 | 2781 | 2641 | 2497 | 2101 | 1988 | 554 | 512 | 646 | 546 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 3462 | 2899 | 1976 | 2170 | 2424 | 2219 | 1650 | 1562 | 140 | 70 | 60 | 42 |
| | None | 3631 | 3641 | 3946 | 4229 | 5111 | 5083 | 5171 | 5641 | 6943 | 7011 | 6654 | 6803 |
| $\theta_S = 2\theta^*$ | All | 3125 | 3358 | 3500 | 3947 | 2742 | 2869 | 3121 | 3281 | 3812 | 3933 | 4171 | 3978 |
| | Correct sg | 657 | 1185 | 1399 | 0 | 527 | 959 | 1138 | 0 | 7 | 11 | 9 | 0 |
| | Other sg | 4826 | 3939 | 3499 | 3880 | 3809 | 3281 | 2747 | 3023 | 551 | 421 | 666 | 551 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 4824 | 3935 | 3489 | 3868 | 3808 | 3261 | 2728 | 2996 | 479 | 209 | 593 | 408 |
| | None | 1392 | 1518 | 1602 | 2173 | 2922 | 2891 | 2994 | 3696 | 5630 | 5635 | 5154 | 5471 |
| | | Stratified Testing | | | | | | | | | | | |
| | | LR-U | | | | LR-A | | | | LR-R | | | |
| | Decision | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta_S = 1.5\theta^*$ | All | 1072 | 1494 | 959 | 1399 | 1207 | 1193 | 1372 | 1408 | 1035 | 1085 | 1186 | 1173 |
| | Correct sg | 211 | 537 | 660 | 0 | 173 | 375 | 506 | 0 | 100 | 220 | 209 | 0 |
| | Other sg | 5261 | 4530 | 4728 | 4802 | 4509 | 4549 | 3681 | 4158 | 4504 | 4522 | 4183 | 4396 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 4493 | 3820 | 2895 | 3235 | 3857 | 3913 | 2408 | 2996 | 2633 | 2593 | 1726 | 1968 |
| | None | 3456 | 3439 | 3653 | 3799 | 4111 | 3883 | 4441 | 4434 | 4361 | 4173 | 4422 | 4431 |
| $\theta_S = 2\theta^*$ | All | 1388 | 1706 | 1166 | 1738 | 1385 | 1340 | 1724 | 1898 | 1352 | 1421 | 1511 | 1587 |
| | Correct sg | 691 | 1457 | 1701 | 0 | 602 | 1196 | 1389 | 0 | 229 | 563 | 532 | 0 |
| | Other sg | 6388 | 5384 | 5624 | 6260 | 6000 | 5569 | 4457 | 5513 | 5943 | 5723 | 5365 | 5692 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 6378 | 5354 | 5572 | 6185 | 5985 | 5497 | 4371 | 5389 | 5538 | 4876 | 4719 | 4802 |
| | None | 1533 | 1453 | 1509 | 2002 | 2013 | 1895 | 2430 | 2589 | 2476 | 2293 | 2592 | 2721 |

Table 3.3: The use of the three versions of LR with testing via stratification or interaction in a simulation study with 10,000 replications. Parameters and values are $n = 200$, $\alpha_F = 0.025$, $\alpha = 0.10$, $K = 4$, $n_{leaves} = 3$, $\omega = 0$, $\beta_0 = 0.30$, $\zeta = 0.50$ and $\theta_S = \{1.5\theta^*, 2\theta^*\}$ with a total of 10,000 replications for each $\theta_S$ and subgroup definition.

section. Simulations use a full population effect of $\eta = \{0, 0.5\theta^*, \theta^*, 1.5\theta^*, 2\theta^*\}$ and subgroup effects with $\zeta = 0.50$ and $\theta_S = \{0, 0.5\theta^*, \theta^*, 1.5\theta^*, 2\theta^*, 2.5\theta^*, 3\theta^*\}$ (with $\eta = 0$). The combinations of $\alpha$-allocation, decision rules, and utilities partition the alternative hypothesis spaces into various decisions.

Figure 3.1 presents the results for the three utilities when $\alpha_F = 0.025$, $K = 4$ and, when applicable, A is the subgroup definition. A homogeneous effect is the truth for the top row of plots and the x-axis is $\theta$. The frequency of rejecting $H_F$ is in black and any $H_{\hat{S}}$ is in gray. The different utilities result in markedly different decisions, particularly under larger values of $\theta$ where significance is likely for both $H_F$ and $H_{\hat{S}}$. The *prefer subgroup* and *prefer all* utilities are the two extremes where we most commonly reject $H_F$ with the former and $H_{\hat{S}}$ with the latter. The *minimum p-value* utility reflects some balance between the two rules, but when $\theta > 0.10$ *minimum p-value* performs more closely with *prefer all*. Compared to an analysis with $\alpha_F = 0.10$ using the same data, represented by the dashed line in the plots, we never achieve comparable power when splitting $\alpha$.

The second row of Figure 3.1 presents results using the three utilities when there is a subgroup

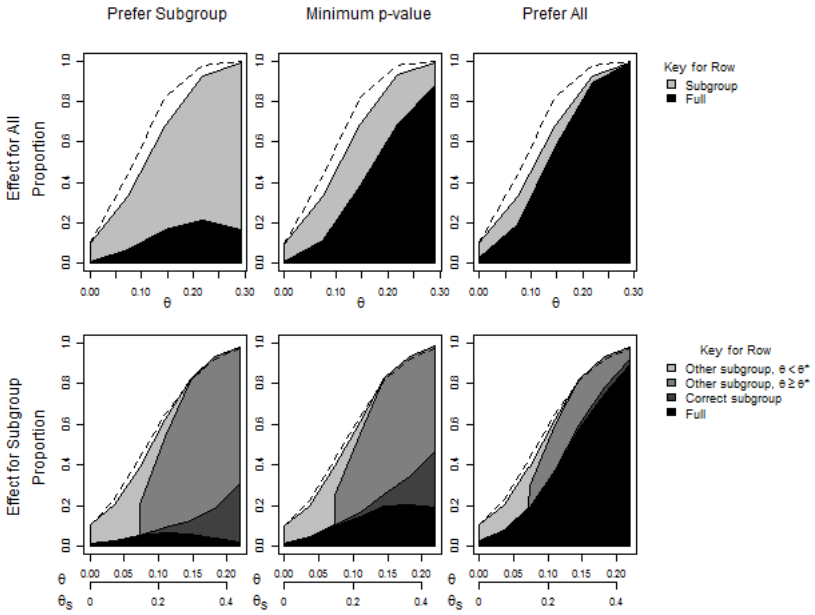



Figure 3.1: Performance of LR with $\alpha_F = 0.025$ and $\alpha = 0.10$ comparing three utilities (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, $K = 4$, and $n_{leaves} = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

effect for 50% of participants based on definition A. These plots have two scales: the subgroup effect, $\theta_S$, and the corresponding marginal effect, $\theta$. In this scenario, $\theta = 0.5\theta_S$. Rejection of $H_F$ is black. As there are subgroup effects, we present three decisions of interest: rejection for the correct subgroup (dark gray), rejection for an incorrect subgroup with an effect greater than or equal to $\theta^*$ (medium gray), and rejection for an incorrect subgroup with an effect less than $\theta^*$ (light gray). Using *prefer subgroup* and *minimum p-value* results in similar performance, with the latter having a somewhat higher rater of rejection for $H_{\hat{S}}$. With both utilities we observe an inverse u-shape for frequency of $H_F$ rejection. Rejection for the correct subgroup is relatively infrequent, occurring at most 20% of the time under an effect of $3\theta^*$ or a difference of 43.8% between arms, but rejection for a subgroup that has a true effect greater than or equal to $\theta^*$ is common, particularly as $\theta_S$ increases. The *prefer all* utility most frequently selects $H_F$, with subgroup decisions occurring 10% or less under most alternatives.

We next consider the same simulations where $\alpha_F = 0.05$ (Figure 3.2). Trends are similar to the previous figure. With *prefer subgroup* and *minimum p-value* the decision to select the full population increases by about 10% for the range of both full population and subgroup effects. With *prefer all*

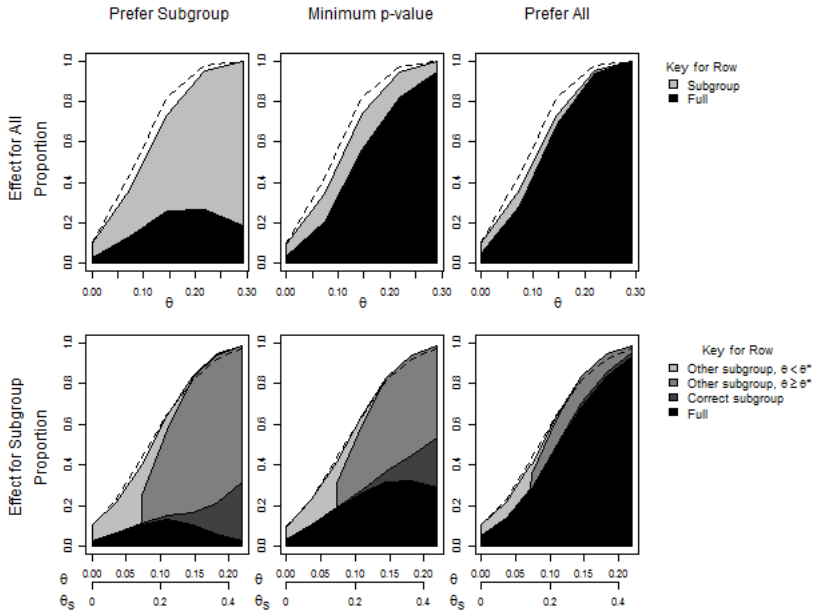



Figure 3.2: Performance of LR with $\alpha_F = 0.05$ and $\alpha = 0.10$ comparing three utilities (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, when $K = 4$, and $n_{leaves} = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

we rarely select any proper subgroup. Setting $\alpha_F = 0.075$, in Figure 3.3 we see another increase in rates of decisions for the full population. Using *minimum p-value* when there are moderate subgroup effects we decide roughly equally between the full population and some subgroup.

Of particular interest is the ability of each utility to discriminate between the presence of full population and subgroup effects for these values of $\alpha_F$. In one direction, *prefer subgroup* results in low rates of rejection of $H_F$ even when there are only full population effects. In another direction, *prefer all* results in high rates of rejection of $H_F$ even when there are only subgroup effects. The selection of how to allocate $\alpha$ and which utility to use is connected to the expectation of the investigators. Depending on their certainty about the presence or absence of heterogeneous effects either approach may be appropriate. We anticipate subgroup search methods to be used in situations where there is some balance between expecting full population effects and heterogeneous effects. The *minimum p-value* utility is a compromise between these two where if there are only full population effects $H_F$ is most frequently rejected and if there are subgroup effects $H_{\hat{S}}$ is most commonly rejected for moderate to large effects. We conclude that for LR with $K = 4$ and $n_{leaves} = 3$, selecting $\alpha_F = 0.025$ and *minimum p-value* provides balanced performance discriminating between subgroup

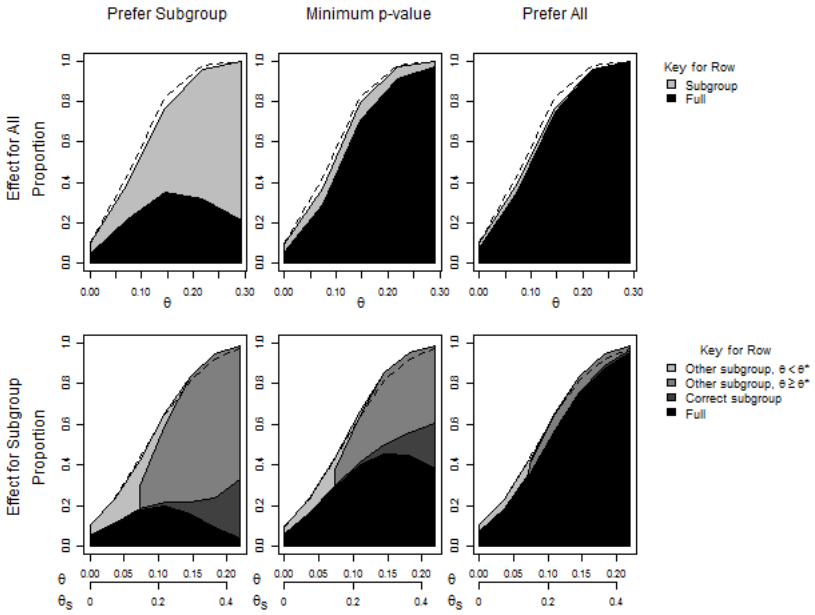



Figure 3.3: Performance of LR with $\alpha_F = 0.075$ and $\alpha = 0.10$ comparing three utilities (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, when $K = 4$, and $n_{leaves} = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

and full population effects.

Concerning identification, in these simulations LR rarely identifies the correct subgroup defined by a single covariate for definition A. LR tends to select a subgroup with two or three variables, i.e. overfitting. We ask if performance is improved when $n_{leaves}$ is equal to the true number of variables used to define the predictive subgroup. Figure 3.4 explores the utilities when using $\alpha_F = 0.025$ and $n_{leaves} = 1$. Decreasing the tuning parameter to 1 improves the frequency of correct subgroup identification, particularly for *prefer subgroup* and *minimum p-value*. The *minimum p-value* utility again balances deciding for the full population or subgroup when those are the correct decisions. With this utility we observe slightly higher rates of subgroup selection under homogeneous effects and slightly lower rates of full population selection under heterogeneous effects when compared to $n_{leaves} = 3$.

### *3.4 Subgroup Definitions*

In the previous section we identified settings that provide reasonable discriminatory performance under subgroup definition A. In this section, we apply these settings, *minimum p-value* and $\alpha_F =$

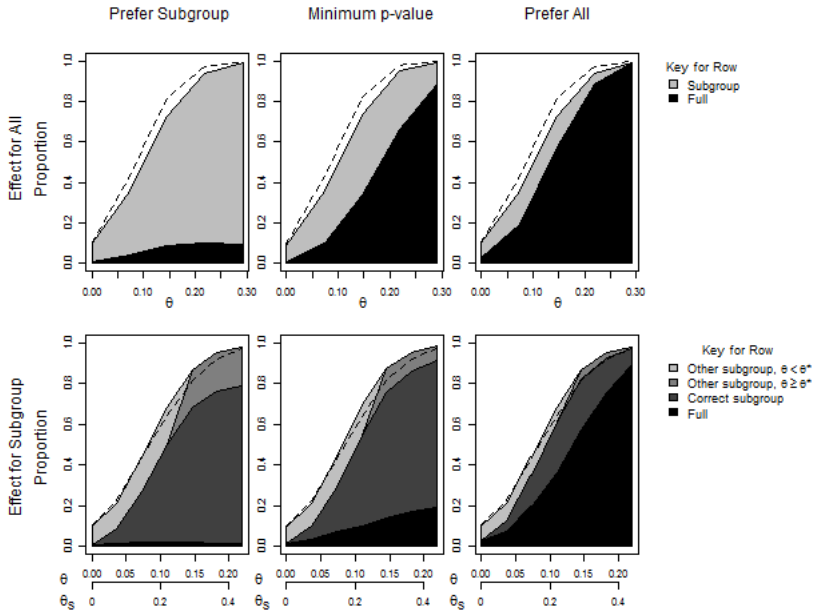



Figure 3.4: Performance of LR with $\alpha_F = 0.025$ and $\alpha = 0.10$ comparing three utilities (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, when $K = 4$, and $n_{leaves} = 1$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

0.025, to the other three definitions. Figure 3.5 compares subgroup definitions A, B, C, and Z when $n_{leaves} = 3$. Under homogeneous effects, modest differences in the covariate distributions as described in Table 3.9 do not change performance and there is always a power loss relative to using a full population analysis at $\alpha = 0.10$.

Under heterogeneous effects performance varies by definition. We observe greater rates of correct identification with subgroups defined by two variables (definitions B and C) than one (definition A). This indicates closer matches between $n_{leaves}$ and the number of covariates in the subgroup definition improves performance. While definition Z is not recoverable with $n_{leaves} = 3$, identification of some subgroup with an effect greater than or equal to $\theta^*$ appears comparable to the other scenarios. For all scenarios there is a relatively quick transition to detecting subgroups with effects less than $\theta^*$ to greater than or equal to $\theta^*$ that occurs between $\theta^*$ and $2\theta^*$ on the $\theta_S$ scale. For definitions A and B there are some values of $\theta^*$ where we reject more often than just testing in the full population. For definitions C and Z our method generally rejects less often than or equal to the full population analysis.

We next consider the various definitions when $n_{leaves} = 1$, shown in Figure 3.6. For definition A,

rates of correct subgroup identification improve. For all other scenarios we cannot recover the correct subgroup and performance appears comparable across the definitions with one small difference: with definition B we identify subgroups with effects greater than or equal to $\theta^*$ at smaller values of $\theta_S$. This is due to the structure of definition B, which is $X_1 \vee X_2$. With $n_{leaves} = 1$, LR can select either $X_1$ or $X_2$ when $\theta_S \geq \theta^*$ and the identified subgroup will have an effect of $\theta_S$. These single-covariate subgroups are subsets of the true predictive subgroup. With definitions C and Z, however, any subgroup defined by a single covariate is a mixture of participants that have either an effect of 0 or $\theta_S$; thus the treatment effect is diluted in any selected subgroup. With these definitions a full population analysis generally rejects more often. Under all scenarios and values of $n_{leaves}$ with large enough $\theta_S$ we frequently select a subgroup with $\theta_{\hat{S}} \geq \theta^*$.

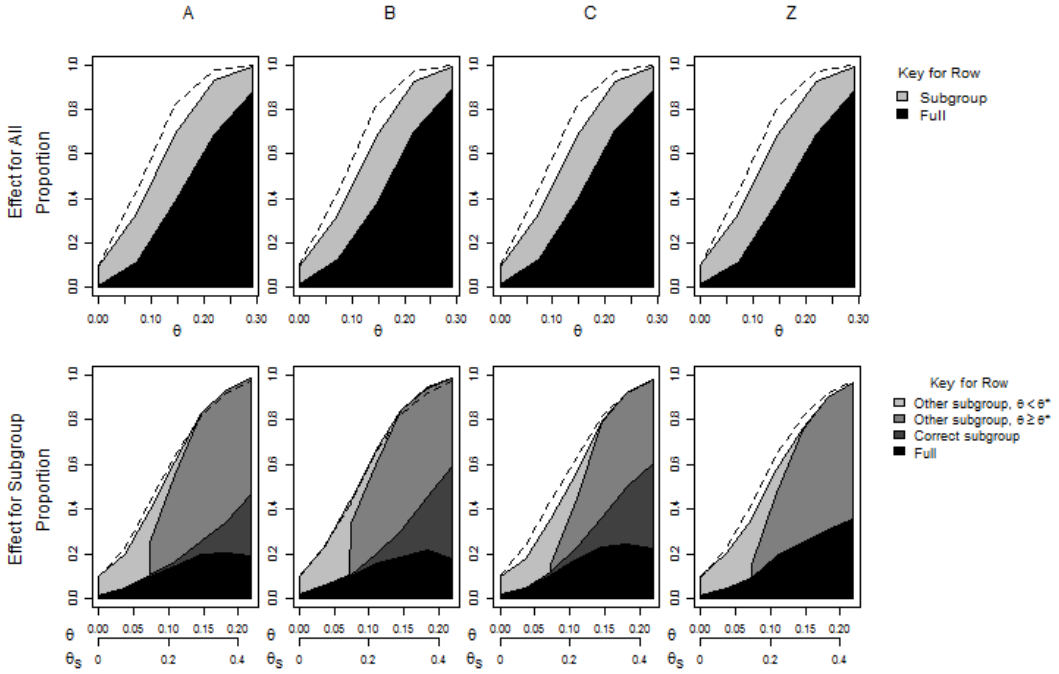



Figure 3.5: Performance of LR with $\alpha_F = 0.025$ and $\alpha = 0.10$ comparing four subgroup definitions (A, B, C and Z - the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, when $K = 4$, and $n_{leaves} = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

In summary, we observe that the definition of a subgroup impacts LR performance. For some values of $\theta_S$ and subgroups definitions A and B we achieve a greater rejection rate than using a full population analysis, while with others we do not.

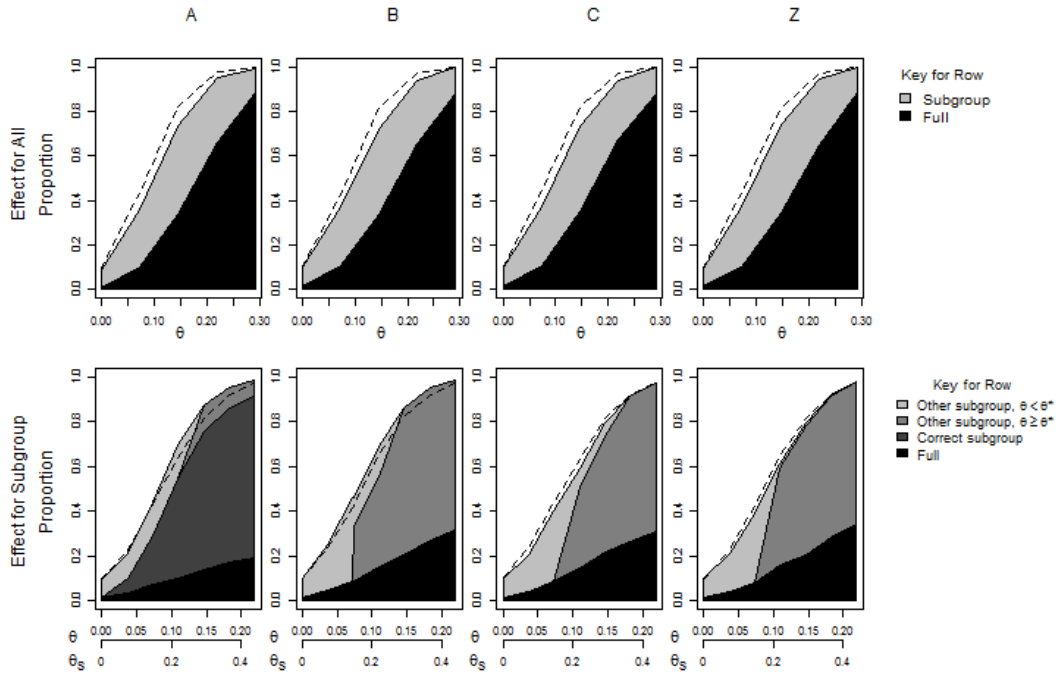



Figure 3.6: Performance of LR with $\alpha_F = 0.025$ and $\alpha = 0.10$ comparing four subgroup definitions (A, B, C and Z - the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, when $K = 4$, and $n_{leaves} = 1$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

### *3.5 Varying $K$ and $n_{leaves}$*

We now evaluate the *minimum p-value* utility, $\alpha_F = 0.025$ , $\alpha = 0.10$, and, when applicable, $\zeta = 0.5$ to the null and alternative scenarios described in the beginning of this chapter. We begin by considering $(K, n_{leaves}) = \{(1,1),(2,2),(4,3),(8,3)\}$. Table 3.4 provides a summary of power, correct group identification and rejection rates for a subgroup with $\theta_{\hat{S}} \geq \theta^*$, i.e. where the true effect in the selected group is greater than or equal to $\theta^*$. Table 3.5 details the decisions with 10,000 replications for each scenario. The entries in the 'Decision' column are the same as in Section 3.2, except in the 'for all' rows, which correspond to homogeneous effects. In this case the decision options are 'All' for the frequency of full population selection, 'Any sg' for the frequency of any subgroup selection and 'None' for when no null hypothesis is rejected. In both tables the 'Full' column reports the results from a full population analysis where $\alpha_F = \alpha = 0.10$, so no subgroups can be selected. We describe results for the null and specified alternative scenarios and for reference include additional values of $\theta = 1.5\theta^*$ and $\theta_S = \theta^*$ in the tables.

| Truth | | Full | $K=1, n_{leaves}=1$ A | B | C | Z | $K=2, n_{leaves}=2$ A | B | C | Z | $K=4, n_{leaves}=3$ A | B | C | Z | $K=8, n_{leaves}=3$ A | B | C | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency of any rejection | | | | | | | | | | | | | | | | | | |
| All | $\theta = 0$ | 9.7 | 10.9 | 9.8 | 10.7 | 10.1 | 10.6 | 10.7 | 10.4 | 9.8 | 9.6 | 9.6 | 9.5 | 10.1 | 10.0 | 9.7 | 10.0 | 10.0 |
| | $\theta = \theta^*$ | 81.5 | 70.8 | 69.9 | 70.7 | 70.5 | 69.1 | 69.3 | 70.4 | 69.8 | 68.7 | 67.5 | 68.2 | 68.8 | 68.8 | 67.6 | 67.2 | 65.5 |
| | $\theta = 1.5\theta^*$ | 97.1 | 92.9 | 93.4 | 93.8 | 93.2 | 92.5 | 93.4 | 93.4 | 93.1 | 93.4 | 92.2 | 92.4 | 92.4 | 92.1 | 92.6 | 91.9 | 90.6 |
| Subgroup | $\theta_S = \theta^*$ | 42.7 | 48.4 | 40.8 | 40.0 | 36.4 | 41.7 | 46.7 | 43.2 | 36.6 | 39.1 | 43.2 | 35.8 | 35.6 | 37.5 | 37.8 | 33.9 | 33.0 |
| | $\theta_S = 1.5\theta^*$ | 63.7 | 74.6 | 63.8 | 64.1 | 57.8 | 66.3 | 71.6 | 69.7 | 59.3 | 61.4 | 66.3 | 56.3 | 57.7 | 58.8 | 59.0 | 54.0 | 52.5 |
| | $\theta_S = 2\theta^*$ | 81.3 | 92.5 | 82.3 | 83.4 | 74.8 | 87.3 | 89.7 | 88.0 | 77.5 | 82.7 | 84.5 | 79.4 | 76.6 | 78.7 | 79.4 | 74.1 | 71.3 |
| Frequency of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ | | | | | | | | | | | | | | | | | | |
| Subgroup | $\theta_S = \theta^*$ | 0.0 | 41.0 | 26.7 | 0.0 | 0.0 | 25.5 | 36.7 | 18.5 | 9.3 | 14.8 | 24.0 | 4.9 | 5.5 | 11.1 | 11.7 | 2.3 | 2.2 |
| | $\theta_S = 1.5\theta^*$ | 0.0 | 66.2 | 40.6 | 53.1 | 35.8 | 51.9 | 58.3 | 53.0 | 31.1 | 40.6 | 44.3 | 26.9 | 28.0 | 30.8 | 30.2 | 21.0 | 18.7 |
| | $\theta_S = 2\theta^*$ | 81.3 | 92.5 | 81.1 | 83.0 | 72.0 | 87.3 | 89.4 | 88.0 | 76.7 | 82.6 | 84.0 | 78.7 | 75.7 | 78.7 | 79.1 | 70.5 | 69.6 |
| Frequency of correct decision | | | | | | | | | | | | | | | | | | |
| All | $\theta = 0$ | 90.3 | 89.1 | 90.2 | 89.3 | 89.9 | 89.4 | 89.3 | 89.6 | 90.2 | 90.4 | 90.4 | 90.5 | 89.9 | 90.0 | 90.4 | 90.0 | 90.0 |
| | $\theta = \theta^*$ | 81.5 | 36.6 | 36.2 | 36.2 | 36.1 | 37.2 | 35.3 | 34.7 | 35.3 | 38.2 | 37.3 | 39.1 | 39.4 | 44.0 | 41.6 | 40.7 | 39.6 |
| | $\theta = 1.5\theta^*$ | 97.1 | 67.2 | 67.7 | 68.3 | 69.3 | 65.5 | 65.0 | 64.3 | 67.0 | 69.0 | 70.0 | 70.5 | 69.1 | 72.5 | 72.2 | 70.3 | 69.6 |
| Subgroup | $\theta_S = \theta^*$ | 0.0 | 41.0 | 0.0 | 0.0 | 0.0 | 16.0 | 18.8 | 18.5 | 0.0 | 0.4 | 1.1 | 1.5 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 |
| | $\theta_S = 1.5\theta^*$ | 0.0 | 66.2 | 0.0 | 0.0 | 0.0 | 34.8 | 39.5 | 39.6 | 0.0 | 1.8 | 4.0 | 5.0 | 0.0 | 0.1 | 0.7 | 1.2 | 0.0 |
| | $\theta_S = 2\theta^*$ | 0.0 | 82.8 | 0.0 | 0.0 | 0.0 | 56.8 | 61.0 | 60.1 | 0.0 | 5.9 | 11.5 | 13.2 | 0.0 | 0.2 | 2.9 | 4.0 | 0.0 |

Table 3.4: Type I error and power for LR for variable $n_{leaves}$ with n=200, $\alpha = 0.10$, $\alpha_F = 0.025$, $\theta^* = 0.146$.

| | | | $K=1, n_{leaves}=1$ | | | | $K=2, n_{leaves}=2$ | | | | $K=4, n_{leaves}=3$ | | | | $K=8, n_{leaves}=3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 969 | 137 | 111 | 138 | 146 | 155 | 121 | 148 | 160 | 120 | 160 | 168 | 144 | 160 | 169 | 146 | 152 |
| for all | Any sg | – | 953 | 874 | 928 | 861 | 906 | 945 | 897 | 822 | 839 | 796 | 782 | 862 | 842 | 796 | 854 | 847 |
| | None | 9031 | 8910 | 9015 | 8934 | 8993 | 8939 | 8934 | 8955 | 9018 | 9041 | 9044 | 9050 | 8994 | 8998 | 9035 | 9000 | 9001 |
| $\theta = \theta^*$ | All | 8148 | 3663 | 3624 | 3620 | 3611 | 3716 | 3533 | 3474 | 3530 | 3824 | 3733 | 3909 | 3941 | 4399 | 4156 | 4068 | 3963 |
| for all | Any sg | – | 3420 | 3368 | 3451 | 3440 | 3195 | 3398 | 3569 | 3451 | 3043 | 3015 | 2914 | 2943 | 2484 | 2600 | 2650 | 2587 |
| | None | 1852 | 2917 | 3008 | 2929 | 2949 | 3089 | 3069 | 2957 | 3019 | 3133 | 3252 | 3177 | 3116 | 3117 | 3244 | 3282 | 3450 |
| $\theta = 1.5\theta^*$ | All | 9707 | 6724 | 6772 | 6826 | 6932 | 6552 | 6496 | 6427 | 6695 | 6901 | 6996 | 7053 | 6906 | 7245 | 7218 | 7026 | 6955 |
| for all | Any sg | – | 2564 | 2568 | 2557 | 2386 | 2698 | 2841 | 2917 | 2612 | 2436 | 2222 | 2186 | 2334 | 1967 | 2040 | 2160 | 2102 |
| | None | 293 | 712 | 660 | 617 | 682 | 750 | 663 | 656 | 693 | 663 | 782 | 761 | 760 | 788 | 742 | 814 | 943 |
| $\theta_S = \theta^*$ | All | 4274 | 632 | 1043 | 631 | 917 | 879 | 784 | 603 | 939 | 996 | 959 | 1036 | 944 | 1241 | 1147 | 1177 | 1085 |
| for 50% | Correct sg | – | 4100 | 0 | 0 | 0 | 1600 | 1877 | 1850 | 0 | 40 | 106 | 147 | 0 | 0 | 6 | 22 | 0 |
| | Other sg | – | 112 | 3035 | 3373 | 2721 | 1693 | 2011 | 1865 | 2724 | 2878 | 3259 | 2393 | 2616 | 2512 | 2625 | 2193 | 2212 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 2670 | 0 | 0 | 947 | 1795 | 0 | 930 | 1435 | 2292 | 343 | 548 | 1112 | 1161 | 205 | 222 |
| | None | 5726 | 5156 | 5922 | 5996 | 6362 | 5828 | 5328 | 5682 | 6337 | 6086 | 5676 | 6424 | 6440 | 6247 | 6222 | 6608 | 6703 |
| $\theta_S = 1.5\theta^*$ | All | 6368 | 821 | 2078 | 1008 | 1761 | 1331 | 1265 | 928 | 1689 | 1443 | 1564 | 1782 | 1921 | 2037 | 1801 | 2078 | 2085 |
| for 50% | Correct sg | – | 6615 | 0 | 0 | 0 | 3482 | 3953 | 3960 | 0 | 180 | 403 | 501 | 0 | 5 | 73 | 120 | 0 |
| | Other sg | – | 25 | 4297 | 5404 | 4019 | 1819 | 1943 | 2082 | 4243 | 4513 | 4662 | 3348 | 3849 | 3836 | 4023 | 3203 | 3166 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 4057 | 5311 | 3576 | 1710 | 1880 | 1340 | 3106 | 3878 | 4031 | 2188 | 2804 | 3077 | 2952 | 1984 | 1872 |
| | None | 3632 | 2539 | 3625 | 3588 | 4220 | 3368 | 2839 | 3030 | 4068 | 3864 | 3371 | 4369 | 4230 | 4122 | 4103 | 4599 | 4749 |
| $\theta_S = 2\theta^*$ | All | 8126 | 964 | 2883 | 1326 | 2743 | 1659 | 1284 | 1044 | 2674 | 1945 | 1860 | 2305 | 2493 | 2586 | 2445 | 2938 | 3020 |
| for 50% | Correct sg | – | 8281 | 0 | 0 | 0 | 5677 | 6103 | 6009 | 0 | 591 | 1154 | 1317 | 0 | 17 | 293 | 395 | 0 |
| | Other sg | – | 3 | 5347 | 7013 | 4736 | 1395 | 1578 | 1748 | 5079 | 5738 | 5432 | 4318 | 5164 | 5265 | 5203 | 4078 | 4114 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 5230 | 6977 | 4454 | 1390 | 1551 | 1745 | 4996 | 5726 | 5385 | 4245 | 5081 | 5265 | 5174 | 3720 | 3944 |
| | None | 1874 | 752 | 1770 | 1661 | 2521 | 1269 | 1035 | 1199 | 2247 | 1726 | 1554 | 2060 | 2343 | 2132 | 2059 | 2589 | 2866 |

Table 3.5: Use of LR with n=200, $\alpha = 0.10$, $\alpha_F = 0.025$, $\theta^* = 0.146$.

**Null:** Simulating under the null, rejection occurs between 9.5 and 10.9% of the time. When we reject, the ratio of decisions for the full population to any proper subgroup is approximately 1:9.

**Homogeneous effects:** With $\theta = \theta^*$, the rejection rate ranges from 65.5 to 70.2% under the various definitions and values of $K$. These rates are always lower than the full population only rate of 81.5%; corresponding to a power loss from 11.3 to 14.0%. When we reject with $K = 1$ we select a proper subgroup about half the time and the full population for the other half. For $K = 8$ this ratio is closer to 2:3. If the effect is $\theta = 1.5\theta^*$, when exploring for subgroups we reject between 90.6 to 93.8% of the time compared to 97.1% for the full population only analysis. The power loss decreases for this alternative: between 3.3 and 6.5%.

**Heterogeneous effects:** With $\theta_S = 1.5\theta^*$, we reject between 52.5 and 74.6% of the trials with the highest rate when $K = 1$ under definition A (with a 66.2% rate of correct subgroup identification) and the lowest rate when $K = 8$ under definition Z (with a 0.0% rate of correct subgroup identification). In comparison, a full population analysis rejects with 63.7% frequency. The frequency of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ ranges from 18.7 to 66.2% compared to a rate of 0.0% for the full population as the marginal effect is $0.75\theta^*$. As evident in the 'Correct sg' rows of Table 3.5, as we consider more non-informative covariates we generally decrease our ability to identify the correct subgroup. For example under scenario A with $K = 1$ we have 66.2% correct identification which decreases to $< 0.1\%$ with $K = 8$.

When $\theta_S = 2\theta^*$, the frequency of any rejection is between 71.3 and 92.5% compared to the full population analysis rate of 81.3%. Similarly, the rate of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ ranges between 69.6 and 92.5%. With $K = \{1, 2\}$, the majority of rejections are for correct subgroups when the definition is recoverable. With $K = \{4, 8\}$ the selected subgroup is typically composed of a either a proper subset of the correct subgroup or a mixture from the correct subgroup and its complement.

We next consider fixing $n_{leaves} = 1$ while varying $K$. Presented in Table 3.7 are results with $\alpha_F = 0.025$, $\alpha = 0.10$ and the *minimum p-value* utility for $K = \{1, 2, 4, 8\}$. Rejection rates are reported in Table 3.6.

**Null**: Under the null distribution, the overall empirical type I error rate ranges from 9.0 to 10.9%. We reject between 1.1 and 1.5% of the time for the full population and between 8.4 and 9.6% for some subgroup.

**Homogeneous effects:** For $\theta = \theta^*$, a full population analysis rejects with 81.5% frequency. With LR the frequency of any rejection increases from an average across the definitions of 70.5% when $K = 1$ to 73.0% when $K = 8$. Underlying the slight increase in rejection are two trends as $K$

| | | | $K=1, n_{leaves}=1$ | | | | $K=2, n_{leaves}=1$ | | | | $K=4, n_{leaves}=1$ | | | | $K=8, n_{leaves}=1$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| Frequency of any rejection | | | | | | | | | | | | | | | | | | |
| All | $\theta = 0$ | 9.7 | 10.9 | 9.8 | 10.7 | 10.1 | 10.0 | 10.1 | 10.6 | 9.5 | 9.0 | 9.9 | 9.8 | 9.8 | 10.7 | 10.0 | 9.5 | 10.6 |
| | $\theta = \theta^*$ | 81.5 | 70.8 | 69.9 | 70.7 | 70.5 | 70.5 | 70.4 | 70.8 | 70.1 | 73.2 | 72.2 | 73.1 | 73.9 | 74.5 | 73.0 | 72.0 | 72.6 |
| | $\theta = 1.5\theta^*$ | 97.1 | 92.9 | 93.4 | 93.8 | 93.2 | 92.8 | 93.7 | 93.8 | 93.1 | 95.2 | 94.0 | 94.1 | 94.3 | 94.4 | 94.4 | 93.8 | 93.8 |
| Subgroup | $\theta_S = \theta^*$ | 42.7 | 48.4 | 40.8 | 40.0 | 36.4 | 45.7 | 43.9 | 41.5 | 37.4 | 43.9 | 46.4 | 39.8 | 38.8 | 43.6 | 44.5 | 36.3 | 38.5 |
| | $\theta_S = 1.5\theta^*$ | 63.7 | 74.6 | 63.8 | 64.1 | 57.8 | 71.0 | 66.9 | 65.0 | 59.2 | 69.7 | 69.0 | 59.3 | 61.4 | 67.7 | 67.2 | 56.6 | 58.9 |
| | $\theta_S = 2\theta^*$ | 81.3 | 92.5 | 82.3 | 83.4 | 74.8 | 90.1 | 85.9 | 84.3 | 76.3 | 87.3 | 86.3 | 79.8 | 79.5 | 86.8 | 84.2 | 75.9 | 77.4 |
| Frequency of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ | | | | | | | | | | | | | | | | | | |
| Subgroup | $\theta_S = \theta^*$ | 0.0 | 41.0 | 26.7 | 0.0 | 0.0 | 31.5 | 32.2 | 0.0 | 0.0 | 21.5 | 24.7 | 0.0 | 0.0 | 18.0 | 19.4 | 0.0 | 0.0 |
| | $\theta_S = 1.5\theta^*$ | 0.0 | 66.2 | 40.6 | 53.1 | 35.8 | 57.0 | 47.9 | 56.9 | 41.4 | 44.6 | 40.7 | 36.2 | 43.8 | 38.9 | 37.3 | 24.0 | 28.9 |
| | $\theta_S = 2\theta^*$ | 81.3 | 92.5 | 81.1 | 83.0 | 72.0 | 90.1 | 85.7 | 84.2 | 74.9 | 87.3 | 86.0 | 73.8 | 78.6 | 86.8 | 84.2 | 64.2 | 66.0 |
| Frequency of correct decision | | | | | | | | | | | | | | | | | | |
| All | $\theta = 0$ | 90.3 | 89.1 | 90.2 | 89.3 | 89.9 | 90.0 | 89.9 | 89.4 | 90.5 | 91.0 | 90.1 | 90.2 | 90.3 | 89.4 | 90.0 | 90.5 | 89.4 |
| | $\theta = \theta^*$ | 81.5 | 36.6 | 36.2 | 36.2 | 36.1 | 36.3 | 35.7 | 35.0 | 35.6 | 34.2 | 33.7 | 35.1 | 34.3 | 33.9 | 33.9 | 33.3 | 32.0 |
| | $\theta = 1.5\theta^*$ | 97.1 | 67.2 | 67.7 | 68.3 | 69.3 | 64.8 | 65.6 | 66.7 | 67.8 | 65.9 | 65.7 | 67.7 | 64.6 | 64.0 | 64.2 | 62.7 | 61.5 |
| Subgroup | $\theta_S = \theta^*$ | 0.0 | 41.0 | 0.0 | 0.0 | 0.0 | 31.5 | 0.0 | 0.0 | 0.0 | 21.5 | 0.0 | 0.0 | 0.0 | 18.0 | 0.0 | 0.0 | 0.0 |
| | $\theta_S = 1.5\theta^*$ | 0.0 | 66.2 | 0.0 | 0.0 | 0.0 | 57.0 | 0.0 | 0.0 | 0.0 | 44.6 | 0.0 | 0.0 | 0.0 | 38.9 | 0.0 | 0.0 | 0.0 |
| | $\theta_S = 2\theta^*$ | 0.0 | 82.8 | 0.0 | 0.0 | 0.0 | 76.6 | 0.0 | 0.0 | 0.0 | 62.1 | 0.0 | 0.0 | 0.0 | 61.2 | 0.0 | 0.0 | 0.0 |

Table 3.6: Type I error and power for LR for $n_{leaves} = 1$ with n=200, $\alpha = 0.10$, $\alpha_F = 0.025$, $\theta^* = 0.146$.

| | | | $K=1, n_{leaves}=1$ | | | | $K=2, n_{leaves}=1$ | | | | $K=4, n_{leaves}=1$ | | | | $K=8, n_{leaves}=1$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 969 | 137 | 111 | 138 | 146 | 141 | 115 | 136 | 152 | 112 | 139 | 141 | 133 | 141 | 124 | 109 | 104 |
| for all | Any sg | – | 953 | 874 | 928 | 861 | 855 | 896 | 923 | 797 | 792 | 852 | 844 | 842 | 924 | 876 | 838 | 959 |
| | None | 9031 | 8910 | 9015 | 8934 | 8993 | 9004 | 8989 | 8941 | 9051 | 9096 | 9009 | 9015 | 9025 | 8935 | 9000 | 9053 | 8937 |
| $\theta = \theta^*$ | All | 8148 | 3663 | 3624 | 3620 | 3611 | 3633 | 3566 | 3498 | 3555 | 3415 | 3366 | 3508 | 3434 | 3385 | 3388 | 3331 | 3202 |
| for all | Any sg | – | 3420 | 3368 | 3451 | 3440 | 3417 | 3473 | 3577 | 3457 | 3903 | 3851 | 3805 | 3954 | 4063 | 3911 | 3871 | 4058 |
| | None | 1852 | 2917 | 3008 | 2929 | 2949 | 2950 | 2961 | 2925 | 2988 | 2682 | 2783 | 2687 | 2612 | 2552 | 2701 | 2798 | 2740 |
| $\theta = 1.5\theta^*$ | All | 9707 | 6724 | 6772 | 6826 | 6932 | 6482 | 6555 | 6667 | 6783 | 6591 | 6571 | 6766 | 6462 | 6403 | 6424 | 6265 | 6153 |
| for all | Any sg | – | 2564 | 2568 | 2557 | 2386 | 2799 | 2811 | 2710 | 2530 | 2924 | 2827 | 2640 | 2964 | 3036 | 3015 | 3115 | 3231 |
| | None | 293 | 712 | 660 | 617 | 682 | 719 | 634 | 623 | 687 | 485 | 602 | 594 | 574 | 561 | 561 | 620 | 616 |
| $\theta_S = \theta^*$ | All | 4274 | 632 | 1043 | 631 | 917 | 692 | 989 | 484 | 854 | 750 | 831 | 828 | 765 | 794 | 788 | 893 | 846 |
| for 50% | Correct sg | – | 4100 | 0 | 0 | 0 | 3153 | 0 | 0 | 0 | 2150 | 0 | 0 | 0 | 1799 | 0 | 0 | 0 |
| | Other sg | – | 112 | 3035 | 3373 | 2721 | 721 | 3400 | 3663 | 2884 | 1494 | 3804 | 3155 | 3119 | 1766 | 3662 | 2735 | 3009 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 2670 | 0 | 0 | 0 | 3223 | 0 | 0 | 0 | 2471 | 0 | 0 | 0 | 1937 | 0 | 0 |
| | None | 5726 | 5156 | 5922 | 5996 | 6362 | 5434 | 5611 | 5853 | 6262 | 5606 | 5365 | 6017 | 6116 | 5641 | 5550 | 6372 | 6145 |
| $\theta_S = 1.5\theta^*$ | All | 6368 | 821 | 2078 | 1008 | 1761 | 956 | 1860 | 771 | 1490 | 1002 | 1480 | 1427 | 1554 | 1255 | 1378 | 1636 | 1541 |
| for 50% | Correct sg | – | 6615 | 0 | 0 | 0 | 5700 | 0 | 0 | 0 | 4460 | 0 | 0 | 0 | 3894 | 0 | 0 | 0 |
| | Other sg | – | 25 | 4297 | 5404 | 4019 | 449 | 4831 | 5733 | 4431 | 1506 | 5418 | 4501 | 4582 | 1621 | 5342 | 4022 | 4350 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 4057 | 5311 | 3576 | 0 | 4786 | 5694 | 4142 | 0 | 4071 | 3624 | 4382 | 0 | 3727 | 2404 | 2895 |
| | None | 3632 | 2539 | 3625 | 3588 | 4220 | 2895 | 3309 | 3496 | 4079 | 3032 | 3102 | 4072 | 3864 | 3230 | 3280 | 4342 | 4109 |
| $\theta_S = 2\theta^*$ | All | 8126 | 964 | 2883 | 1326 | 2743 | 1166 | 2437 | 992 | 2317 | 1385 | 2047 | 2150 | 2025 | 1470 | 2041 | 2544 | 2326 |
| for 50% | Correct sg | – | 8281 | 0 | 0 | 0 | 7664 | 0 | 0 | 0 | 6214 | 0 | 0 | 0 | 6124 | 0 | 0 | 0 |
| | Other sg | – | 3 | 5347 | 7013 | 4736 | 176 | 6149 | 7437 | 5311 | 1132 | 6581 | 5825 | 5923 | 1091 | 6380 | 5049 | 5413 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 5230 | 6977 | 4454 | 176 | 6137 | 7433 | 5173 | 1132 | 6551 | 5233 | 5831 | 1091 | 6374 | 3875 | 4277 |
| | None | 1874 | 752 | 1770 | 1661 | 2521 | 994 | 1414 | 1571 | 2372 | 1269 | 1372 | 2025 | 2052 | 1315 | 1579 | 2407 | 2261 |

Table 3.7: Use of LR with n=200, $\alpha = 0.10$, $\alpha_F = 0.025$, $\theta^* = 0.146$.

increases: a decrease in rejection for the full population from an average of 36.3 to 33.3% and an increase in subgroup rejection from 34.2 to 39.8%. Similar trends are present when $\theta = 1.5\theta^*$.

**Heterogeneous effects:** For $\theta_S = 1.5\theta^*$ and subgroup definition A, the identification of the correct subgroup ranges from 66.2% when $K = 1$ to 38.9% when $K = 8$. Here we assume the correct variable, $X_1$, is always included in the analysis making this result a best case scenario. As a comparison, if the correct variable is $X_4$ the correct subgroup rates are decreased to 0.0% with $K = \{1, 2\}$. For this covariate distribution, value of $\theta_S$ and scenario the proportion of rejections for the correct subgroup and for a subgroup with $\theta_{\hat{S}} \geq \theta^*$ are equal.

For the other subgroup definitions we never recover the correct subgroup because $n_{leaves} = 1$. Identifying a subgroup with $\theta_{\hat{S}} \geq \theta^*$ occurs most frequently when $K$ corresponds to the number of covariates in the subgroup definition. For example, with definition C and $\theta_S = 1.5\theta^*$ the performance is best when $K = 2$ with a rate of 56.9% compared to when $K = 1$ with a rate of 53.1% or when $K = 4$ with a rate of 36.2%. With $\theta_S = 2\theta^*$, we observe a similar trend.

**Summary and general trends for heterogeneous effects:** We now summarize trends using overall rejection, power for $\theta^*$ and correct subgroup identification rates as $K$ and $n_{leaves}$ varies. In the following plots, $K = 0$ or $n_{leaves} = 0$ represent an analysis of only the full population. Figure 3.7 correponds to results shown in Table 3.4 for $\theta_S = 1.5\theta^*$ and Figure 3.8 to Table 3.6. Additionally, Figure 3.9 considers $n_{leaves} = \{0, 1, 2, 3, 4\}$ for $K = 4$. We generally observe that performance is related to the agreement between $K$, $n_{leaves}$ and the subgroup definition. The best performance occurs when these agree, but there is wide variation dependending on our metric.

When the correct subgroup is recoverable, rates of correct identification are highest when the subgroup definition, $n_{leaves}$ and $K$ agree, as represented by the green peaks. As $n_{leaves}$ or $K$ increases beyond this, performance drops to near 0 for most scenarios, except with definition A and $n_{leaves} = 1$ where rates decrease slowly. This suggests restricting to only informative variables is critical if the goal is correct subgroup identification. Perhaps LR is best used to derive the subgroup definition when we know which variables are important, their number is small, and we want to determine their Boolean relationship. Figure 3.9 also suggests more complicated definitions like Z are rarely identified with the simulation parameters, even when $K$ and $n_{leaves}$ are correct. As $K$ or $n_{leaves}$ increases, rates of any rejection decrease much slower and more consistently than other metrics, as represented by the slight negative slope of the black lines in all three figures. This relationship occurs across subgroup definitions and for both $\theta_S = 1.5\theta^*$ and $\theta_S = 2\theta^*$. When $\theta_S = 2\theta^*$, identification of a group with $\theta_{\hat{S}} \geq \theta^*$ is similar to rates of any rejection. When $\theta_S = 1.5\theta^*$ the rates mimic the correct subgroup identification rates when $n_{leaves} = 3$ but tends to decrease much less as $n_{leaves}$

or $K$ increases. Under heterogeneous effects we sometimes do better and sometimes do worse than if we only analyzed the full population. With rates of any rejection we see a bump at the true $n_{leaves}$ except with scenario Z. With power for $\theta^*$ we always do better looking for subgroups when $\theta_S = 1.5\theta^*$ as the marginal effect is less than $\theta^*$, but with $\theta_S = 2\theta^*$ the marginal effect is equal to $\theta^*$ so an advantage is not always present.

Overall, these results suggest that if our priority is any rejection, we will not experience much difference using LR, but for correct subgroup identification or power for $\theta^*$ we have gains. In the remaining sections on LR we use $K = 4$ with $n_{leaves} = 3$ as these values provide a balance between a completely exploratory and a confirmatory analysis.

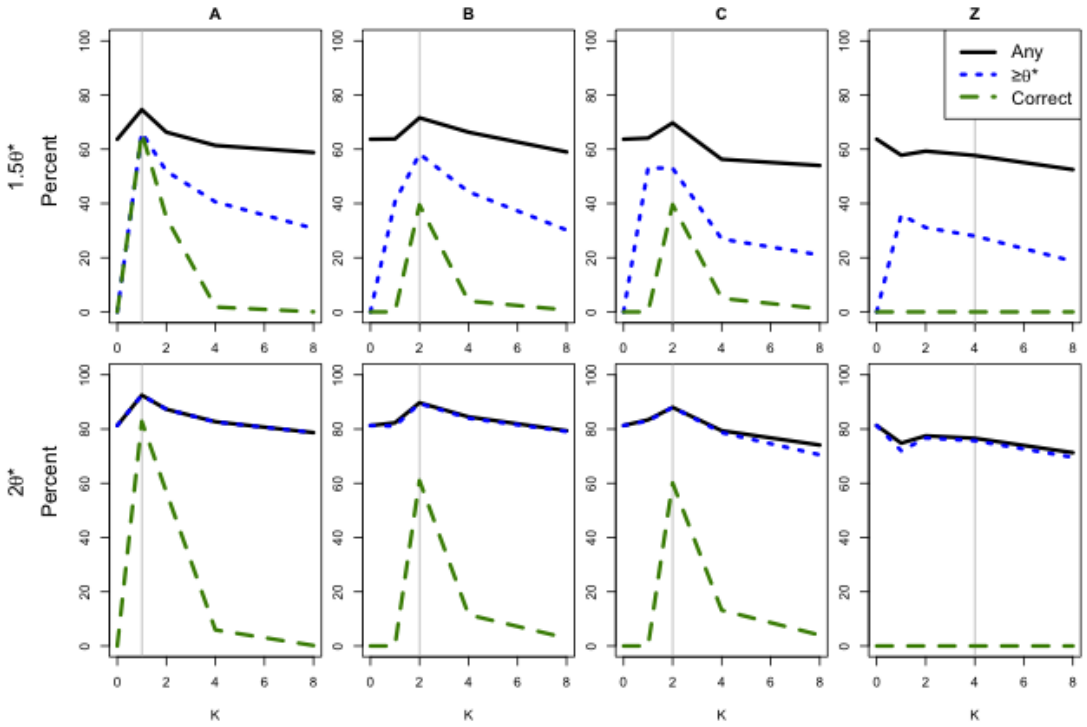



Figure 3.7: Performance of LR with $n_{leaves} = minimum(3, K)$ when $\theta_S = 1.5\theta^*$ (top row) and $\theta_S = 2\theta^*$ (bottom row) for increasing K in terms of any rejection (Any), rejection for a subgroup with a moderate effect ($\geq \theta^*$ ) or for the correct subgroup (Correct). The gray vertical line represents the true number of covariates in the subgroup definition.

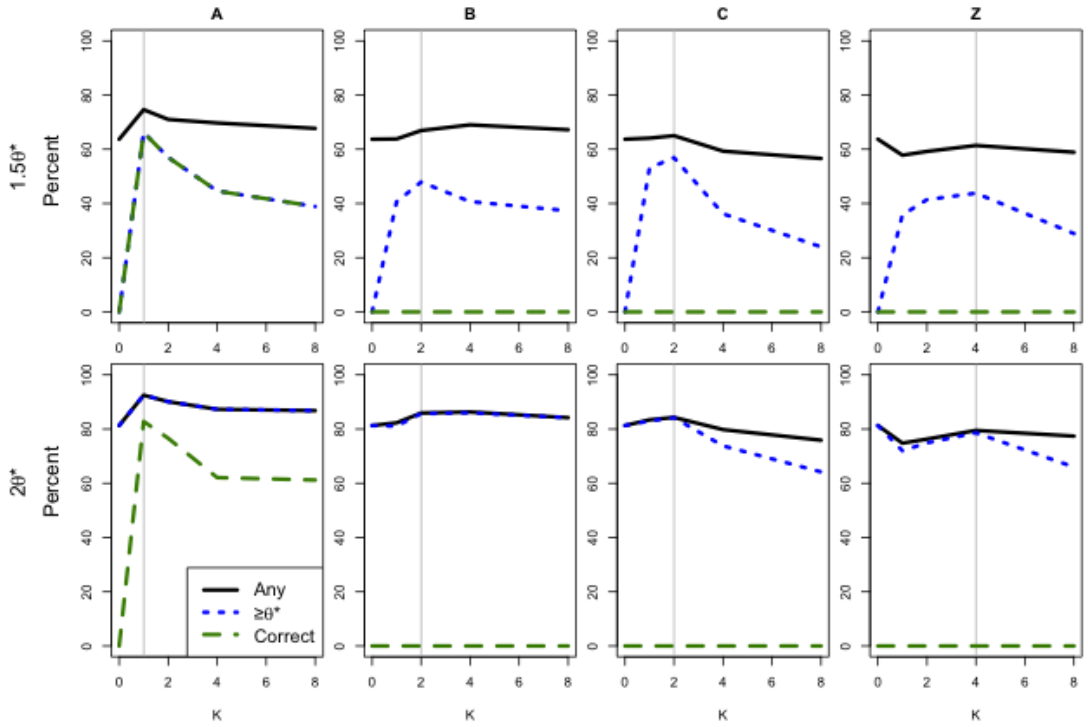



Figure 3.8: Performance of LR with $n_{leaves} = 1$ when $\theta_S = 1.5\theta^*$ (top row) and $\theta_S = 2\theta^*$ (bottom row) for increasing K in terms of any rejection (Any), rejection for a subgroup with a moderate effect ($\geq \theta^*$ ) or for the correct subgroup (Correct). The gray vertical line represents the true number of covariates in the subgroup definition.

### *3.6 Varying the Proportion of the Subgroup ($\zeta$)*

In previous sections, we simulate a subgroup effect for $\zeta = 0.50$ of the enrolled population; however, uncertainty about the definition of a predictive subgroup implies uncertainty about the subgroup size. The collection of candidate subgroup sizes is determined by LR via the tuning parameter $n_{leaves}$ and the observed covariate distribution. For each candidate subgroup, the observed and true prevalence relative to the enrolled population ranges from 0 to 1. Among these candidates, subgroups with prevalences in the middle, say between 0.20 and 0.80 is of greatest interest as there is sufficient information; we might expect a smaller identified subgroup is a random high and a larger identified group is the result of removing a subgroup that is a random low.

To explore how subgroup prevalence influences performance in our simulation studies we use

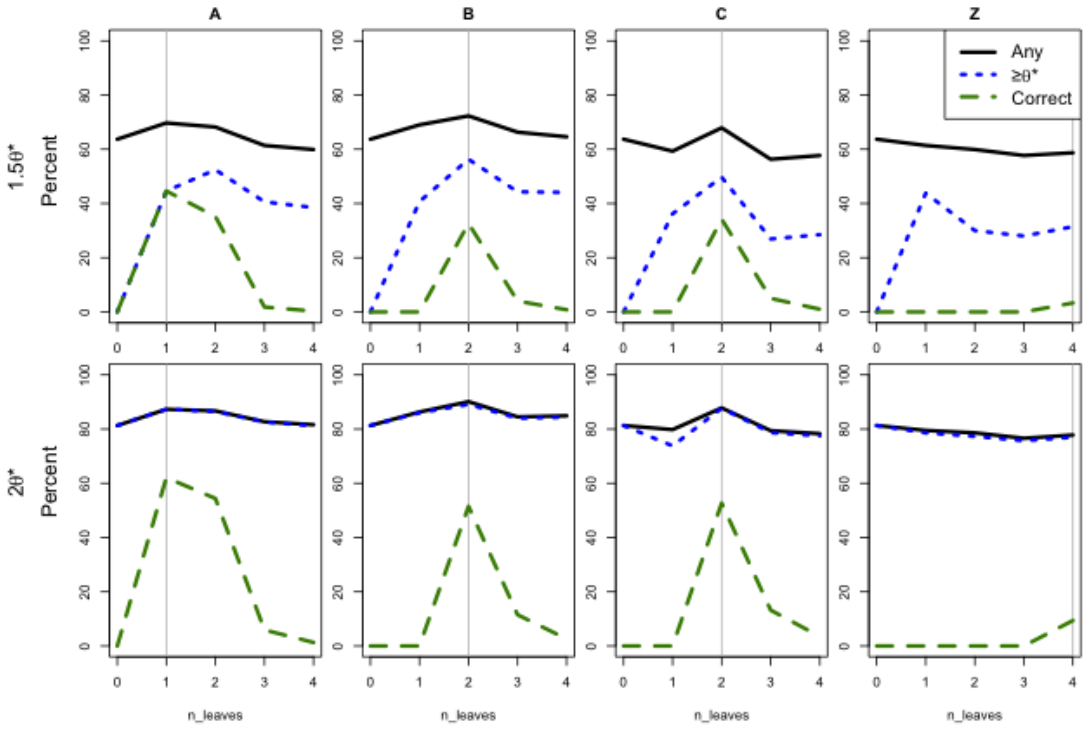



Figure 3.9: Performance of LR with $K = 4$ when $\theta_S = 1.5\theta^*$ (top row) and $\theta_S = 2\theta^*$ (bottom row) for increasing $n_{leaves}$ in terms of any rejection (Any), rejection for a subgroup with a moderate effect ($\geq \theta^*$ ) or for the correct subgroup (Correct). The gray vertical line represents the true number of covariates in the subgroup definition.

$\zeta = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. To aid interpretation we take two general approaches for describing the effect: fixing $\theta_S$ or $\theta$. With the former, we consider $\theta_S = 1.5\theta^*$, which corresponds to values of $\theta$ from 0.002 to 0.22 for the range of $\zeta$ and $\theta_S = 2\theta^*$, corresponding to $\theta$ from 0.003 to 0.29. This approach may be consistent with an investigator who hypothesizes the magnitude of a subgroup effect yet is uncertain as to the size of the subgroup. Alternatively, we fix $\theta$ and allow $\theta_S$ to vary for the values of $\zeta$. Here we do not consider the full range of $\zeta$ as we must have $\beta_0 + \theta_S < 1$ where $\beta_0 = 0.3$. We first consider $\theta = 0.75\theta^*$, which is equivalent to various combinations of $(\zeta, \theta_S)$ ranging from (0.2, 0.55) to (0.99, 0.11), including the previous case of (0.5, $1.5\theta^*$). We also fix $\theta = \theta^*$ with underlying combinations ranging from (0.3, 0.49) to (0.99, 0.15), including the case presented in previous sections of (0.5, $2\theta^*$). The fixed $\theta$ approach may be applicable when one is

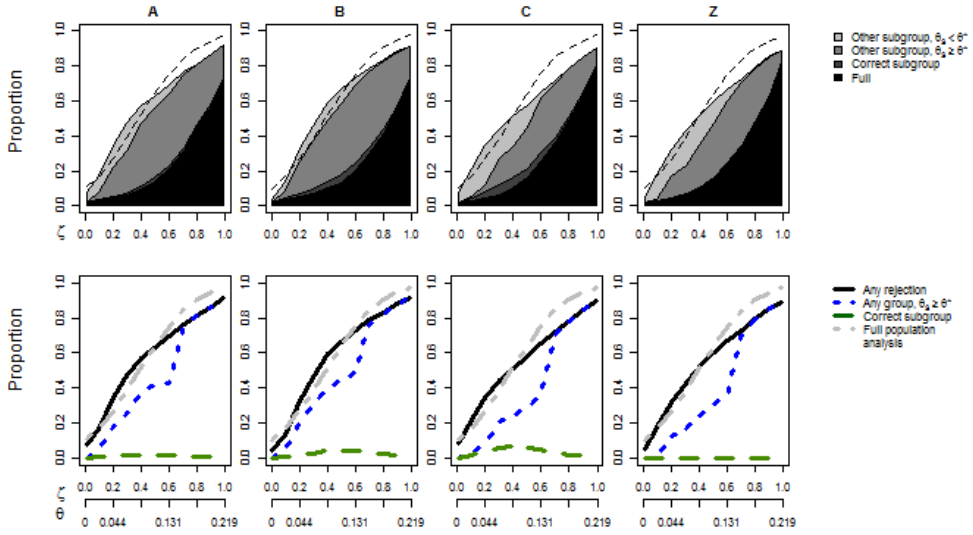



Figure 3.10: Performance of LR with $K = 4$, $n_{leaves} = 3$ when holding $\theta_S = 1.5\theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

hypothesizing an overall effect, but has uncertainty about the presence and magnitude of subgroup effects.

In simulations we again use the *minimum p-value* utility with $\alpha_F = 0.025$, $K = 4$ and $n_{leaves} = 3$. Figure 3.10 shows results when $\theta_S = 1.5\theta^*$ and Figure 3.11 when $\theta_S = 2\theta^*$. The top row of plots shows the various decisions for the values of $\zeta$ with the dashed line indicating the rejection rate of the full population analysis. The bottom row of the plots presents the performance of the various performance metrics. For all subgroup definitions and both values of $\theta_S$ for most values of $\zeta < 0.5$ LR has a higher rejection rate than the full population analysis and a lower rate for values of $\zeta > 0.5$. As $\zeta$ approaches 1, the rate of rejection for all increases sharply, but the rate of correct subgroup identification decreases, having peaked around $\zeta = 0.40$ for most subgroup definitions. In terms of power for $\theta^*$ the full population analysis is constrained by design so that non-zero power occurs for $\zeta \geq \frac{2}{3}$ when $\theta_S = 1.5\theta^*$ or for $\zeta \geq 0.5$ when $\theta_S = 2\theta^*$. LR achieves non-zero power for the range of $\zeta$.

Figure 3.12 and Figure 3.13 present results for fixing $\theta = \{0.75\theta^*, \theta^*\}$, respectively. We again see a switch around 50%, where for $\zeta < 0.50$ we reject more often searching for subgroups and for $\zeta > 0.50$ we reject more often only analyzing the full population. For $\zeta = \{0.20, 0.30\}$ we observe the highest rates of decision for the correct subgroup: at or greater than 40% for most definitions. As

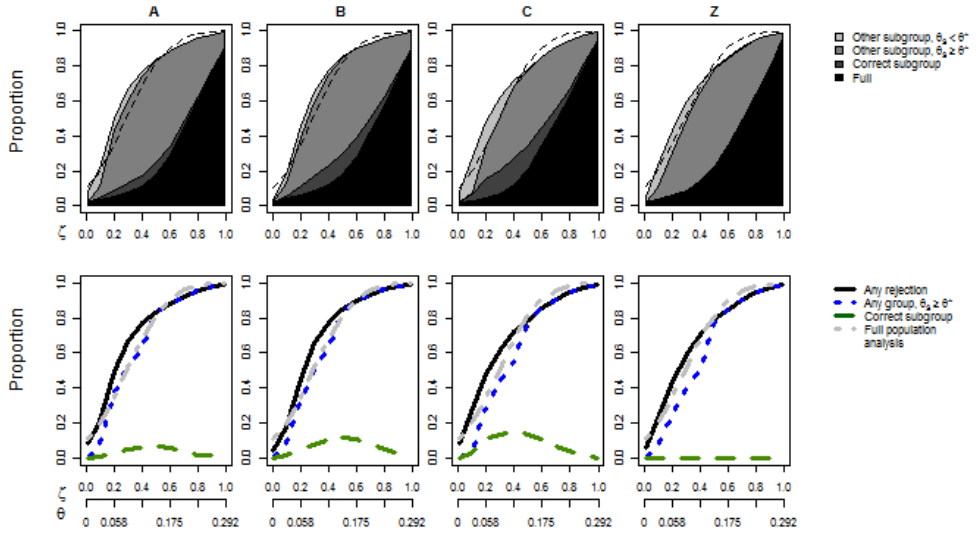



Figure 3.11: Performance of LR with $K = 4$, $n_{leaves} = 3$ when holding $\theta_S = 2\theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

$\zeta$ approaches 1 there is an increasing rate of decisions for the entire population while concurrently correct subgroup identification drops to 0% for $\zeta > 0.80$. This is driven by the decreasing $\theta^*$, as the truth underlying the fixed $\theta$ transitions from a stong heterogeneous effect to a moderate homogeneous effect. By design when $\theta = 0.75\theta^*$, power for $\theta^*$ is always 0% when only using a full population analysis, implying improved performance using LR. However when $\theta = \theta^*$ the power for the full population analysis is equal to its rejection rate, which dominates LR for most values of $\zeta$.

From these results we highlight three items. First, under this simulation set-up, $\zeta = 0.50$ is a transition point. With $\zeta < 0.50$ we see greater rejection rates for subgroup exploration than compared to the full population analysis. Second, the power to detect a group with $\theta_{\hat{S}} \geq \theta^*$ is sensitive to the marginal effects. Finally, subgroup effects of $1.5\theta^*$ and $2\theta^*$ may not be sufficient for reasonable rates of correct subgroup identification with n=200; with larger rates, such as $\frac{10}{3}\theta^*$ performance improves.

### *3.7 Varying the Baseline Rate ($\beta_0$)*

In simulations thus far we assume that the baseline response rate in the control arm is $\beta_0 = 0.30$. We use this baseline response rate in conjunction with the tuning parameter and covariate distribution to determine a critical value for the subgroup that ensures control of the experiment-wise type I

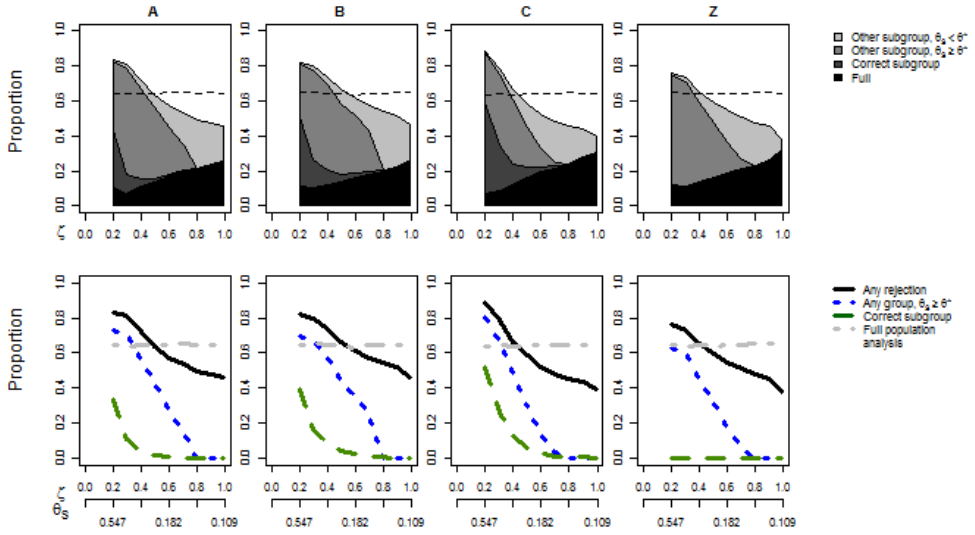



Figure 3.12: Performance of LR with $K = 4$, $n_{leaves} = 3$ when holding the marginal effect at $\theta = 0.75\theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

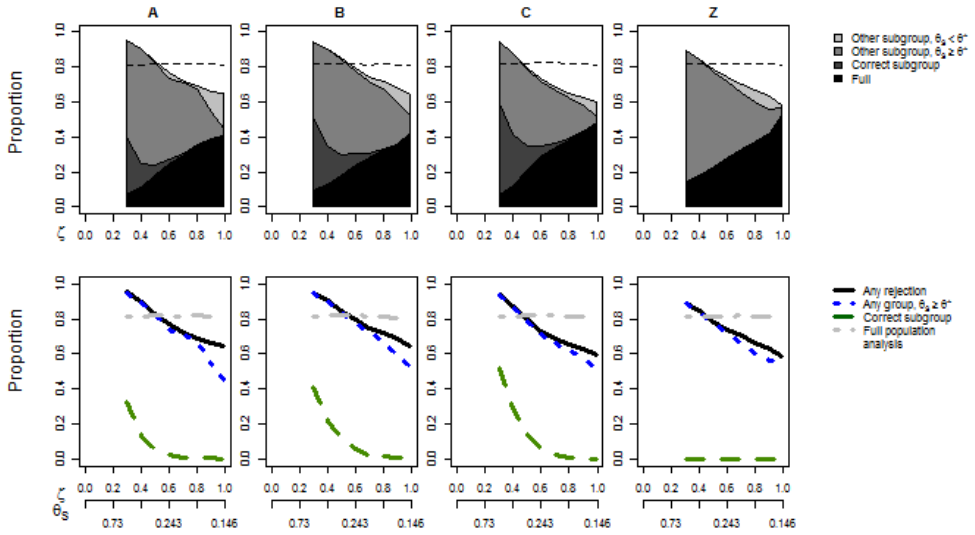



Figure 3.13: Performance of LR with $K = 4$, $n_{leaves} = 3$ when holding the marginal effect at $\theta = \theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

error rate, as described in Chapter 2. However, when planning a clinical trial uncertainty about the baseline rate is common. Given the mean-variance relationship for a binary response, we can expect different results for the alternatives depending on $\beta_0$. Table 3.8 presents results under the null and the three design alternatives for $\beta_0 = \{0.10, 0.30, 0.50\}$ when $K = 4$ and $n_{leaves} = 3$ with the same decision settings as in previous sections. The critical values use the correct baseline rates. As the variance when $\beta_0 = 0.10$ is less than the variance when $\beta = 0.30$ resulting in more power under each alternative, we considered holding power constant while simulating under different values of $\theta^*$ for each value of $\beta_0$. This approach, however, is inconsistent with the order of events when planning a clinical trial where an investigator determines sample size from an assumed baseline rate and treatment effect. So, these simulations represent the situation where an investigator planned on $\beta_0 = 0.30$ and $\theta^* = 0.146$ but found later that the control arm rate was lower or higher.

Under the null hypothesis, type I error is maintained at the nominal level with the majority of rejections for a proper subgroup for all values of $\beta_0$. With a homogeneous effect of $\theta = \theta^*$, both a full population only approach and LR have the highest rejection rates when $\beta_0 = 0.10$, where the binomial variance is the smallest out of the $\beta_0$ values we consider. There is minimal differences between $\beta_0 = 0.30$ and $\beta_0 = 0.50$. The lowest power loss for any rejection relative to a full population analysis is an average difference of 7.6% when $\beta_0 = 0.10$. The percentage of decisions for the entire study population relative to any rejection is greatest when $\beta_0 = 0.10$ with an average rate across the definitions of 72.8% compared to 56.5% when $\beta_0 = 0.30$.

When $\theta_S = 1.5\theta^*$, we again observe the best performance with $\beta_0 = 0.10$ for both any rejection (where the differences are driven by an increase in rejection of $H_F$) and power for $\theta^*$ (where the differences are slight). Rates of correct subgroup identification are around 1% or less as LR often includes more covariates than necessary in subgroup definitions. When $\theta_S = 2\theta^*$, LR performs similarly to the full population analysis with respect to any rejection or power for $\theta^*$, with the exception of a slight decrease with definition Z. The best performance is again with $\beta_0 = 0.10$, but with this baseline rate the frequency of rejection of $H_F$ is nearly double the other baseline rates. LR identifies the correct subgroup with definition C most often when $\beta_0 = 0.10$ (a rate of 22.5%) but with definition B it performs best with $\beta_0 = 0.50$ (a rate of 18.2%).

In summary, we see that the baseline rate directly influence power. Additionally, there is evidence that method calibration may depend on $\beta_0$ as we observe different decision ratios between the full population and a subgroup for the values of $\beta_0$.

| | | $\beta_0 = 0.10$ | | | | | $\beta_0 = 0.30$ | | | | | $\beta_0 = 0.50$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | Full | A | B | C | Z | Full | A | B | C | Z |
| $\theta = 0$ | All | 989 | 160 | 161 | 159 | 233 | 953 | 120 | 160 | 168 | 144 | 919 | 149 | 156 | 119 | 158 |
| for all | Any sg | – | 790 | 836 | 841 | 800 | – | 839 | 796 | 782 | 862 | – | 857 | 865 | 853 | 897 |
| | None | 9011 | 9050 | 9003 | 9000 | 8967 | 9047 | 9041 | 9044 | 9050 | 8994 | 9081 | 8994 | 8979 | 9028 | 8945 |
| | **Any rejection,** % | 9.9 | 9.5 | 10.0 | 10.0 | 10.3 | 9.5 | 9.6 | 9.6 | 9.5 | 10.1 | 9.2 | 10.1 | 10.2 | 9.7 | 10.6 |
| $\theta = \theta^*$ | All | 9347 | 6174 | 6234 | 6084 | 6516 | 8192 | 3824 | 3733 | 3909 | 3941 | 7916 | 3661 | 3635 | 3708 | 3499 |
| for all | Any sg | – | 2378 | 2276 | 2453 | 2234 | – | 3043 | 3015 | 2914 | 2943 | – | 3060 | 3060 | 2970 | 3249 |
| | None | 653 | 1448 | 1490 | 1463 | 1250 | 1808 | 3133 | 3252 | 3177 | 3116 | 2084 | 3279 | 3305 | 3322 | 3252 |
| | **Any rejection,** % | 93.5 | 85.5 | 85.1 | 85.4 | 87.5 | 81.9 | 68.7 | 67.5 | 68.2 | 68.8 | 79.2 | 67.2 | 67.0 | 66.8 | 67.5 |
| $\theta_S = 1.5\theta^*$ | All | 8148 | 2875 | 2843 | 3335 | 3728 | 6392 | 1443 | 1564 | 1782 | 1921 | 6048 | 1423 | 1447 | 1625 | 1604 |
| for 50% | Correct sg | – | 273 | 396 | 1133 | 0 | – | 180 | 403 | 501 | 0 | – | 217 | 625 | 427 | 0 |
| | Other sg | – | 4727 | 4755 | 3013 | 3867 | – | 4513 | 4662 | 3348 | 3849 | – | 4451 | 4329 | 3595 | 4059 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 4455 | 4493 | 2540 | 3311 | – | 3878 | 4031 | 2188 | 2804 | – | 3629 | 3555 | 1986 | 2604 |
| | None | 1852 | 2125 | 2006 | 2519 | 2405 | 3608 | 3864 | 3371 | 4369 | 4230 | 3952 | 3909 | 3599 | 4353 | 4337 |
| | **Any rejection,** % | 81.5 | 78.8 | 79.9 | 74.8 | 76.0 | 63.9 | 61.4 | 66.3 | 56.3 | 57.7 | 60.5 | 60.9 | 64.0 | 56.5 | 56.6 |
| | **Power for $\theta^*$,** % | 0.0 | 47.3 | 48.9 | 36.7 | 33.1 | 0.0 | 40.6 | 44.3 | 26.9 | 28.0 | 0.0 | 38.5 | 41.8 | 24.1 | 26.0 |
| $\theta_S = 2\theta^*$ | All | 9294 | 3619 | 3595 | 4183 | 4827 | 8177 | 1945 | 1860 | 2305 | 2493 | 8063 | 1783 | 1906 | 1861 | 2271 |
| for 50% | Correct sg | – | 684 | 887 | 2245 | 0 | – | 591 | 1154 | 1317 | 0 | – | 676 | 1819 | 1166 | 0 |
| | Other sg | – | 4993 | 4797 | 2697 | 4291 | – | 5738 | 5432 | 4318 | 5164 | – | 5866 | 4807 | 4935 | 5472 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 4987 | 4780 | 2672 | 4274 | – | 5726 | 5385 | 4245 | 5081 | – | 5858 | 4746 | 4883 | 5373 |
| | None | 706 | 704 | 721 | 875 | 882 | 1823 | 1726 | 1554 | 2060 | 2343 | 1937 | 1675 | 1468 | 2038 | 2257 |
| | **Any rejection,** % | 92.9 | 93.0 | 92.8 | 91.2 | 91.2 | 81.8 | 82.7 | 84.5 | 79.4 | 76.6 | 80.6 | 83.2 | 85.3 | 79.6 | 77.4 |
| | **Power for $\theta^*$,** % | 92.9 | 92.9 | 92.6 | 91.0 | 91.0 | 81.8 | 82.6 | 84.0 | 78.7 | 75.7 | 80.6 | 83.2 | 84.7 | 79.1 | 76.4 |

Table 3.8: LR for various $\beta_0$ with $\alpha_F = 0.025$, $K = 4$, and $n_{leaves} = 3$.

### *3.8 Prognostic Effects* ($\omega$)

Covariates that define a subgroup might be prognostic rather than predictive. We explore the influence of prognostic effects by adding a prognostic definition to A and C, as described in Table 3.9, with a prognostic effect prevalence of 50%. In A, the definitions for the predictive and prognostic subgroups are independent whereas with C they are correlated with $X_1 = T$ required for both. The previous simulations assumed homogeneity for the control group with $\beta_0 = 0.30$ and implictly $\omega = 0.0$. For comparability, we hold the average baseline rate at 0.30 and use prognostic effects of $\omega = \{0.05, 0.10, 0.20\}$ with corresponding $\beta_0 = \{0.275, 0.25, 0.20\}$. With *LR-A* the tuning parameter for the prognostic subgroup fit was set to $n_{leaves} = 3$.

| | Predictive | Prognostic | $Pr(X_1 = 1)$ | $Pr(X_2 = 1)$ | $Pr(X_3 = 1)$ | $Pr(X_4 = 1)$ |
|---|---|---|---|---|---|---|
| A | $X_1$ | $X_2$ | 0.50 | 0.50 | 0.50 | 0.50 |
| C | $X_1 \wedge X_2$ | $X_1 \wedge X_3$ | 0.71 | 0.71 | 0.71 | 0.50 |

Table 3.9: Boolean expressions for subgroup definitions A, B, C and Z. The probabilities for each covariate correspond to an overall subgroup prevalence of 0.5.

Table 3.10 has the results with the other simulation parameter settings as before. The critical values assume $\beta_0 = 0.30$ with no prognostic effects. Under the null, the rate of rejection exceeds the nominal level of 10% to 36.9% for A and 22.7% for C under the greatest prognostic effects. The type I error inflation translates to increased rejection rates under all of the alternatives. When considering correct subgroup identification, performance improves with an increasing prognostic effect with definition C and worsens with definition A, suggesting effects that are both prognostic and predictive improve correct subgroup identification, although without proper type I error calibration the results under the alternatives cannot be believed.

### *3.9 Continuous Endpoint*

We next explore the use of LR with a continuous endpoint. Section 3.1 introduces a data generation mechanism from a normal model: here we use the model with variance $\epsilon = 1$. A simple sample size calculation yields that a one-sided test of level $\alpha = 0.10$ with $n = 200$ has 80% power to detect an effect of $\theta^* = 0.300$.

Table 3.11 contains the results from the null, full population and two subgroup alternative hypotheses. Under homogeneous effects with $\theta = \theta^*$ we observe a power loss using LR from 9.4 to 14.9% compared to a full population analysis. Decisions are approximately split between the full

| Truth | Decision | $\beta_0 = 0.30, \omega = 0.00$ Full | A | C | $\beta_0 = 0.275, \omega = 0.05$ Full | A | C | $\beta_0 = 0.25, \omega = 0.10$ Full | A | C | $\beta_0 = 0.20, \omega = 0.20$ Full | A | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0$ | All | 953 | 120 | 168 | 938 | 94 | 141 | 956 | 101 | 118 | 990 | 103 | 150 |
| for all | Any sg | – | 839 | 782 | – | 1299 | 874 | – | 2085 | 1453 | – | 3586 | 2117 |
| | None | 9047 | 9041 | 9050 | 9062 | 8607 | 8985 | 9044 | 7814 | 8429 | 9010 | 6311 | 7733 |
| | **Any rejection,** % | 9.5 | 9.6 | 9.5 | 9.4 | 13.9 | 10.2 | 9.6 | 21.9 | 15.7 | 9.9 | 36.9 | 22.7 |
| $\theta = \theta^*$ | All | 8192 | 3824 | 3909 | 8191 | 3423 | 3725 | 8196 | 3061 | 3466 | 8164 | 2966 | 3745 |
| for all | Any sg | – | 3043 | 2914 | – | 3677 | 3116 | – | 4392 | 3672 | – | 4644 | 3372 |
| | None | 1808 | 3133 | 3177 | 1809 | 2900 | 3159 | 1804 | 2547 | 2862 | 1836 | 2390 | 2883 |
| | **Any rejection,** % | 81.9 | 68.7 | 68.2 | 81.9 | 71.0 | 68.4 | 82.0 | 74.5 | 71.4 | 81.6 | 76.1 | 71.2 |
| $\theta_S = 1.5\theta^*$ | All | 6392 | 1443 | 1782 | 6449 | 1327 | 1441 | 6436 | 1227 | 1123 | 6368 | 1213 | 1299 |
| for 50% | Correct sg | – | 180 | 501 | – | 152 | 698 | – | 121 | 646 | – | 106 | 1037 |
| | Other sg | – | 4513 | 3348 | – | 5015 | 4214 | – | 5421 | 4917 | – | 5680 | 4431 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 3878 | 2188 | – | 4146 | 3362 | – | 4139 | 4054 | – | 4075 | 3941 |
| | None | 3608 | 3864 | 4369 | 3551 | 3506 | 3647 | 3564 | 3231 | 3314 | 3632 | 3001 | 3233 |
| | **Any rejection,** % | 63.9 | 61.4 | 56.3 | 64.5 | 64.9 | 63.5 | 64.4 | 67.7 | 66.9 | 63.7 | 70.0 | 67.7 |
| | **Power for $\theta^*$,** % | 0.0 | 40.6 | 26.9 | 0.0 | 43.0 | 40.6 | 0.0 | 42.6 | 47.0 | 0.0 | 41.8 | 49.8 |
| $\theta_S = 2\theta^*$ | All | 8177 | 1945 | 2305 | 8108 | 1743 | 1695 | 8134 | 1549 | 1493 | 8164 | 1471 | 1569 |
| for 50% | Correct sg | – | 591 | 1317 | – | 507 | 1555 | – | 382 | 1489 | – | 388 | 2152 |
| | Other sg | – | 5738 | 4318 | – | 6038 | 5084 | – | 6538 | 5542 | – | 6698 | 4833 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 5726 | 4245 | – | 6021 | 5053 | – | 6513 | 5500 | – | 6681 | 4819 |
| | None | 1823 | 1726 | 2060 | 1892 | 1712 | 1666 | 1866 | 1531 | 1476 | 1836 | 1443 | 1446 |
| | **Any rejection,** % | 81.8 | 82.7 | 79.4 | 81.1 | 82.9 | 83.3 | 81.3 | 84.7 | 85.2 | 81.6 | 85.6 | 85.5 |
| | **Power for $\theta^*$,** % | 81.7 | 82.6 | 78.7 | 81.1 | 82.7 | 83.0 | 81.3 | 84.4 | 84.8 | 81.6 | 85.4 | 85.4 |

Table 3.10: Presence of prognostic effects with $\alpha_F = 0.025$, $K = 4$, $n_{leaves} = 3$.

and subgroup alternatives when $K = 1$ but increase in $K$ in favor of the full. Under heterogeneous effects the highest correct subgroup identification occurs with definition A at 66.8% when $\theta_S = 1.5\theta^*$ and 86.0% when $\theta_S = 2\theta^*$. For $K = \{4, 8\}$ correct subgroup identification occurs less than 5% of the time when $\theta_S = 1.5\theta^*$ and less than 15% of the time when $\theta_S = 2\theta^*$. Overall, in comparing these results to those from a binary endpoint, we conclude performance is quite similar.

### *3.10 Summary*

Logic Regression shows some potential promise to identify predictive subgroups. Our findings include:

- Rates of correct subgroup identification are dependent upon agreement between the $n_{leaves}$ tuning parameters and the number of covariates in the true subgroup definition. Identification decreases quickly when they do not agree.
- The value of $\alpha_F$ requires some consideration as selecting a value too high will result in always selecting the full population and too low will always select a subgroup. We implement $\alpha_F = 0.025$ in this chapter to provide some balance between these two extremes.
- LR does not identify subgroups with effects of $\theta_S = \{1.5\theta^*, 2\theta^*\}$ often as might be required by an investigator before adopting this method. Larger effects or sample sizes are required for performance that might be accepted by an investigator.
- There exists a relationship between the subgroup definition, $K$ and $n_{leaves}$, with our results being too optimistic as we preferably selected the correct $K$ variables for definitions A, B, and C. By assuming we are uncertain about the correct subgroup definition we must also be uncertain about which variables matter.
- We generally have lower power under homogeneous effects, which reflects the cost of searching for a subgroup. However, it is possible to gain power with heterogeneous effects when the exploration is limited.
- Performance is similar for a continuous endpoint.
- $\zeta = 0.50$ functions as a change point in these simulations. Less than this value and we gain power relative to a full population analysis, greater we lose power.
- Type I error is greatly inflated when there are prognostic effects explained by the $K$ covariates. This suggests that when implementing LR only covariates that are not prognostic can be used

| Truth | Decision | Full | $K=1, n_{leaves}=1$ A | B | C | Z | $K=2, n_{leaves}=2$ A | B | C | Z | $K=4, n_{leaves}=3$ A | B | C | Z | $K=8, n_{leaves}=3$ A | B | C | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0$ | All | 1030 | 107 | 146 | 116 | 146 | 125 | 143 | 112 | 144 | 144 | 136 | 133 | 124 | 150 | 153 | 159 | 162 |
| for all | Any sg | – | 838 | 889 | 888 | 928 | 853 | 861 | 893 | 874 | 828 | 928 | 903 | 842 | 850 | 846 | 841 | 838 |
| | None | 8970 | 9055 | 8965 | 8996 | 8926 | 9022 | 8996 | 8995 | 8982 | 9028 | 8936 | 8964 | 9034 | 9000 | 9001 | 9000 | 9000 |
| | **Any rejection,** % | 10.3 | 9.4 | 10.3 | 10.0 | 10.7 | 9.8 | 10.0 | 10.1 | 10.2 | 9.7 | 10.6 | 10.4 | 9.7 | 10.0 | 10.0 | 10.0 | 10.0 |
| $\theta = \theta^*$ | All | 8038 | 3388 | 3356 | 3285 | 3541 | 3460 | 3248 | 3179 | 3465 | 3383 | 3518 | 3530 | 3292 | 3465 | 3600 | 3608 | 3685 |
| for all | Any sg | – | 3711 | 3596 | 3689 | 3314 | 3406 | 3566 | 3699 | 3227 | 3474 | 3026 | 3102 | 3359 | 3081 | 2917 | 2863 | 2999 |
| | None | 1962 | 2901 | 3048 | 3026 | 3145 | 3134 | 3186 | 3122 | 3308 | 3143 | 3456 | 3368 | 3349 | 3454 | 3483 | 3529 | 3316 |
| | **Any rejection,** % | 80.4 | 71.0 | 69.5 | 69.7 | 68.5 | 68.7 | 68.1 | 68.8 | 66.9 | 68.6 | 65.4 | 66.3 | 66.5 | 65.5 | 65.2 | 64.7 | 66.8 |
| $\theta = 1.5\theta^*$ | All | 6134 | 741 | 1665 | 802 | 1674 | 1199 | 978 | 773 | 1685 | 1261 | 1245 | 1603 | 1757 | 1591 | 1589 | 1958 | 1963 |
| for 50% | Correct sg | – | 6682 | 0 | 0 | 0 | 3521 | 3932 | 3920 | 0 | 160 | 433 | 444 | 0 | 2 | 79 | 67 | 0 |
| | Other sg | – | 20 | 4585 | 5342 | 3775 | 1831 | 2181 | 2045 | 4044 | 4604 | 4482 | 3182 | 3989 | 3835 | 4080 | 2693 | 3182 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 4318 | 5216 | 3360 | 1724 | 2094 | 1346 | 3052 | 3847 | 3884 | 2200 | 2771 | 2848 | 3022 | 1651 | 1802 |
| | None | 3866 | 2557 | 3750 | 3856 | 4551 | 3449 | 2909 | 3262 | 4271 | 3975 | 3840 | 4771 | 4254 | 4572 | 4252 | 5282 | 4855 |
| | **Any rejection,** % | 61.3 | 74.4 | 62.5 | 61.4 | 54.5 | 65.5 | 70.9 | 67.4 | 57.3 | 60.2 | 61.6 | 52.3 | 57.5 | 54.3 | 57.5 | 49.9 | 51.5 |
| | **Power for $\theta^*$,** % | 0.0 | 66.8 | 43.2 | 52.2 | 33.6 | 52.5 | 60.3 | 52.7 | 30.5 | 40.1 | 43.2 | 26.4 | 27.7 | 28.5 | 31.0 | 17.2 | 18.0 |
| $\theta = 2\theta^*$ | All | 7864 | 568 | 2299 | 926 | 2327 | 1163 | 924 | 723 | 2172 | 1533 | 1454 | 1950 | 1985 | 1919 | 1949 | 2595 | 2568 |
| for 50% | Correct sg | – | 8597 | 0 | 0 | 0 | 5903 | 6313 | 6234 | 0 | 544 | 1210 | 1172 | 0 | 15 | 292 | 281 | 0 |
| | Other sg | – | 3 | 5840 | 7206 | 5074 | 1458 | 1756 | 1809 | 5407 | 5957 | 5424 | 4321 | 5475 | 5580 | 5430 | 3956 | 4382 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 5727 | 7164 | 4709 | 1455 | 1735 | 1806 | 5300 | 5934 | 5360 | 4256 | 5361 | 5574 | 5408 | 3917 | 4244 |
| | None | 2136 | 832 | 1861 | 1868 | 2599 | 1476 | 1007 | 1234 | 2421 | 1966 | 1912 | 2557 | 2540 | 2486 | 2329 | 3168 | 3050 |
| | **Any rejection,** % | 78.6 | 91.7 | 81.4 | 81.3 | 74.0 | 85.2 | 89.9 | 87.7 | 75.8 | 80.3 | 80.9 | 74.4 | 74.6 | 75.1 | 76.7 | 68.3 | 69.5 |
| | **Power for $\theta^*$,** % | 78.6 | 91.7 | 80.3 | 80.9 | 70.4 | 85.2 | 89.7 | 87.6 | 74.7 | 80.1 | 80.2 | 73.8 | 73.5 | 75.1 | 76.5 | 67.9 | 68.1 |

Table 3.11: Use of LR with a continuous endpoint, n=200, $\alpha = 0.10$, $\alpha_F = 0.025$, $\theta^* = 0.30$.

to avoid inflation of the type I error. Additional work is needed in this area to determine if statistical adjustment can correct this inflation.

Overall, our simulations and results provide a general overview of the use of LR for our specific setting. There are many additional settings that could be explored, related to specific drugs and under consideration.

# Chapter 4

# SIMULATION STUDIES AND PERFORMANCE EVALUATION FOR SHAPES

In this chapter we explore the performance of SHAPES. In general, we take a similar approach as in the previous chapter and begin by considering the same $\alpha$-allocation, decision rules and utility functions used with LR to identify a reasonable trade-off between subgroup and full population selection. We introduce a variation on a heat map or quilt plot specific to the identification of subgroups. We then proceed to consider performance when varying specific aspects of the data generation mechanism. Throughout, when relevant, we highlight differences with LR and conclude with a summary of the performance of SHAPES.

## *4.1 Decision Rules and Utility Functions for $K = 4$ and $\ell = 3$*

We apply SHAPES to the same scenarios as described in Section 3.1. When determining the critical values for each subgroup depth we apportion $(\alpha - \alpha_F)$ equally to each depth. Figure 4.1 displays results with the three utility functions when $\alpha_F = 0.025$ and the subgroup definition is A. Similar to LR, under both heterogeneous and homogeneous effects *prefer subgroup* and *prefer all* most frequently select the corresponding preferred group. The *minimum p-value* utility performs similarly to *prefer subgroup* where even with a homogeneous effect of $\theta = 2\theta^*$ rejection of $H_F$ is less than 30% compared to near 100% with *prefer all*.

Increasing to $\alpha_F = 0.05$ (Figure 4.2), there is little appreciable difference in performance for *prefer subgroup* and *prefer all*. The *minimum p-value* utility shows better discrimination between the full population and the subgroup, with rejection greater than 60% under homogeneous effects when $\theta = 2\theta^*$. Under heterogeneous effects when $\theta_S = 3\theta^*$ correct subgroup identification occurs a majority of the time with the full population selected less than 20% of the time for the range of subgroup effects. Increasing to $\alpha_F = 0.075$ (Figure 4.3), we rarely select any subgroup with the *prefer all* rule. With the *prefer subgroup* rule under both homogeneous and heterogeneous effects we select the full population less than 20% of the time, which approaches 0% as the effect increases. Using *minimum p-value* we approach 90% power for homogeneous effects and for heterogeneous effects we approach a 40% rejection rate of $H_F$.

The nine combinations formed by the three values of $\alpha_F$ and the three utility functions result in

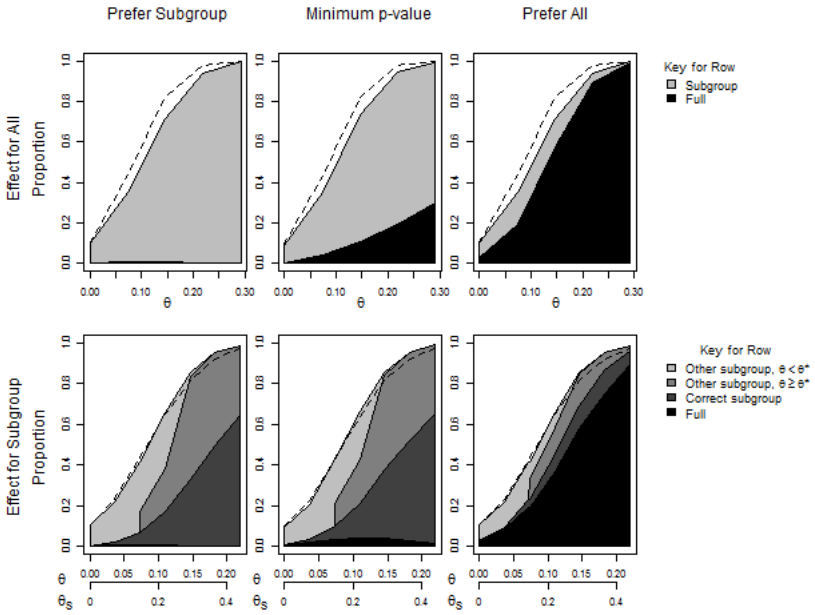



Figure 4.1: Performance of SHAPES with $\alpha_F = 0.025$ and $\alpha = 0.10$ comparing three utility functions (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, $K = 4$, and $\ell = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

a range of different selection rates of subgroup and full population. We take a similar approach to LR and seek an approach that provides reasonable discrimination and correct performance between homogeneous and heterogeneous effects; which is provided by *minimum p-value* and $\alpha_F = 0.05$. With LR we selected the same rule but $\alpha_F = 0.025$. The difference here is likely due to two competing factors: 1) SHAPES selects a subgroup based on the p-value rather than the residual sum of squares, which ensures smaller subgroup p-values under the null and hence a greater $\alpha_F$ for similar calibration compared to LR and 2) the decrease in the number of subgroup definitions under consideration with SHAPES helps to limit the competition of random high subgroups, i.e. subgroups perform well in terms of the LR scoring function by chance rather than a true effect.

### *4.2 Subgroup Definitions and Quilt Plots*

In Figure 4.4 are the results using $\alpha_F = 0.05$ and the *minimum p-value* utility function for the four subgroup definitions of A, B, C, and Z. With homogeneous effects the performance of these metrics is similar across the subtle differences in covariate distributions. There is always some power loss compared to an analysis that only tests in the full population. Under heterogeneous effects, the

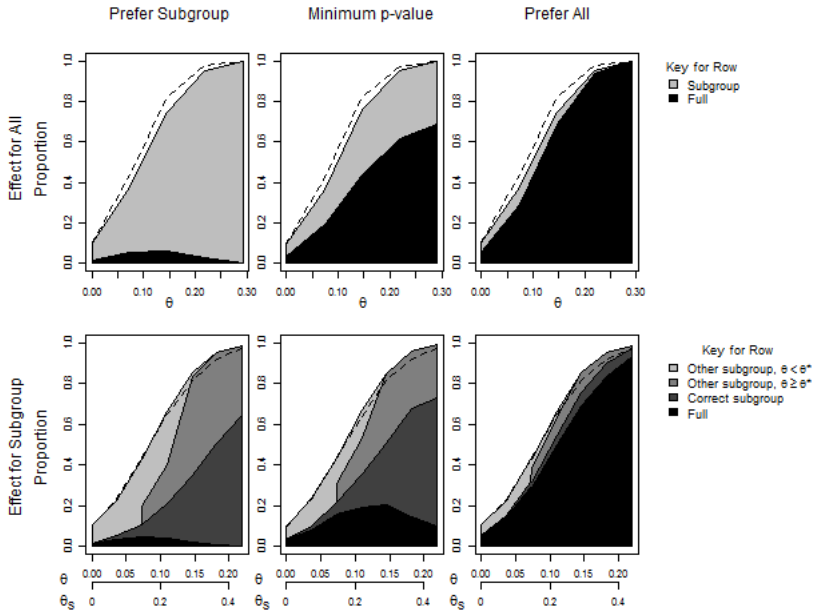



Figure 4.2: Performance of SHAPES with $\alpha_F = 0.05$ and $\alpha = 0.10$ comparing three utility functions (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, $K = 4$, and $\ell = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

overall rates of rejection are similar across the definitions, represented by the the total non-white area. For the range of $\theta_S$ there is an increase in power compared to only testing in the full population. Differences arise in the composition of the decisions for each definition. The highest rate of correct subgroup identification occurs with definition A and the lowest rate with Z, constrained to be 0 by design. Identification of a subgroup with an effect greater than or equal to $\theta^*$ increases rapidly for $\theta_S \geq \theta^*$ and when $\theta_S \geq 2\theta^*$ nearly every identified subgroup has a true effect greater than $\theta^*$.

We now describe a plot that is a variation of a heat map or quilt plot; examples can be seen in Figures 4.5 and 4.6. The purpose of the plot is to provide a summary of the prevalence and effects of the subgroups and their frequency of selection under null, heterogeneous and homogeneous effects. The plot features four horizontal regions separated by the thick black lines. The upper two regions represent homogeneous effects at two different values of $\theta^*$. Below these regions is the largest region that corresponds to a specific heterogeneous effect, covariate distribution and subgroup definition. The lowest region corresponds to the null. Within each region the location of the correct group decision is noted with a black circle. Each region of the plot is divided into four subregions based on the prevalence of the subgroup with cut-offs of 0 (representing failing to reject the null), 20 and 80%.

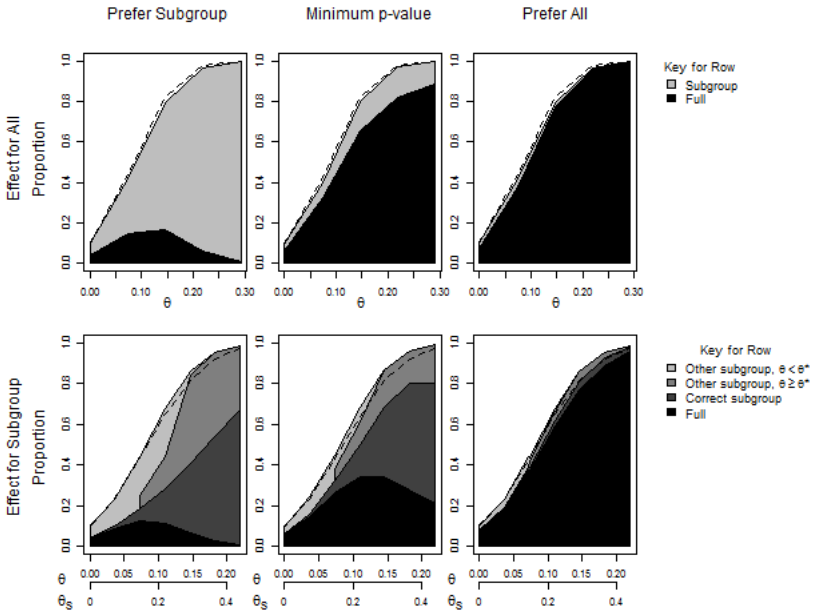



Figure 4.3: Performance of SHAPES with $\alpha_F = 0.075$ and $\alpha = 0.10$ comparing three utility functions (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, subgroup definition A, $K = 4$, and $\ell = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

For the heterogeneous case the regions are further split by the true effect in the subgroup with a cut-off of $\theta^*$. An integer on the plot indicates the number of unique subgroup definitions of the same prevalence and effect. For example, a '4' indicates there are four unique subgroup definitions with the same prevalence and effect. Colors signify the frequency of rejection from 10,000 repetitions, with the color spectrum ranging from white (for exactly 0), to yellow, green and blue. A small colored square around each integer indicates the total number of rejections for subgroups with the same size and effect and the color of each subregion shows the frequency of rejection for any subgroup within the subregion. The frequency is also reported.

Many observations can be made from these plots. Under heterogeneous effects, the bivariate distribution of $(\theta_S, \zeta)$ is different for the various definitions (and the inherent covariates). Relatedly the proximity of neighbors is different for the various definitions and the number of candidate subgroups in each subregion varies. The proximity of subgroups to one another represents some measure of relationship or correlation. For definitions A, B, and C the rejection rates for the subregions with an effect greater than or equal to $\theta^*$ and prevalence greater than 0.20 are similar. The number of proper subgroups of the true subgroup is also different, indicated by the sum of the subgroups directly to

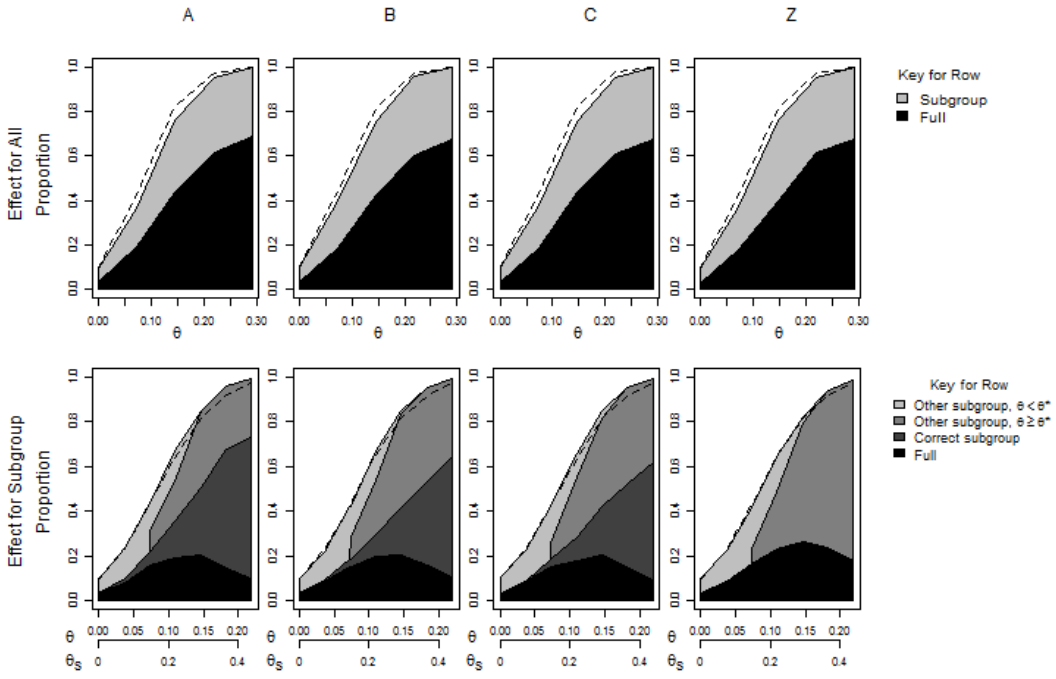



Figure 4.4: Performance of SHAPES with $\alpha_F = 0.05$ and $\alpha = 0.10$, *minimum p-value* utility function comparing four subgroup definitions (the columns) under full population and subgroup effects (the rows) for 10,000 replications with n=200, $K = 4$, and $n_{leaves} = 3$. The dashed line represents the frequency of rejection when testing only the full population at level $\alpha$.

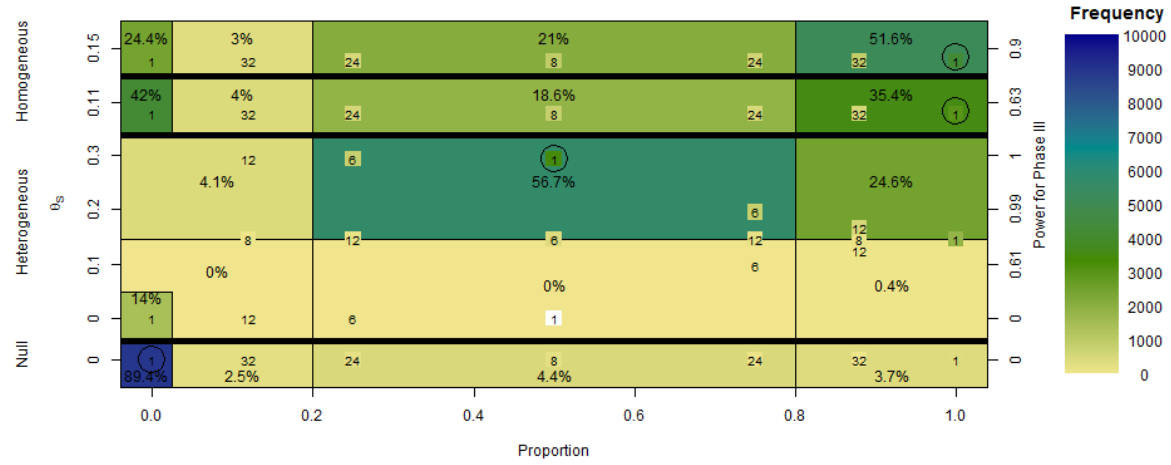
the left of the black circle in the middle subregion. Overall, the selection of subgroups with $\theta_S < \theta^*$ is relatively infrequent.

Under homogeneous effects, rejection tends to be for groups with higher prevalences and most often for the full population. The remaining rejections are spread among the range of possible subgroups with similar rejection frequencies for each possible subgroup and subregion. We see a similar spread of rejections under the null hypothesis, as well.

### *4.3 Varying $K$ and $\ell$*

We now evaluate SHAPES with the *minimum p-value* utility function, $\alpha_F = 0.05$, $\alpha = 0.10$, and, when applicable, $\zeta = 0.5$ to the null and alternative scenarios while varying $K$ and $\ell$. We use $(K, \ell) = \{(1,1), (2,2), (4,3), (8,3)\}$. Table 4.1 provides a summary of power, correct group identification and rejection rates for a subgroup with $\theta_{\hat{S}} \geq \theta^*$, i.e. where the true effect in the selected group is greater than or equal to $\theta^*$. Table 4.2 details the decisions with 10,000 replications for each scenario. A

Definition A




Definition B

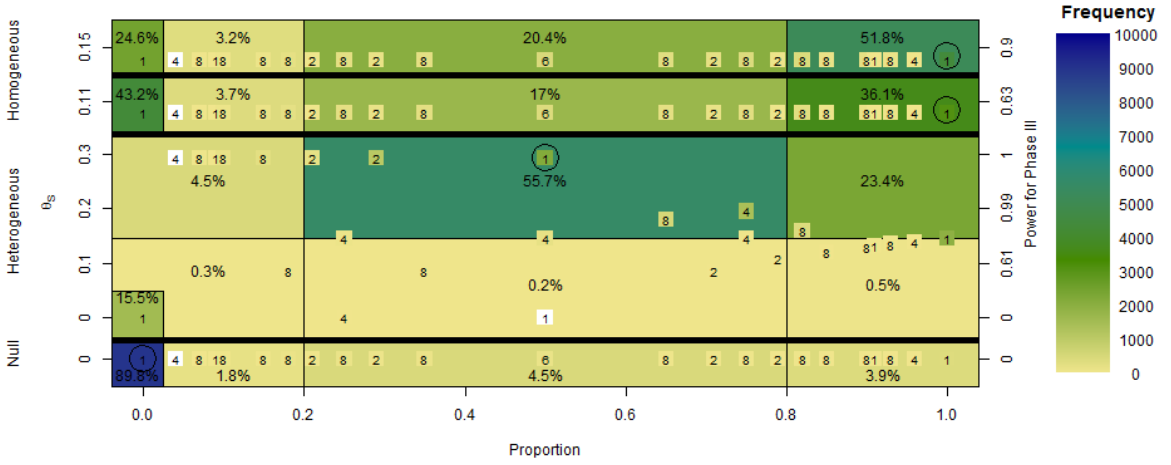
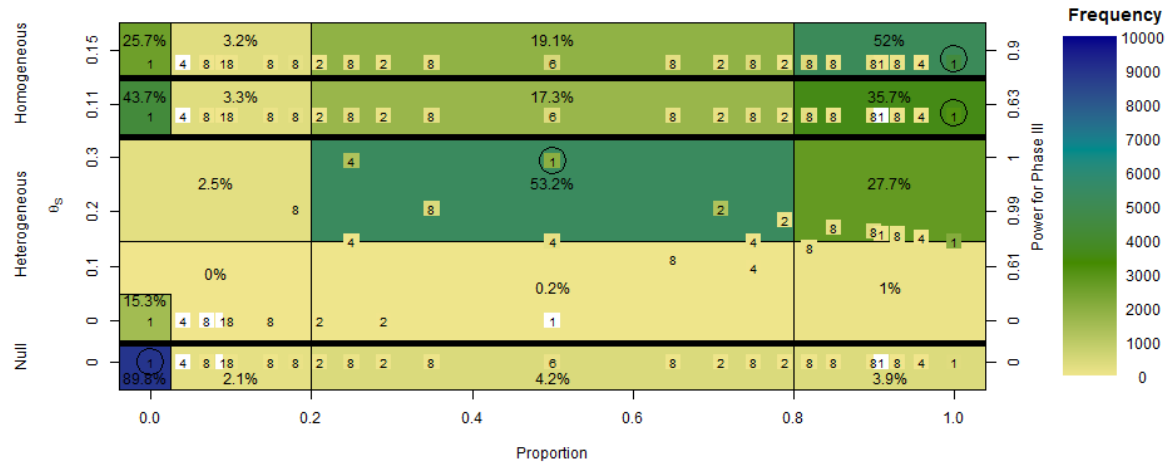



Figure 4.5: Quilt plot for definitions A and B using SHAPES with $\ell = 3$, $K = 4$, $\alpha_F = 0.05$.

Definition C




Definition Z

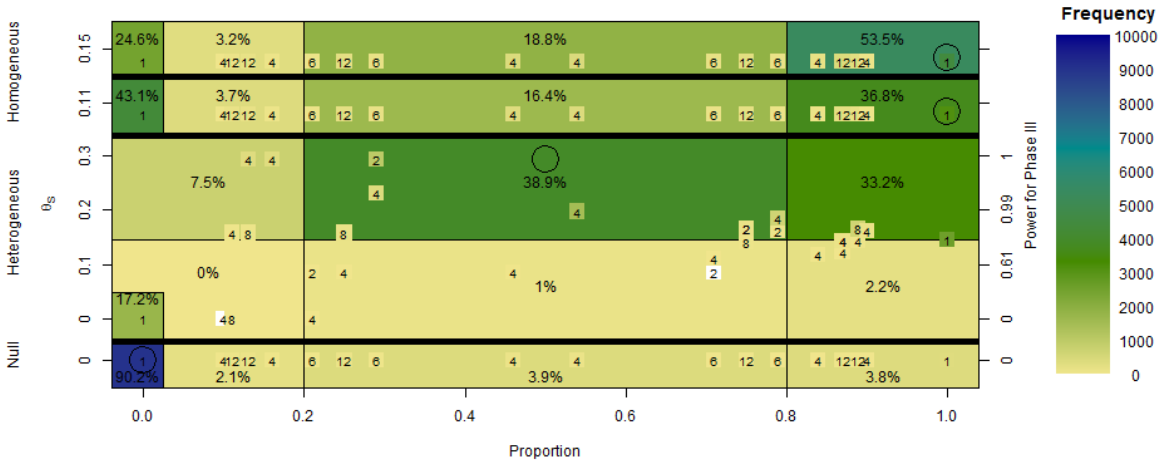



Figure 4.6: Quilt plot for definitions C and Z using SHAPES with $\ell = 3$, $K = 4$, $\alpha_F = 0.05$.

description of the 'Decisions' column is available in Section 3.2. In both tables the 'Full' column reports the results from a full population analysis where $\alpha_F = \alpha = 0.10$. We describe results for the null and specified alternative scenarios and for reference include additional values of $\theta = 1.5\theta^*$ and $\theta_S = \theta^*$ in the tables.

**Null:** Simulating under the null, rejection occurs between 9.1 and 11.0% of the time. When we reject, the ratio of decisions for the full population to any proper subgroup is approximately 3:7.

**Homogeneous effects:** With $\theta = \theta^*$, the rejection rate ranges from 74.1 to 77.1% under the various definitions and values of $K$, with the highest rates for definition A. These rates are always lower than the full population only rate of 81.5%; corresponding to a power loss from 4.4 to 7.4%. With $K = 1$ on average we select the full population 52.0% of the time and a proper subgroup 22.8% of the time. For $K = 8$ this ratio is closer to 4:3. If the effect is $\theta = 1.5\theta^*$, when exploring for subgroups we reject between 94.9 to 95.9% of the time compared to 97.1% for the full population. The power loss decreases for this alternative: between 1.2 and 2.2%.

**Heterogeneous effects:** With $\theta_S = 1.5\theta^*$, we reject between 60.0 and 73.8% of the trials with the highest rate when $K = 1$ under definition A (with a 58.0% rate of correct subgroup identification) and the lowest rate when $K = 8$ under definition Z (with a 0.0% rate of correct subgroup identification). In comparison, a full population analysis rejects with 63.7% frequency. The frequency of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ ranges from 19.7 to 58.0% compared to a rate of 0.0% for the full population as the marginal effect is $0.75\theta^*$. As evident in the 'Correct sg' rows of Table 4.2, as we consider more non-informative covariates we generally decrease our ability to identify the correct subgroup. For example under scenario A with $K = 1$ we have 58.0% correct identification which decreases to 15.0% with $K = 8$.

When $\theta_S = 2\theta^*$, the frequency of any rejection is between 77.1 and 91.3% compared to the full population analysis rate of 81.3%. Similarly, the rate of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ ranges between 72.7 and 91.3%. With $K = \{1, 2\}$, the majority of these rejections are for correct subgroups when the definition is recoverable but with $K = \{4, 8\}$ these rejections are for some other subgroups: composed of a either a proper subset of the correct subgroup or a mixture from the correct subgroup and its complement.

We next consider fixing $\ell = 1$ while varying $K$. Presented in Table 4.4 are results with $\alpha_F = 0.025$, $\alpha = 0.10$ and the *minimum p-value* rule for $K = \{1, 2, 4, 8\}$. Rejection rates are reported in Table 4.3.

**Null**: Under the null distribution, the overall empirical type I error rate ranges from 9.5 to 10.8%. We reject between 2.4 and 3.4% of the time for the full population and between 6.3 and 8.2%

| | | | $K=1,\ell=1$ | | | | $K=2,\ell=2$ | | | | $K=4,\ell=3$ | | | | $K=8,\ell=3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| Frequency of any rejection | | | | | | | | | | | | | | | | | | |
| All | $\theta = 0$ | 10.1 | 10.8 | 10.4 | 10.5 | 9.5 | 11.0 | 10.0 | 9.5 | 9.5 | 9.3 | 10.0 | 10.3 | 9.6 | 10.2 | 9.9 | 10.3 | 9.1 |
| | $\theta = \theta^*$ | 81.5 | 74.8 | 74.1 | 75.3 | 75.1 | 75.9 | 74.4 | 75.0 | 74.3 | 76.1 | 75.5 | 75.2 | 75.9 | 77.1 | 75.7 | 76.2 | 75.0 |
| | $\theta = 1.5\theta^*$ | 97.1 | 94.9 | 95.3 | 95.5 | 95.3 | 95.4 | 95.3 | 95.5 | 95.1 | 95.2 | 95.6 | 95.2 | 95.4 | 95.9 | 95.6 | 95.2 | 95.5 |
| Subgroup | $\theta_S = \theta^*$ | 42.7 | 48.0 | 43.0 | 40.9 | 38.5 | 45.7 | 46.5 | 42.3 | 40.0 | 43.2 | 42.1 | 42.2 | 41.3 | 42.5 | 40.6 | 43.4 | 41.2 |
| | $\theta_S = 1.5\theta^*$ | 63.7 | 73.8 | 65.0 | 62.7 | 60.0 | 71.2 | 71.5 | 67.5 | 62.0 | 67.1 | 67.1 | 66.0 | 64.2 | 67.1 | 63.0 | 66.3 | 61.8 |
| | $\theta_S = 2\theta^*$ | 81.3 | 91.3 | 83.8 | 82.6 | 77.1 | 88.8 | 88.9 | 86.8 | 79.9 | 85.0 | 84.8 | 85.2 | 82.2 | 85.4 | 82.0 | 85.0 | 81.0 |
| Frequency of rejection for a group with $\theta_{\hat{S}} \geq \theta^*$ | | | | | | | | | | | | | | | | | | |
| Subgroup | $\theta_S = \theta^*$ | 0.0 | 33.9 | 20.2 | 0.0 | 0.0 | 21.5 | 28.9 | 14.5 | 6.5 | 14.9 | 13.1 | 10.5 | 6.3 | 11.1 | 7.2 | 5.0 | 3.0 |
| | $\theta_S = 1.5\theta^*$ | 0.0 | 58.0 | 32.1 | 41.2 | 25.9 | 51.1 | 50.0 | 42.8 | 23.9 | 34.0 | 34.2 | 37.7 | 25.4 | 27.2 | 19.7 | 30.6 | 21.0 |
| | $\theta_S = 2\theta^*$ | 81.3 | 91.3 | 82.3 | 82.3 | 74.8 | 88.4 | 87.8 | 86.7 | 78.5 | 84.7 | 83.7 | 81.6 | 79.6 | 85.3 | 81.4 | 72.7 | 73.2 |
| Frequency of correct decision | | | | | | | | | | | | | | | | | | |
| All | $\theta = 0$ | 89.9 | 89.2 | 89.6 | 89.5 | 90.5 | 89.0 | 90.0 | 90.5 | 90.5 | 90.7 | 90.0 | 89.7 | 90.4 | 89.8 | 90.1 | 89.7 | 90.9 |
| | $\theta = \theta^*$ | 81.5 | 50.8 | 51.9 | 53.5 | 52.0 | 45.5 | 46.1 | 46.9 | 46.3 | 43.8 | 42.7 | 43.6 | 39.8 | 42.4 | 44.0 | 41.8 | 46.1 |
| | $\theta = 1.5\theta^*$ | 97.1 | 79.9 | 76.1 | 76.1 | 77.7 | 67.6 | 69.6 | 68.3 | 70.4 | 61.5 | 60.5 | 60.8 | 61.7 | 58.1 | 60.2 | 57.8 | 61.1 |
| Subgroup | $\theta_S = \theta^*$ | 0.0 | 33.9 | 0.0 | 0.0 | 0.0 | 12.6 | 13.1 | 14.5 | 0.0 | 5.5 | 3.1 | 3.1 | 0.0 | 4.9 | 1.3 | 1.2 | 0.0 |
| | $\theta_S = 1.5\theta^*$ | 0.0 | 58.0 | 0.0 | 0.0 | 0.0 | 29.3 | 29.9 | 30.3 | 0.0 | 16.4 | 10.1 | 10.2 | 0.0 | 15.0 | 4.7 | 4.3 | 0.0 |
| | $\theta_S = 2\theta^*$ | 0.0 | 77.7 | 0.0 | 0.0 | 0.0 | 46.8 | 49.5 | 47.7 | 0.0 | 30.2 | 20.9 | 21.9 | 0.0 | 29.6 | 12.7 | 11.8 | 0.0 |

Table 4.1: Type I error and power for SHAPES with n=200, $\alpha = 0.10$, $\alpha_F = 0.05$, $\theta^* = 0.146$.

| | | | $K=1,\ell=1$ | | | | $K=2,\ell=2$ | | | | $K=4,\ell=3$ | | | | $K=8,\ell=3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 1006 | 334 | 321 | 335 | 322 | 317 | 340 | 320 | 314 | 325 | 346 | 338 | 304 | 294 | 329 | 309 | 356 |
| for all | Any sg | – | 747 | 717 | 718 | 632 | 785 | 660 | 631 | 640 | 604 | 654 | 692 | 652 | 731 | 661 | 718 | 555 |
| | None | 8994 | 8919 | 8962 | 8947 | 9046 | 8898 | 9000 | 9049 | 9046 | 9071 | 9000 | 8970 | 9044 | 8975 | 9010 | 8973 | 9089 |
| $\theta = \theta^*$ | All | 8148 | 5081 | 5187 | 5352 | 5195 | 4551 | 4606 | 4694 | 4628 | 4376 | 4267 | 4360 | 3982 | 4241 | 4400 | 4181 | 4614 |
| for all | Any sg | – | 2394 | 2221 | 2178 | 2312 | 3039 | 2830 | 2807 | 2798 | 3231 | 3281 | 3164 | 3607 | 3472 | 3173 | 3436 | 2886 |
| | None | 1852 | 2525 | 2592 | 2470 | 2493 | 2410 | 2564 | 2499 | 2574 | 2393 | 2452 | 2476 | 2411 | 2287 | 2427 | 2383 | 2500 |
| $\theta = 1.5\theta^*$ | All | 9707 | 7990 | 7606 | 7614 | 7772 | 6763 | 6959 | 6829 | 7036 | 6151 | 6046 | 6081 | 6171 | 5805 | 6019 | 5779 | 6108 |
| for all | Any sg | – | 1500 | 1921 | 1939 | 1759 | 2781 | 2570 | 2716 | 2471 | 3368 | 3513 | 3442 | 3373 | 3781 | 3543 | 3742 | 3437 |
| | None | 293 | 510 | 473 | 447 | 469 | 456 | 471 | 455 | 493 | 481 | 441 | 477 | 456 | 414 | 438 | 479 | 455 |
| $\theta_S = \theta^*$ | All | 4274 | 1303 | 1996 | 1375 | 1848 | 1289 | 1508 | 1239 | 1655 | 1593 | 1511 | 1508 | 1648 | 1553 | 1623 | 1533 | 1830 |
| for 50% | Correct sg | – | 3392 | 0 | 0 | 0 | 1256 | 1306 | 1452 | 0 | 547 | 313 | 307 | 0 | 494 | 128 | 114 | 0 |
| | Other sg | – | 101 | 2302 | 2719 | 2005 | 2029 | 1840 | 1538 | 2349 | 2182 | 2382 | 2405 | 2485 | 2208 | 2314 | 2693 | 2294 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 2022 | 0 | 0 | 895 | 1585 | 0 | 645 | 945 | 994 | 739 | 635 | 612 | 593 | 382 | 297 |
| | None | 5726 | 5204 | 5702 | 5906 | 6147 | 5426 | 5346 | 5771 | 5996 | 5678 | 5794 | 5780 | 5867 | 5745 | 5935 | 5660 | 5876 |
| $\theta_S = 1.5\theta^*$ | All | 6368 | 1538 | 3062 | 2071 | 3076 | 1642 | 1958 | 1427 | 2482 | 1906 | 1962 | 1778 | 2320 | 2107 | 2291 | 2203 | 2739 |
| for 50% | Correct sg | – | 5801 | 0 | 0 | 0 | 2932 | 2986 | 3029 | 0 | 1640 | 1006 | 1017 | 0 | 1497 | 471 | 425 | 0 |
| | Other sg | – | 41 | 3440 | 4196 | 2922 | 2546 | 2206 | 2293 | 3717 | 3167 | 3740 | 3808 | 4100 | 3104 | 3535 | 4004 | 3439 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 3208 | 4122 | 2593 | 2176 | 2014 | 1249 | 2387 | 1763 | 2419 | 2754 | 2539 | 1221 | 1500 | 2635 | 2101 |
| | None | 3632 | 2620 | 3498 | 3733 | 4002 | 2880 | 2850 | 3251 | 3801 | 3287 | 3292 | 3397 | 3580 | 3292 | 3703 | 3368 | 3822 |
| $\theta_S = 2\theta^*$ | All | 8126 | 1363 | 3945 | 2243 | 4124 | 1521 | 2042 | 1217 | 3161 | 2061 | 2027 | 2020 | 2615 | 2165 | 2492 | 2494 | 3284 |
| for 50% | Correct sg | – | 7767 | 0 | 0 | 0 | 4678 | 4945 | 4770 | 0 | 3018 | 2086 | 2194 | 0 | 2956 | 1273 | 1175 | 0 |
| | Other sg | – | 4 | 4437 | 6014 | 3583 | 2679 | 1907 | 2696 | 4825 | 3426 | 4367 | 4309 | 5607 | 3418 | 4432 | 4828 | 4821 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 4283 | 5991 | 3361 | 2645 | 1794 | 2687 | 4691 | 3389 | 4258 | 3943 | 5346 | 3406 | 4377 | 3603 | 4037 |
| | None | 1874 | 866 | 1618 | 1743 | 2293 | 1122 | 1106 | 1317 | 2014 | 1495 | 1520 | 1477 | 1778 | 1461 | 1803 | 1503 | 1895 |

Table 4.2: Use of SHAPES with n=200, $\alpha = 0.10$, $\alpha_F = 0.05$, $\theta^* = 0.146$.

| Truth | | Full | $K=1,\ell=1$ A | B | C | Z | $K=2,\ell=1$ A | B | C | Z | $K=4,\ell=1$ A | B | C | Z | $K=8,\ell=1$ A | B | C | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency of any rejection | | | | | | | | | | | | | | | | | | |
| All | $\theta=0$ | 10.1 | 10.8 | 10.4 | 10.5 | 9.5 | 10.3 | 9.6 | 9.5 | 9.8 | 10.0 | 10.3 | 10.1 | 10.9 | 10.4 | 9.5 | 10.6 | 9.7 |
| | $\theta=\theta^*$ | 81.5 | 74.8 | 74.1 | 75.3 | 75.1 | 76.3 | 74.8 | 75.5 | 74.5 | 75.6 | 75.9 | 75.1 | 75.7 | 77.6 | 75.8 | 75.9 | 76.0 |
| | $\theta=1.5\theta^*$ | 97.1 | 94.9 | 95.3 | 95.5 | 95.3 | 95.4 | 95.4 | 95.5 | 95.2 | 95.3 | 95.7 | 95.4 | 95.7 | 95.8 | 95.5 | 95.3 | 95.9 |
| Subgroup | $\theta_S=\theta^*$ | 42.7 | 48.0 | 43.0 | 40.9 | 38.5 | 46.9 | 45.6 | 40.8 | 40.0 | 43.4 | 44.1 | 38.9 | 41.3 | 42.6 | 42.1 | 40.3 | 40.3 |
| | $\theta_S=1.5\theta^*$ | 63.7 | 73.8 | 65.0 | 62.7 | 60.0 | 71.9 | 69.1 | 65.0 | 60.5 | 67.4 | 68.6 | 61.7 | 62.9 | 66.9 | 64.9 | 60.8 | 60.6 |
| | $\theta_S=2\theta^*$ | 81.3 | 91.3 | 83.8 | 82.6 | 77.1 | 89.7 | 87.5 | 84.1 | 78.6 | 86.0 | 85.5 | 81.7 | 81.1 | 85.6 | 83.0 | 80.6 | 79.4 |
| Frequency of rejection for a group with $\theta_{\hat{S}}\geq\theta^*$ | | | | | | | | | | | | | | | | | | |
| Subgroup | $\theta_S=\theta^*$ | 0.0 | 33.9 | 20.2 | 0.0 | 0.0 | 24.0 | 25.5 | 0.0 | 0.0 | 15.4 | 16.4 | 0.0 | 0.0 | 10.7 | 10.3 | 0.0 | 0.0 |
| | $\theta_S=1.5\theta^*$ | 0.0 | 58.0 | 32.1 | 41.2 | 25.9 | 46.7 | 41.9 | 49.1 | 30.3 | 34.4 | 30.9 | 32.1 | 37.7 | 26.7 | 20.1 | 21.2 | 24.0 |
| | $\theta_S=2\theta^*$ | 81.3 | 91.3 | 82.3 | 82.3 | 74.8 | 89.7 | 86.8 | 83.9 | 76.9 | 86.0 | 85.2 | 72.3 | 79.4 | 85.6 | 82.9 | 61.7 | 64.8 |
| Frequency of correct decision | | | | | | | | | | | | | | | | | | |
| All | $\theta=0$ | 89.9 | 89.2 | 89.6 | 89.5 | 90.5 | 89.7 | 90.4 | 90.5 | 90.2 | 90.0 | 89.7 | 89.9 | 89.1 | 89.6 | 90.5 | 89.4 | 90.3 |
| | $\theta=\theta^*$ | 81.5 | 50.8 | 51.9 | 53.5 | 52.0 | 48.3 | 48.2 | 49.0 | 49.5 | 47.9 | 46.0 | 46.3 | 46.0 | 42.5 | 44.1 | 42.4 | 43.1 |
| | $\theta=1.5\theta^*$ | 97.1 | 79.9 | 76.1 | 76.1 | 77.7 | 74.5 | 74.5 | 74.2 | 76.6 | 74.4 | 70.8 | 72.1 | 72.3 | 67.7 | 70.6 | 68.4 | 68.7 |
| Subgroup | $\theta_S=\theta^*$ | 0.0 | 33.9 | 0.0 | 0.0 | 0.0 | 24.0 | 0.0 | 0.0 | 0.0 | 15.4 | 0.0 | 0.0 | 0.0 | 10.7 | 0.0 | 0.0 | 0.0 |
| | $\theta_S=1.5\theta^*$ | 0.0 | 58.0 | 0.0 | 0.0 | 0.0 | 46.7 | 0.0 | 0.0 | 0.0 | 34.4 | 0.0 | 0.0 | 0.0 | 26.7 | 0.0 | 0.0 | 0.0 |
| | $\theta_S=2\theta^*$ | 0.0 | 77.7 | 0.0 | 0.0 | 0.0 | 67.0 | 0.0 | 0.0 | 0.0 | 55.2 | 0.0 | 0.0 | 0.0 | 47.5 | 0.0 | 0.0 | 0.0 |

Table 4.3: Type I error and power to identify a group with $\theta\geq\theta^*=0.146$ using SHAPES based on 10,000 replications.

| | | | $K=1,\ell=1$ | | | | $K=2,\ell=1$ | | | | $K=4,\ell=1$ | | | | $K=8,\ell=1$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 1006 | 334 | 321 | 335 | 322 | 290 | 316 | 332 | 309 | 325 | 280 | 325 | 313 | 237 | 279 | 243 | 257 |
| for all | Any sg | – | 747 | 717 | 718 | 632 | 737 | 647 | 621 | 674 | 674 | 753 | 683 | 777 | 804 | 667 | 821 | 714 |
| | None | 8994 | 8919 | 8962 | 8947 | 9046 | 8973 | 9037 | 9047 | 9017 | 9001 | 8967 | 8992 | 8910 | 8959 | 9054 | 8936 | 9029 |
| $\theta = \theta^*$ | All | 8148 | 5081 | 5187 | 5352 | 5195 | 4825 | 4819 | 4903 | 4952 | 4787 | 4595 | 4628 | 4586 | 4250 | 4412 | 4235 | 4308 |
| for all | Any sg | – | 2394 | 2221 | 2178 | 2312 | 2809 | 2663 | 2650 | 2501 | 2773 | 2991 | 2880 | 2987 | 3511 | 3169 | 3354 | 3288 |
| | None | 1852 | 2525 | 2592 | 2470 | 2493 | 2366 | 2518 | 2447 | 2547 | 2440 | 2414 | 2492 | 2427 | 2239 | 2419 | 2411 | 2404 |
| $\theta = 1.5\theta^*$ | All | 9707 | 7990 | 7606 | 7614 | 7772 | 7454 | 7453 | 7420 | 7663 | 7442 | 7080 | 7214 | 7230 | 6773 | 7056 | 6838 | 6870 |
| for all | Any sg | – | 1500 | 1921 | 1939 | 1759 | 2083 | 2083 | 2133 | 1854 | 2085 | 2488 | 2324 | 2340 | 2804 | 2497 | 2692 | 2716 |
| | None | 293 | 510 | 473 | 447 | 469 | 463 | 464 | 447 | 483 | 473 | 432 | 462 | 430 | 423 | 447 | 470 | 414 |
| $\theta_S = \theta^*$ | All | 4274 | 1303 | 1996 | 1375 | 1848 | 1383 | 1730 | 1254 | 1726 | 1466 | 1645 | 1354 | 1441 | 1301 | 1566 | 1341 | 1440 |
| for 50% | Correct sg | – | 3392 | 0 | 0 | 0 | 2400 | 0 | 0 | 0 | 1542 | 0 | 0 | 0 | 1074 | 0 | 0 | 0 |
| | Other sg | – | 101 | 2302 | 2719 | 2005 | 911 | 2830 | 2823 | 2272 | 1333 | 2766 | 2537 | 2693 | 1882 | 2645 | 2691 | 2591 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 2022 | 0 | 0 | 0 | 2552 | 0 | 0 | 0 | 1636 | 0 | 0 | 0 | 1031 | 0 | 0 |
| | None | 5726 | 5204 | 5702 | 5906 | 6147 | 5306 | 5440 | 5923 | 6002 | 5659 | 5589 | 6109 | 5866 | 5743 | 5789 | 5968 | 5969 |
| $\theta_S = 1.5\theta^*$ | All | 6368 | 1538 | 3062 | 2071 | 3076 | 1682 | 2563 | 1509 | 2716 | 1911 | 2540 | 1869 | 2209 | 1827 | 2544 | 1939 | 2198 |
| for 50% | Correct sg | – | 5801 | 0 | 0 | 0 | 4669 | 0 | 0 | 0 | 3435 | 0 | 0 | 0 | 2668 | 0 | 0 | 0 |
| | Other sg | – | 41 | 3440 | 4196 | 2922 | 841 | 4348 | 4986 | 3339 | 1394 | 4324 | 4297 | 4080 | 2194 | 3942 | 4145 | 3862 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 3208 | 4122 | 2593 | 0 | 4191 | 4911 | 3026 | 0 | 3095 | 3213 | 3774 | 0 | 2012 | 2125 | 2401 |
| | None | 3632 | 2620 | 3498 | 3733 | 4002 | 2808 | 3089 | 3505 | 3945 | 3260 | 3136 | 3834 | 3711 | 3311 | 3514 | 3916 | 3940 |
| $\theta_S = 2\theta^*$ | All | 8126 | 1363 | 3945 | 2243 | 4124 | 1620 | 3291 | 1335 | 3607 | 2020 | 3149 | 1904 | 2788 | 2026 | 3413 | 2287 | 2861 |
| for 50% | Correct sg | – | 7767 | 0 | 0 | 0 | 6698 | 0 | 0 | 0 | 5517 | 0 | 0 | 0 | 4746 | 0 | 0 | 0 |
| | Other sg | – | 4 | 4437 | 6014 | 3583 | 653 | 5456 | 7072 | 4251 | 1066 | 5404 | 6262 | 5319 | 1785 | 4885 | 5774 | 5083 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 4283 | 5991 | 3361 | 647 | 5391 | 7057 | 4087 | 1064 | 5371 | 5328 | 5150 | 1785 | 4879 | 3882 | 3617 |
| | None | 1874 | 866 | 1618 | 1743 | 2293 | 1029 | 1253 | 1593 | 2142 | 1397 | 1447 | 1834 | 1893 | 1443 | 1702 | 1939 | 2056 |

Table 4.4: Use of SHAPES with n=200, $\alpha = 0.10$, $\alpha_F = 0.05$, $\theta^* = 0.146$.

for some subgroup.

**Homogeneous effects:** For $\theta = \theta^*$, a full population analysis rejects with 81.5% frequency. With SHAPES the frequency of any rejection increases from an average across the definitions of 74.8% when $K = 1$ to 76.3% when $K = 8$. Underlying the slight increase in rejection are two trends as $K$ increases: a decrease in rejection for the full population from an average of 52.0 to 43.0% and an increase in subgroup rejection from 22.8 to 33.3%. Even under homogeneous effects, we are more likely to find a promising subgroup as we explore more. Similar trends are present when $\theta = 1.5\theta^*$.

**Heterogeneous effects:** For $\theta_S = 1.5\theta^*$ and subgroup definition A, the identification of the correct subgroup ranges from 58.0% when $K = 1$ to 26.7% when $K = 8$. Here we assume the correct variable, $X_1$, is always included in the analysis. For the other subgroup definitions we never recover the correct subgroup because $n_{leaves} = 1$. Identifying a subgroup with $\theta_{\hat{S}} \geq \theta^*$ occurs most frequently when $K$ corresponds to the number of covariates in the subgroup definition. For example, with definition C and $\theta_S = 1.5\theta^*$ the performance is best when $K = 2$ with a rate of 65.0% compared to when $K = 1$ a rate of 62.7% or when $K = 4$ a rate of 61.7%. With $\theta_S = 2\theta^*$, we observe a similar trend.

**Summary and general trends for heterogeneous effects:** We now summarize trends using overall rejection, power for $\theta^*$ and correct subgroup identification rates as $K$ and $\ell$ varies. In the following plots, $K = 0$ or $\ell = 0$ represent an analysis of only the full population. Figure 4.7 corresponds to results shown in Table 4.1 for $\theta_S = 1.5\theta^*$ and Figure 4.8 to Table 4.3. Additionally, Figure 4.9 considers $\ell = \{0, 1, 2, 3\}$ for $K = 4$. We generally observe that performance is related to the agreement between $K$, $\ell$ and the subgroup definition. The best performance occurs when these agree, but there is wide variation depending on our metric. When the correct subgroup is recoverable, rates of correct identification are highest when the subgroup definition, $\ell$ and $K$ agree, as represented by the green peaks. As $\ell$ or $K$ increases beyond this, performance decreases but remains relatively stable. As $K$ or $\ell$ increases, rates of any rejection decrease more slowly and consistently than other metrics, as represented by the slight negative slope of the black lines in all three figures. This relationship occurs across subgroup definitions and for both $\theta_S = 1.5\theta^*$ and $\theta_S = 2\theta^*$. Of particular note, as $\ell$ increases, rejection rates change little. When $\theta_S = 2\theta^*$, identification of a group with $\theta_{\hat{S}} \geq \theta^*$ is similar to rates of any rejection. When $\theta_S = 1.5\theta^*$ the rates of $\theta^*$-power mimic the correct subgroup identification rates when $\ell = 3$ but tend to decrease much less as $\ell$ or $K$ increases.

Regarding comparing SHAPES to a full population analysis we see that under heterogeneous effects we sometimes do better and sometimes do worse than if we only analyzed the full population. With rates of any rejection we see a bump at the true $\ell$ except with scenario Z. With $\theta^*$-power

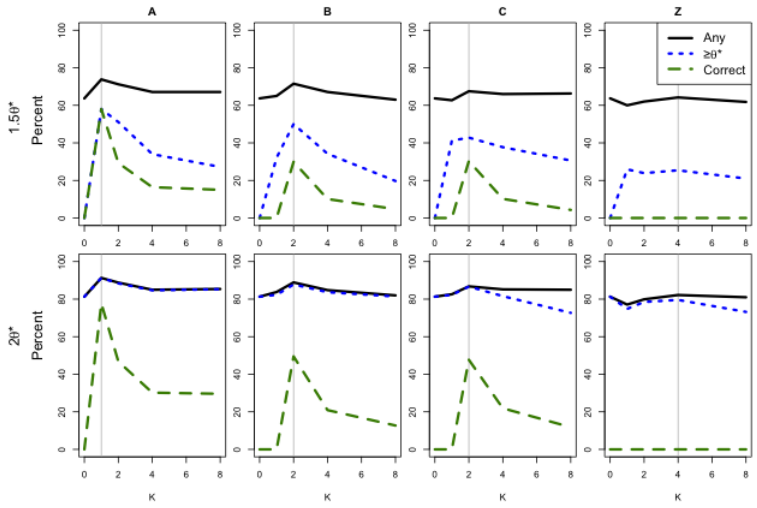



Figure 4.7: Performance of SHAPES with $\ell = minimum(3, K)$ when $\theta_S = 1.5\theta^*$ (top row) and $\theta_S = 2\theta^*$ (bottom row) for increasing K in terms of any rejection (Any), rejection for a subgroup with a moderate effect ($\geq \theta^*$ ) or for the correct subgroup (Correct).

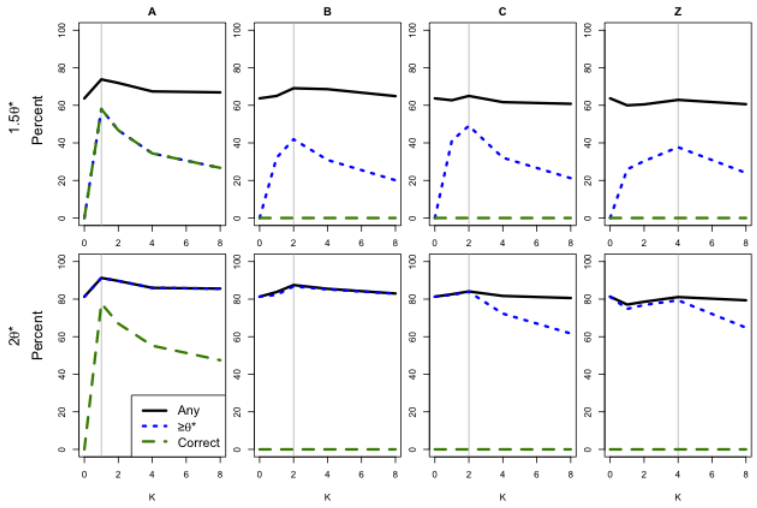



Figure 4.8: Performance of SHAPES with $\ell = 1$ when $\theta_S = 1.5\theta^*$ (top row) and $\theta_S = 2\theta^*$ (bottom row) for increasing K in terms of any rejection (Any), rejection for a subgroup with a moderate effect ($\geq \theta^*$ ) or for the correct subgroup (Correct).

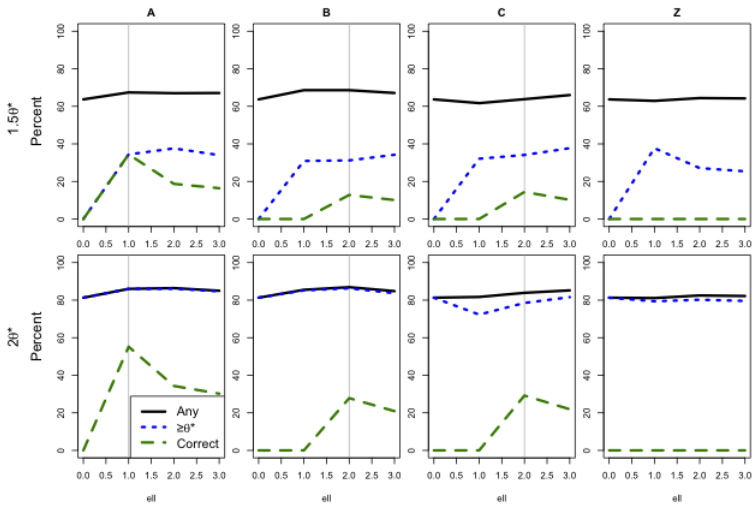



Figure 4.9: Performance of SHAPES with $K = 4$ when $\theta_S = 1.5\theta^*$ (top row) and $\theta_S = 2\theta^*$ (bottom row) for increasing $\ell$ in terms of any rejection (Any), rejection for a subgroup with a moderate effect ($\geq \theta^*$ ) or for the correct subgroup (Correct).

we always do better looking for subgroups when $\theta_S = 1.5\theta^*$ as the marginal effect is less than $\theta^*$, but with $\theta_S = 2\theta^*$ the marginal effect is equal to $\theta^*$ so an advantage is not always present and decreases as $\ell$ or $K$ increases corresponding to greater exploration. Overall, these results suggest that if our priority is any rejection, we will not experience much difference using SHAPES, but for correct subgroup identification or power for $\theta^*$ we have gains. In the remainder of the chapter we use $K = 4$ with $\ell = 3$ as these values provide a balance between a completely exploratory and a confirmatory analysis.

### *4.4 Varying the Proportion of the Subgroup ($\zeta$)*

We next consider performance as the subgroup definition is fixed but the prevalence of the subgroup varies, with values $\zeta = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. We use the same set-up as with LR and hold either $\theta_S$ or $\theta$ constant. With LR we found the highest correct subgroup identification occurs at $\zeta = 0.50$ with a significant reduction in correct subgroup identification closer to 0 or 1, i.e. as the truth gets closer to the null or homogeneous effects. In simulations we again use $K = 4$, $\ell = 3$, $\alpha_F = 0.05$ and the *minimum p-value* utility function.

When holding $\theta_S$ fixed, results are shown in Figure 4.10 for $\theta_S = 1.5\theta^*$ and Figure 4.11 for

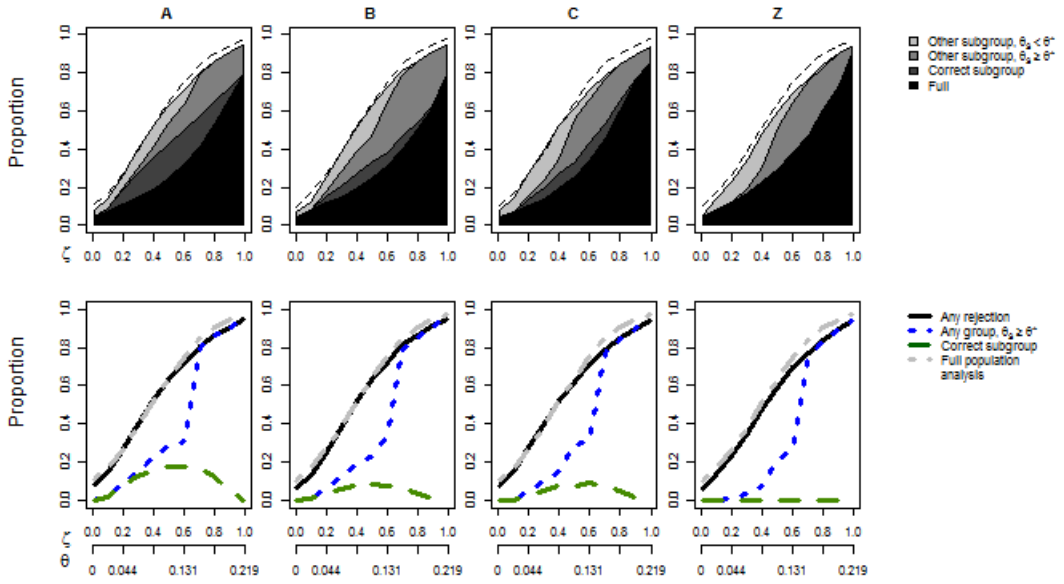



Figure 4.10: Performance of shapes with $K = 4$, $\ell = 3$ when holding $\theta_S = 1.5\theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

$\theta_S = 2\theta^*$. The top row of plots highlights the group that was selected and the bottom row the performance for the three metrics with the dotted line indicating the rejection rate for the full population. For all subgroup definitions, both values of $\theta_S$, and most values of $\zeta < 0.5$, SHAPES has a higher rejection rate than the full population analysis. The rate is sometimes lower for values of $\zeta > 0.5$ (we note that it remains higher or roughly equivalent for $\theta_S = 2\theta^*$ with definitions A, B and C). As $\zeta$ approaches 1, the rate of rejection for all increases sharply, but the rate of correct subgroup identification decreases, peaking between $\zeta = 0.40$ and $\zeta = 0.60$ for most subgroup definitions with a maximum rate of 40% for A and 20% for B and C. In terms of $\theta^*$-power, the full population analysis achieves non-zero power only for $\zeta \geq \frac{2}{3}$ when $\theta_S = 1.5\theta^*$ or for $\zeta \geq 0.5$ when $\theta_S = 2\theta^*$, whereas SHAPES achieves non-zero power for the range of $\zeta$.

When holding $\theta$ fixed, results are shown in Figure 4.12 for $\theta = 0.75\theta^*$ and Figure 4.13 for $\theta = \theta^*$. There is again a switch around 50%, where for $\zeta < 0.50$ we reject more often searching for subgroups and for $\zeta > 0.50$ we reject more often only analyzing the full population. For $\zeta = \{0.20, 0.30\}$ we observe the highest rates of decision for the correct subgroup: at or greater than 40% for the identifiable definitions. As $\zeta$ approaches 1 there is an increasing rate of decisions for the entire population while concurrently correct subgroup identification drops to 0% for $\zeta > 0.80$. This is driven

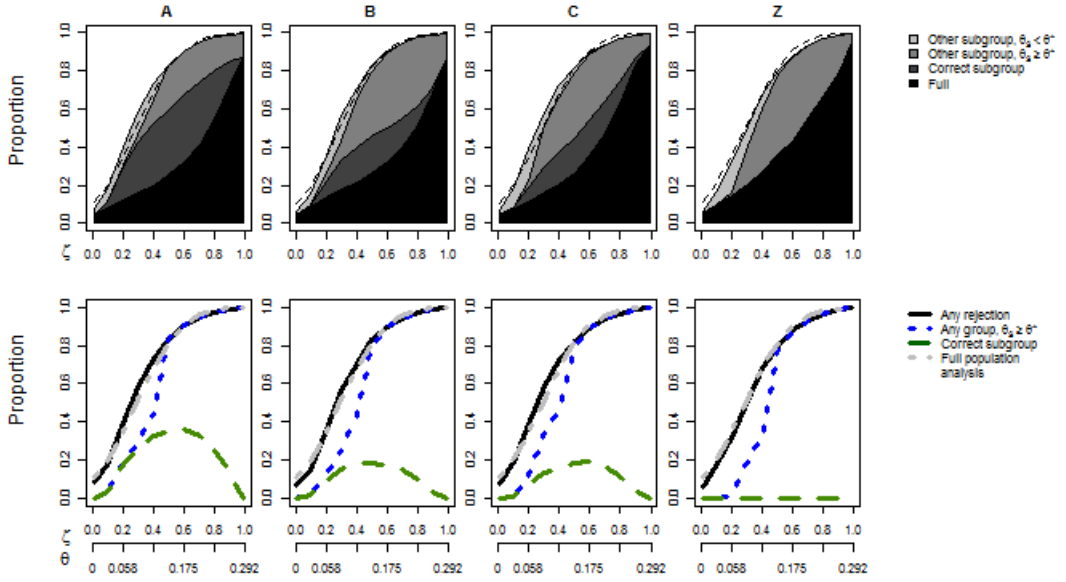



Figure 4.11: Performance of shapes with $K = 4$, $\ell = 3$ when holding $\theta_S = 2\theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

by the decreasing $\theta^*$, as the truth underlying the fixed $\theta$ transitions from a strong heterogeneous effect to a moderate homogeneous effect.

Beyond describing rates of correct group identification, we also aim to describe the frequency and characteristics of other subgroups that are selected as prevalences vary. Figures 4.14 and 4.15 contain quilt plots of the results with definition A and $\zeta = \{0.20, 0.40, 0.60, 0.80\}$. For these plots we fix $\theta_S = 2\theta^*$ for the heterogeneous effect and display homogeneous effects of $\theta = \zeta\theta_S$ and $\theta = 0.75\zeta\theta_S$ in the upper two regions of the figure.

With heterogeneous effects and definition A, there are always 12 unique subgroups that have the same effect as the correct subgroup but a smaller prevalence. When holding $\theta_S$ fixed, the value of $\zeta$ drives the number of subgroups with $\theta_S \geq \theta^*$: when $\zeta = 0.20$, all subgroups except the correct subgroup or proper subsets of the correct subgroup have $\theta_S < \theta^*$ and when $\zeta = 0.80$ only the subgroups corresponding to $X_1^C$ and subsets thereof have an effect of $\theta_S < \theta^*$. Regarding proximity to the correct subgroup in $(\theta_S, \zeta)$-space, with small values of $\zeta$, we observe that only proper subsets of the correct subgroup are near the truth and as $\zeta$ increases we observe an increasing number of other candidate subgroups closer to the true subgroup. This suggests that when we select a small subgroup under this heterogeneous effects scenario, we have either selected the true predictive subgroup or a

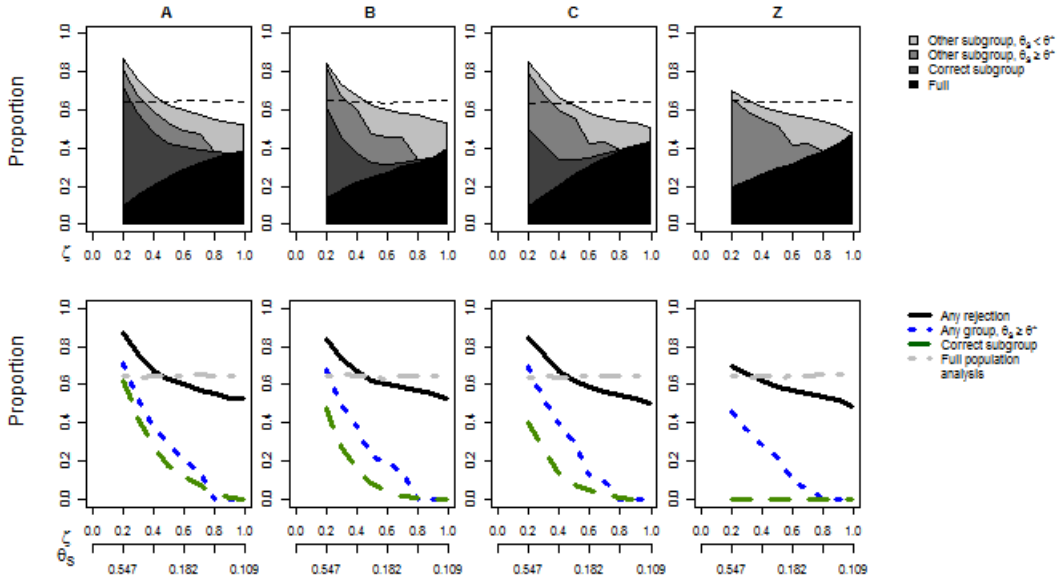



Figure 4.12: Performance of shapes with $K = 4$, $\ell = 3$ when holding the marginal effect at $\theta = 0.75\theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

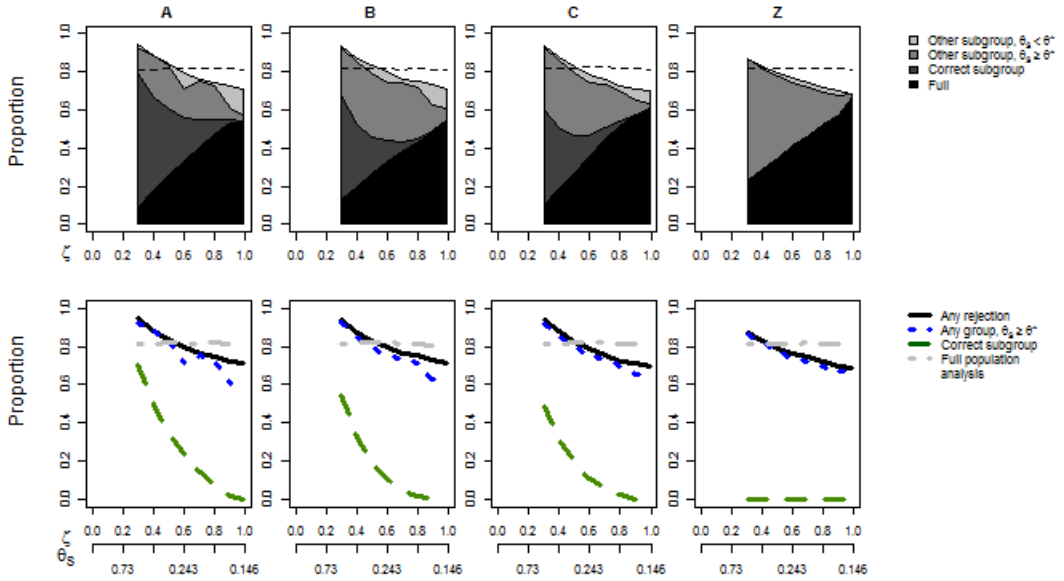



Figure 4.13: Performance of shapes with $K = 4$, $\ell = 3$ when holding the marginal effect at $\theta = \theta^*$ and varying the prevalence of the predictive subgroup ($\zeta$) from 0 to 1.00.

subset thereof (with $\zeta = 0.20$ we select a proper subset with frequency 2.8% and the correct subgroup 16.8%) but when we select a larger subgroup it is either the true predictive subgroup or a mixture of the true group and its complement (with $\zeta = 0.80$ we never select a subgroup with $\theta_{\hat{S}} < \theta^*$).

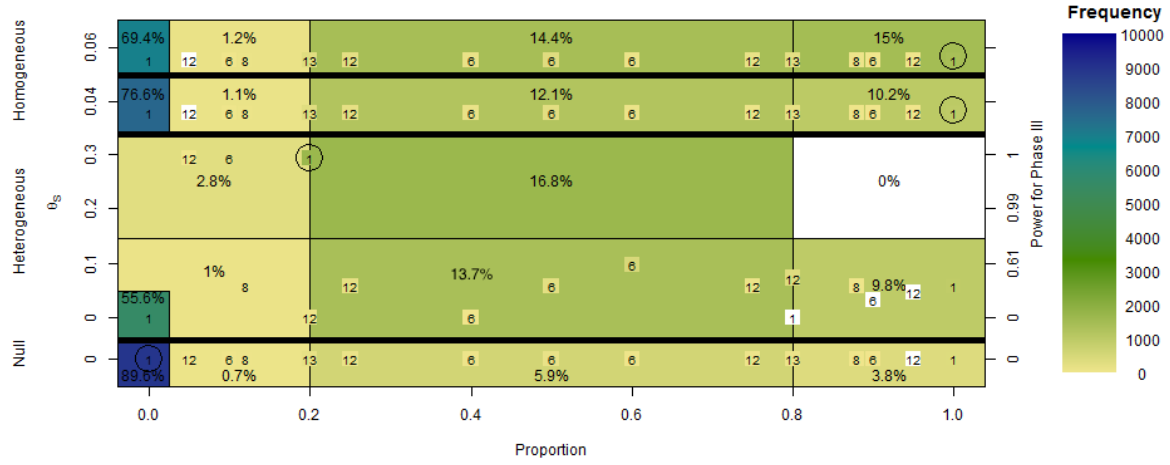
The quilt plots highlight that the number of candidate subgroups in each subregion depends on $\theta_S$ and the covariate distribution with corresponding $\zeta$. Ideally, we will select the truth, but we have previously shown this is not always possible, so secondly we may aim to be in the same subregion as the truth. With $\zeta = 0.40$ we observe a 44.2% of rejection for the correct subgroup or for a candidate in the same subregion. Here there are a total of 13 possible candidates. With $\zeta = 0.60$, the number of candidates increases to 44 and the rate is 54.9%; with more candidates in the true subregion we improve our chances of getting something closer to the truth. We also generally observe that rejection rates remain low, $< 4\%$, for candidate subgroups that are not in the true subregion.

In considering the varying prevalences of a predictive subgroup, we summarize several points. First, under this simulation set-up, $\zeta = 0.50$ is a transition point, similar to LR. With $\zeta < 0.50$ we see greater rejection rates for subgroup exploration than compared to the full population analysis. Relatedly, we observe the power to detect a group with $\theta_{\hat{S}} \geq \theta^*$ is sensitive to the marginal effects. If the primary aim is to detect a correct subgroup, we note that subgroup effects of $1.5\theta^*$ and $2\theta^*$ may not be sufficient for reasonable rates of correct subgroup identification with $n = 200$, $\ell = 3$; with larger rates, such as $\frac{10}{3}\theta^*$ performance improves. Finally, we make the observation that the distribution of candidate subgroups depends on $\zeta$, and is likely different for various covariate distributions, true subgroup definitions, and values of $\theta_S$.

### *4.5 Varying the Baseline Rate ($\beta_0$)*

In Table 4.5 we simulate with baseline rates $\beta_0 = \{0.10, 0.30, 0.50\}$. We implement our usual simulation settings and as with LR we hold $\theta^* = 0.146$, and therefore anticipate gains in power when $\beta_0$ decreases from 0.30 because of the reduction in variance for extreme probability values, i.e. close to 0 or 1. Critical values were derived assuming the correct covariate distribution and corresponding $\beta_0$.

$\zeta = 0.20$




$\zeta = 0.40$

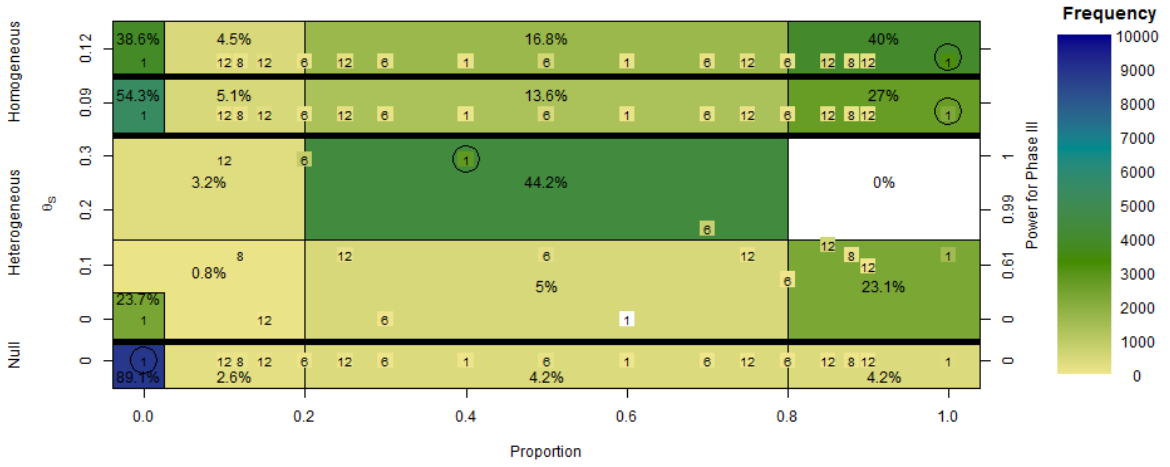
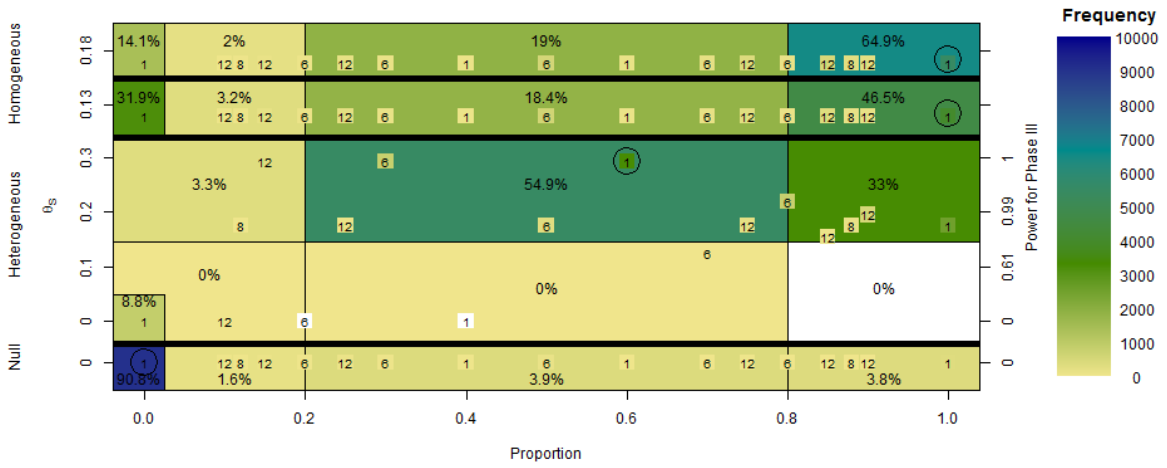



Figure 4.14: Quilt plot for definition A using SHAPES and $\zeta = 0.20$ and 0.40 with $\ell = 3$, $K = 4$, $\alpha_F = 0.05$ and $\theta_S = 2\theta^*$.

$\zeta = 0.60$




$\zeta = 0.80$

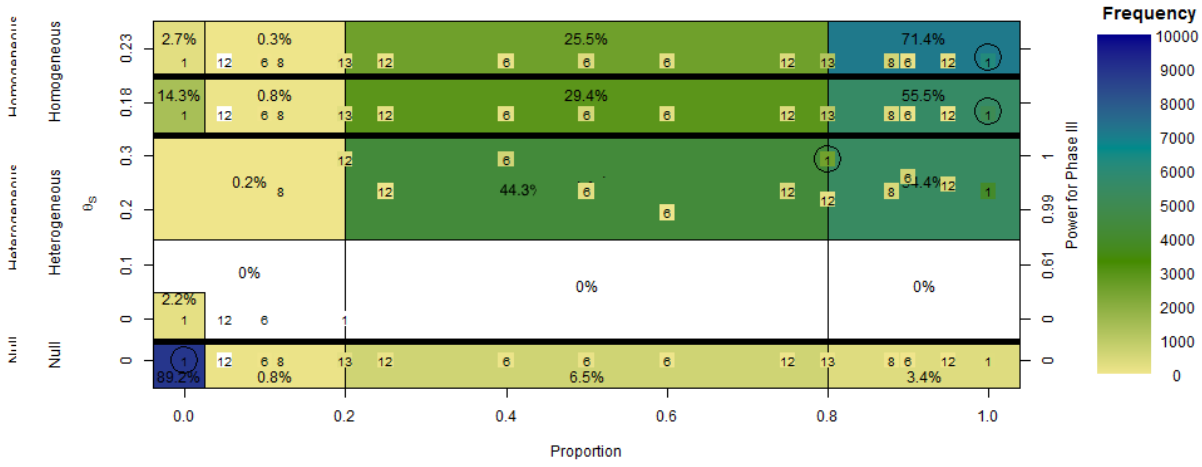



Figure 4.15: Quilt plot for definition A using SHAPES and $\zeta = 0.60$ and 0.80 with $\ell = 3$, $K = 4$, $\alpha_F = 0.05$ and $\theta_S = 2\theta^*$.

| | | $\beta_0 = 0.10$ | | | | | $\beta_0 = 0.30$ | | | | | $\beta_0 = 0.50$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | Full | A | B | C | Z | Full | A | B | C | Z |
| $\theta = 0$ | All | 989 | 147 | 200 | 297 | 161 | 953 | 325 | 346 | 338 | 304 | 919 | 353 | 369 | 363 | 386 |
| for all | Any sg | – | 828 | 769 | 717 | 849 | – | 604 | 654 | 692 | 652 | – | 677 | 700 | 647 | 589 |
| | None | 9011 | 9025 | 9031 | 8986 | 8990 | 9047 | 9071 | 9000 | 8970 | 9044 | 9081 | 8970 | 8931 | 8990 | 9025 |
| | Any rej., % | 9.9 | 9.8 | 9.7 | 10.1 | 10.1 | 9.5 | 9.3 | 10.0 | 10.3 | 9.6 | 9.2 | 10.3 | 10.7 | 10.1 | 9.8 |
| $\theta = \theta^*$ | All | 9347 | 2478 | 2956 | 2836 | 2467 | 8192 | 4376 | 4267 | 4360 | 3982 | 7916 | 4858 | 4915 | 4983 | 5009 |
| for all | Any sg | – | 6853 | 6347 | 6483 | 6902 | – | 3231 | 3281 | 3164 | 3607 | – | 2359 | 2300 | 2244 | 2118 |
| | None | 653 | 669 | 697 | 681 | 631 | 1808 | 2393 | 2452 | 2476 | 2411 | 2084 | 2783 | 2785 | 2773 | 2873 |
| | Any rej., % | 93.5 | 93.3 | 93.0 | 93.2 | 93.7 | 81.9 | 76.1 | 75.5 | 75.2 | 75.9 | 79.2 | 72.2 | 72.2 | 72.3 | 71.3 |
| $\theta_S = 1.5\theta^*$ | All | 8148 | 968 | 1224 | 1122 | 1011 | 6392 | 1906 | 1962 | 1778 | 2320 | 6048 | 2198 | 2213 | 2336 | 2680 |
| for 50% | Correct sg | – | 2232 | 1511 | 1727 | 0 | – | 1640 | 1006 | 1017 | 0 | – | 1893 | 879 | 890 | 0 |
| | Other sg | – | 5617 | 6025 | 5904 | 7624 | – | 3167 | 3740 | 3808 | 4100 | – | 2240 | 3141 | 3007 | 3164 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 1843 | 3346 | 2427 | 2318 | – | 1763 | 2419 | 2754 | 2539 | – | 1177 | 2167 | 2300 | 2258 |
| | None | 1852 | 1183 | 1240 | 1247 | 1365 | 3608 | 3287 | 3292 | 3397 | 3580 | 3952 | 3669 | 3767 | 3767 | 4156 |
| | Any rej., % | 81.5 | 88.2 | 87.6 | 87.5 | 86.3 | 63.9 | 67.1 | 67.1 | 66.0 | 64.2 | 60.5 | 63.3 | 62.3 | 62.3 | 58.4 |
| | $\theta^*$-power, % | 0.0 | 40.8 | 48.6 | 41.5 | 23.2 | 0.0 | 34.0 | 34.2 | 37.7 | 25.4 | 0.0 | 30.7 | 30.5 | 31.9 | 22.6 |
| $\theta_S = 2\theta^*$ | All | 9294 | 990 | 1277 | 1088 | 1044 | 8184 | 2061 | 2027 | 2020 | 2615 | 8063 | 2198 | 2392 | 2281 | 3131 |
| for 50% | Correct sg | – | 3297 | 2465 | 2677 | 0 | – | 3018 | 2086 | 2194 | 0 | – | 4053 | 2304 | 2294 | 0 |
| | Other sg | – | 5420 | 5947 | 5909 | 8601 | – | 3426 | 4367 | 4309 | 5607 | – | 2318 | 3734 | 3780 | 4772 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 5213 | 5515 | 5417 | 7665 | – | 3389 | 4258 | 3943 | 5346 | – | 2304 | 3685 | 3455 | 4621 |
| | None | 706 | 293 | 311 | 326 | 355 | 1816 | 1495 | 1520 | 1477 | 1778 | 1937 | 1431 | 1570 | 1645 | 2097 |
| | Any rej., % | 92.9 | 97.1 | 96.9 | 96.7 | 96.5 | 81.8 | 85.0 | 84.8 | 85.2 | 82.2 | 80.6 | 85.7 | 84.3 | 83.5 | 79.0 |
| | $\theta^*$-power, % | 92.9 | 95.0 | 92.6 | 91.8 | 87.1 | 81.8 | 84.7 | 83.7 | 81.6 | 79.6 | 80.6 | 85.5 | 83.8 | 80.3 | 77.5 |

Table 4.5: Varying baseline rate using SHAPES with $K = 4, \ell = 3, \alpha_F = 0.05$ and $\alpha = 0.10$.

Under the null, type I error control holds. When rejection occurs, the ratio of rejection for the full population compared to a subgroup is roughly 3:6 for $\beta_0 = \{0.30, 0.50\}$ and shifts to 2:7 for $\beta_0 = 0.10$ with variability by subgroup definition. We see a similar shift under homogeneous effects, where for $\beta_0 = 0.30$ the percent of rejections for a subgroup, averaged across the definitions, is 43.9% and with $\beta_0 = 0.10$ it is 71.2%. Any rejection rates increase from 75.7 to 93.3% for the same $\beta_0$ values.

Under heterogeneous effects with $\theta_S = 1.5\theta^*$, we observe a u-shaped relationship between rates of correct subgroup identification and $\beta_0$ for definition A and a negative relationship for definitions B and C. The negative relationship generally holds for all definitions for any rejection and $\theta^*$ power. With LR we also observed the u-shaped relationship for A. With $\theta_S = 2\theta^*$, the u-shaped relationship for correct subgroup identification exists for A, B, and C, but the relationship again appears negative with respect to any rejection and $\theta^*$-power. Under both values of $\theta_S$ and all baseline rates we observe higher rejection rates with SHAPES than using a full population analysis, except sometimes with definition Z. Making the same comparison with $\theta^*$-power and $\theta_S = 2\theta^*$, we find improvements with definitions A and B but losses with C and Z compared to the full population analysis.

Overall, we see baseline rate directly influences performance for most metrics. Taking advantage of gains in power as variance decreases requires method re-calibration.

### *4.6 Prognostic Effects* ($\omega$)

In simulating the presence of prognostic effects among the $K$ covariates, we used the same model from Section 3.8. Table 4.6 reports results from SHAPES when increasing the prognostic effect $\omega$ and holding the overall baseline rate at 0.30 with $\ell = 3$, $\alpha_F = 0.05$, and $K = 4$. With LR, we observed extreme type I error inflation with large prognostic effects. This inflation is less expected with SHAPES as this method directly evaluates a stratified test comparing treatment and control patients in the same subgroup, meaning that prognostic effects should cancel each other out. We see comparing to using LR, the rates of type I error are essentially not affected with SHAPES. Under the null, as the prognostic effect increases the type I error rate inflates slightly under definition A, but remains under 12% and holds under definition C compared to the nominal 10% rate.

Under a homogeneous effect of $\theta = \theta^*$ the overall rejection rate is similar by definition as $\omega$ increases, with a slight increase in the frequency of rejection for a subgroup. Under heterogeneous effects rates of any rejection and power for $\theta^*$ are similar for various values of $\omega$. There is an increase in correct subgroup identification when $\omega = 0.40$ for both definitions, which is particularly strong with definition C when $\theta_S = 2\theta^*$ where it goes from 19.9 to 30.5%. Potentially the correlation of

| Truth | Decision | $\beta_0 = 0.30$, $\omega = 0.00$ A | C | $\beta_0 = 0.275$, $\omega = 0.05$ A | C | $\beta_0 = 0.25$, $\omega = 0.10$ A | C | $\beta_0 = 0.20$, $\omega = 0.20$ A | C | $\beta_0 = 0.10$, $\omega = 0.40$ A | C |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0$ | All | 335 | 340 | 313 | 302 | 288 | 326 | 280 | 306 | 305 | 343 |
| for all | Any sg | 725 | 679 | 756 | 638 | 834 | 685 | 912 | 739 | 830 | 705 |
| | None | 8940 | 8981 | 8931 | 9060 | 8878 | 8989 | 8808 | 8955 | 8865 | 8952 |
| | **Any rej.,** % | 10.6 | 10.2 | 10.7 | 9.4 | 11.2 | 10.1 | 11.9 | 10.4 | 11.4 | 10.5 |
| $\theta = \theta^*$ | All | 4248 | 4441 | 4237 | 4456 | 4005 | 4374 | 4010 | 4343 | 4117 | 4112 |
| for all | Any sg | 3315 | 2985 | 3296 | 2885 | 3452 | 3035 | 3432 | 3052 | 3473 | 3262 |
| | None | 2437 | 2574 | 2467 | 2659 | 2543 | 2591 | 2558 | 2605 | 2410 | 2626 |
| | **Any rej.,** % | 75.6 | 74.3 | 75.3 | 73.4 | 74.6 | 74.1 | 74.4 | 74.0 | 75.9 | 73.7 |
| $\theta_S = 1.5\theta^*$ | All | 1931 | 2067 | 1840 | 2045 | 1815 | 1962 | 1793 | 1880 | 1731 | 1570 |
| for 50% | Correct sg | 1754 | 899 | 1742 | 827 | 1748 | 804 | 1729 | 893 | 1949 | 1281 |
| | Other sg | 3015 | 3606 | 3130 | 3544 | 3095 | 3613 | 3144 | 3767 | 2929 | 3992 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 1762 | 2698 | 1766 | 2617 | 1793 | 2706 | 1712 | 2804 | 1454 | 2869 |
| | None | 3300 | 3428 | 3288 | 3584 | 3342 | 3621 | 3334 | 3460 | 3391 | 3157 |
| | **Any rej.,** % | 67.0 | 65.7 | 67.1 | 64.2 | 66.6 | 63.8 | 66.7 | 65.4 | 66.1 | 68.4 |
| | $\theta^*$**-power,** % | 35.2 | 36.0 | 35.1 | 34.4 | 35.4 | 35.1 | 34.4 | 37.0 | 34.0 | 41.5 |
| $\theta_S = 2\theta^*$ | All | 1827 | 2120 | 1858 | 1949 | 1782 | 1906 | 1636 | 1755 | 1523 | 1316 |
| for 50% | Correct sg | 3438 | 1990 | 3346 | 1973 | 3322 | 2028 | 3521 | 2139 | 3997 | 3051 |
| | Other sg | 3331 | 4357 | 3411 | 4466 | 3439 | 4476 | 3410 | 4607 | 3173 | 4579 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 3284 | 4233 | 3383 | 4366 | 3406 | 4373 | 3374 | 4511 | 3138 | 4484 |
| | None | 1404 | 1533 | 1385 | 1612 | 1457 | 1590 | 1433 | 1499 | 1307 | 1054 |
| | **Any rej.,** % | 86.0 | 84.7 | 86.2 | 83.9 | 85.4 | 84.1 | 85.7 | 85.0 | 86.9 | 89.5 |
| | $\theta^*$**-power,** % | 85.5 | 83.4 | 85.9 | 82.9 | 85.1 | 83.1 | 85.3 | 84.0 | 86.6 | 88.5 |

Table 4.6: SHAPES with increasing prognostic rates with $K = 4$, $\ell = 3$, $\alpha_F = 0.05$, and $\alpha = 0.10$.

the prognostic and predictive effects is boosting the performance of SHAPES here.

### *4.7 Continuous Endpoint*

SHAPES can also be used with a continuous endpoint. Instead of assuming a baseline response rate, to generate the critical values we assume (correctly) the variance is $\epsilon = 1$ and we set $\theta^* = 0.30$, corresponding to the difference we have with 80% power to detect when $n = 200$. We set $\alpha_F = 0.05$ and $\alpha = 0.10$ and use values $(K, \ell) = \{(1,1), (2,2), (4,3), (8,3)\}$ with the typical subgroup definitions.

| | | | $K=1,\ell=1$ | | | | $K=2,\ell=2$ | | | | $K=4,\ell=3$ | | | | $K=8,\ell=3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | Full | A | B | C | Z | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 1023 | 340 | 341 | 337 | 331 | 330 | 376 | 354 | 330 | 346 | 365 | 404 | 394 | 358 | 417 | 368 | 348 |
| for all | Any sg | – | 665 | 631 | 669 | 612 | 620 | 643 | 676 | 674 | 650 | 668 | 602 | 622 | 703 | 544 | 592 | 651 |
| | None | 8977 | 8995 | 9028 | 8994 | 9057 | 9050 | 8981 | 8970 | 8996 | 9004 | 8967 | 8994 | 8984 | 8939 | 9039 | 9040 | 9001 |
| | Any rej., % | 10.2 | 10.1 | 9.7 | 10.1 | 9.4 | 9.5 | 10.2 | 10.3 | 10.0 | 10.0 | 10.3 | 10.1 | 10.2 | 10.6 | 9.6 | 9.6 | 10.0 |
| $\theta = \theta^*$ | All | 8038 | 5081 | 5009 | 4980 | 5182 | 4855 | 4874 | 5093 | 4759 | 4885 | 5162 | 5462 | 5142 | 4619 | 5215 | 4921 | 4797 |
| for all | Any sg | – | 2493 | 2361 | 2461 | 2229 | 2506 | 2385 | 2223 | 2717 | 2487 | 2091 | 1818 | 2194 | 2668 | 2018 | 2259 | 2525 |
| | None | 1962 | 2426 | 2630 | 2559 | 2589 | 2639 | 2741 | 2684 | 2524 | 2628 | 2747 | 2720 | 2664 | 2713 | 2767 | 2820 | 2678 |
| | Any rej., % | 80.4 | 75.7 | 73.7 | 74.4 | 74.1 | 73.6 | 72.6 | 73.2 | 74.8 | 73.7 | 72.5 | 72.8 | 73.4 | 72.9 | 72.3 | 71.8 | 73.2 |
| $\theta_S = 1.5\theta^*$ | All | 6134 | 1486 | 2699 | 1951 | 2971 | 1879 | 2071 | 1893 | 2433 | 2210 | 2509 | 2895 | 2744 | 2408 | 2944 | 2979 | 2908 |
| for 50% | Correct sg | – | 5898 | 0 | 0 | 0 | 2278 | 2037 | 2622 | 0 | 1796 | 369 | 604 | 0 | 1180 | 100 | 210 | 0 |
| | Other sg | – | 41 | 3810 | 4154 | 2835 | 2484 | 2844 | 1784 | 3496 | 2369 | 3376 | 1894 | 3030 | 2504 | 2890 | 2110 | 2668 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 3537 | 4043 | 2500 | 2245 | 2774 | 1201 | 2545 | 1456 | 2764 | 1232 | 2379 | 1080 | 2065 | 950 | 1732 |
| | None | 3866 | 2575 | 3491 | 3895 | 4194 | 3359 | 3048 | 3701 | 4071 | 3625 | 3746 | 4607 | 4226 | 3908 | 4066 | 4701 | 4424 |
| | Any rej., % | 61.3 | 74.2 | 65.1 | 61.0 | 58.1 | 66.4 | 69.5 | 63.0 | 59.3 | 63.8 | 62.5 | 53.9 | 57.7 | 60.9 | 59.3 | 53.0 | 55.8 |
| | $\theta^*$-power, % | 0.0 | 59.0 | 35.4 | 40.4 | 25.0 | 45.2 | 48.1 | 38.2 | 25.4 | 32.5 | 31.3 | 18.4 | 23.8 | 22.6 | 21.6 | 11.6 | 17.3 |
| $\theta_S = 2\theta^*$ | All | 7864 | 1196 | 3565 | 2213 | 3937 | 1677 | 1889 | 1559 | 3055 | 2115 | 2911 | 3122 | 3117 | 2495 | 3482 | 3547 | 3475 |
| for 50% | Correct sg | – | 7880 | 0 | 0 | 0 | 4074 | 3681 | 4582 | 0 | 3414 | 987 | 1640 | 0 | 2562 | 339 | 712 | 0 |
| | Other sg | – | 12 | 4787 | 5849 | 3712 | 2777 | 3142 | 2225 | 4811 | 2725 | 4250 | 2791 | 4578 | 2991 | 4007 | 3059 | 3962 |
| | $\theta_{\hat{S}} \geq \theta^*$ | – | 0 | 4667 | 5805 | 3455 | 2767 | 3100 | 2162 | 4615 | 2700 | 4228 | 2620 | 4471 | 2977 | 4000 | 2806 | 3777 |
| | None | 2136 | 912 | 1648 | 1938 | 2351 | 1472 | 1288 | 1634 | 2134 | 1746 | 1852 | 2447 | 2305 | 1952 | 2172 | 2682 | 2563 |
| | Any rej., % | 78.6 | 90.9 | 83.5 | 80.6 | 76.5 | 85.3 | 87.1 | 83.7 | 78.7 | 82.5 | 81.5 | 75.5 | 77.0 | 80.5 | 78.3 | 73.2 | 74.4 |
| | $\theta^*$-power, % | 78.6 | 90.8 | 82.3 | 80.2 | 73.9 | 85.2 | 86.7 | 83.0 | 76.7 | 82.3 | 81.3 | 73.8 | 75.9 | 80.3 | 78.2 | 70.7 | 72.5 |

Table 4.7: Use of SHAPES with a continuous endpoint, n=200, $\alpha = 0.10$, $\alpha_F = 0.05$, $\theta^* = 0.30$.

Table 4.7 shows results from the simulations. Type I error is maintained in this setting, with most rejections being for a subgroup. Under homogeneous effects of $\theta = \theta^*$, we see lower rejection rates compared to the full population analysis with the greatest loss of power of 8.6% and an average loss of power across the definitions when using $(K, \ell) = (4, 3)$ of 7.3 %. Under heterogeneous effects, definition A has the highest rates of correct subgroup rejection at 59.0 and 78.8% for $\theta_S = \{1.5\theta^*, 2\theta^*\}$, respectively. Definitions B and C do best with the setting of $(K, \ell) = (2, 2)$ with rates of 20.4 and 26.2%, respectively, when $\theta_S = 1.5\theta^*$ and rates of 36.8 and 45.8%, respectively, when $\theta_S = 2\theta^*$. Regarding rates of any rejection and $\theta^*$-power, there is improvement for small values of $K$ but with increasing values of $K$ we tend to perform worse for both values of $\theta_S$. In effect, this represents the cost of exploring too much.

Generally, we observe results using a continuous endpoint are similar to a binary endpoint. In this section we have not considered different values of $\epsilon$, but it is likely that the results from Section 4.5 are relevant and we can expect the need for appropriate tuning of the methods based on the baseline rate.

### *4.8 Increasing $n$*

We next explore the effect of increasing $n$ and consider two settings: 1) holding $\theta^*$ fixed at 0.146 and 2) holding the power fixed at 80% while changing $\theta^*$. Table 4.9 contains the results from the first setting where critical values are updated based on the $n$. Correct group identification rates improve with $n$ and type I error control is maintained. At $n = 1000$, under homogeneous effects the overall rejection rate is 100.0% with an average correct full population identification rate of 82%. Under heterogeneous effects with $\theta_S = 1.5\theta^*$ the average rate of correct subgroup identification for the recoverable definitions is 65.9% and when $\theta_S = 2\theta^*$ it is 87.0% with 92.1% for definition A.

Table 4.8 shows results from the second setting. With a nominal type I error rate of $\alpha = 0.10$ and power 80%, we set $\theta^* = 0.095$ with $n = 500$ and $\theta^* = 0.064$ with $n = 1000$. Rates of type I error, any rejection, $\theta^*$-power remain similar as $n$ increases. However, under heterogeneous effects with $\theta_S = 2\theta^*$, correct subgroup identification for definitions B and C drops from 21.6 and 19.9% to 16.4 and 17.6%, respectively. With definition A, we see a 5% improvement in correct identification with $n = 500$ compared to the other values.

### *4.9 Summary and Comparison of Logic Regression and SHAPES*

We begin by summarizing the primary results with applying SHAPES to the identification of subgroups.

| | | $n = 200$ | | | | $n = 500$ | | | | $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 335 | 353 | 340 | 332 | 318 | 354 | 396 | 359 | 328 | 349 | 371 | 347 |
| for all | Any sg | 725 | 663 | 679 | 647 | 666 | 668 | 638 | 625 | 734 | 634 | 598 | 628 |
| | None | 8940 | 8984 | 8981 | 9021 | 9016 | 8978 | 8966 | 9016 | 8938 | 9017 | 9031 | 9025 |
| | **Any rej.,** % | 10.6 | 10.2 | 10.2 | 9.8 | 9.8 | 10.2 | 10.3 | 9.8 | 10.6 | 9.8 | 9.7 | 9.8 |
| $\theta = \theta^*$ | All | 4248 | 4353 | 4441 | 4424 | 7494 | 7405 | 7340 | 7555 | 8145 | 8333 | 8346 | 8320 |
| for all | Any sg | 3315 | 3186 | 2985 | 3118 | 2188 | 2308 | 2376 | 2143 | 1850 | 1664 | 1652 | 1669 |
| | None | 2437 | 2461 | 2574 | 2458 | 318 | 287 | 284 | 302 | 5 | 3 | 2 | 11 |
| | **Any rej.,** % | 75.6 | 75.4 | 74.3 | 75.4 | 96.8 | 97.1 | 97.2 | 97.0 | 100.0 | 100.0 | 100.0 | 99.9 |
| $\theta = 1.5\theta^*$ | All | 1931 | 1952 | 2067 | 2314 | 2025 | 2280 | 2118 | 3345 | 725 | 1154 | 931 | 2155 |
| for 50% | Correct sg | 1754 | 981 | 899 | 0 | 4975 | 2879 | 3015 | 0 | 7683 | 6174 | 5924 | 0 |
| | Other sg | 3015 | 3642 | 3606 | 4149 | 2368 | 4154 | 4063 | 5755 | 1578 | 2636 | 3103 | 7799 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 1762 | 2435 | 2698 | 2665 | 1731 | 3297 | 3352 | 4216 | 1321 | 2165 | 2612 | 5036 |
| | None | 3300 | 3425 | 3428 | 3537 | 632 | 687 | 804 | 900 | 14 | 36 | 42 | 46 |
| | **Any rej.,** % | 67.0 | 65.8 | 65.7 | 64.6 | 93.7 | 93.1 | 92.0 | 91.0 | 99.9 | 99.6 | 99.6 | 99.5 |
| | $\theta^*$**-power,** % | 35.2 | 34.2 | 36.0 | 26.6 | 67.1 | 61.8 | 63.7 | 42.2 | 90.0 | 83.4 | 85.4 | 50.4 |
| $\theta = 2\theta^*$ | All | 1827 | 1949 | 2120 | 2571 | 978 | 1300 | 1028 | 2452 | 135 | 262 | 197 | 889 |
| for 50% | Correct sg | 3438 | 2158 | 1990 | 0 | 7320 | 5512 | 5364 | 0 | 9211 | 8689 | 8213 | 0 |
| | Other sg | 3331 | 4344 | 4357 | 5706 | 1660 | 3136 | 3526 | 7443 | 654 | 1049 | 1590 | 9111 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 3284 | 4247 | 4233 | 5385 | 1660 | 3120 | 3512 | 7381 | 654 | 1045 | 1590 | 9095 |
| | None | 1404 | 1549 | 1533 | 1723 | 42 | 52 | 82 | 105 | 0 | 0 | 0 | 0 |
| | **Any rej.,** % | 86.0 | 84.5 | 84.7 | 82.8 | 99.6 | 99.5 | 99.2 | 99.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | $\theta^*$**-power,** % | 75.5 | 83.5 | 83.4 | 79.6 | 99.6 | 99.3 | 99.0 | 98.3 | 100.0 | 100.0 | 100.0 | 99.8 |

Table 4.8: SHAPES with increasing sample sizes for $\theta^* = 0.146$ with $K = 4$, $\ell = 3$, and $\alpha_F = 0.05$.

| | | $\theta^* = 0.146$ $n = 200$ | | | | $\theta^* = 0.095$ $n = 500$ | | | | $\theta^* = 0.064$ $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | Decision | A | B | C | Z | A | B | C | Z | A | B | C | Z |
| $\theta = 0$ | All | 335 | 353 | 340 | 332 | 318 | 354 | 396 | 359 | 328 | 349 | 371 | 347 |
| for all | Any sg | 725 | 663 | 679 | 647 | 666 | 668 | 638 | 625 | 734 | 634 | 598 | 628 |
| | None | 8940 | 8984 | 8981 | 9021 | 9016 | 8978 | 8966 | 9016 | 8938 | 9017 | 9031 | 9025 |
| | **Any rej.,** % | 10.6 | 10.2 | 10.2 | 9.8 | 9.8 | 10.2 | 10.3 | 9.8 | 10.6 | 9.8 | 9.7 | 9.8 |
| $\theta = \theta^*$ | All | 4248 | 4353 | 4441 | 4424 | 5279 | 5385 | 5296 | 5406 | 5098 | 5239 | 5227 | 5107 |
| for all | Any sg | 3315 | 3186 | 2985 | 3118 | 2481 | 2424 | 2454 | 2339 | 2482 | 2138 | 2196 | 2331 |
| | None | 2437 | 2461 | 2574 | 2458 | 2240 | 2191 | 2250 | 2255 | 2420 | 2623 | 2577 | 2562 |
| | **Any rej.,** % | 75.6 | 75.4 | 74.3 | 75.4 | 77.6 | 78.1 | 77.5 | 77.4 | 75.8 | 73.8 | 74.2 | 74.4 |
| $\theta = 1.5\theta^*$ | All | 1931 | 1952 | 2067 | 2314 | 2359 | 2588 | 2461 | 2903 | 2193 | 2502 | 2536 | 2794 |
| for 50% | Correct sg | 1754 | 981 | 899 | 0 | 2000 | 760 | 866 | 0 | 1789 | 635 | 738 | 0 |
| | Other sg | 3015 | 3642 | 3606 | 4149 | 2376 | 3541 | 3198 | 3502 | 2539 | 3381 | 2782 | 3331 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 1762 | 2435 | 2698 | 2665 | 1474 | 2638 | 2557 | 2636 | 1600 | 2567 | 2226 | 2626 |
| | None | 3300 | 3425 | 3428 | 3537 | 3265 | 3111 | 3475 | 3595 | 3479 | 3482 | 3944 | 3875 |
| | **Any rej.,** % | 67.0 | 65.8 | 65.7 | 64.6 | 67.4 | 68.9 | 65.2 | 64.0 | 65.2 | 65.2 | 60.6 | 61.2 |
| | $\theta^*$**-power,** % | 35.2 | 34.2 | 36.0 | 26.6 | 34.7 | 34.0 | 34.2 | 26.4 | 33.9 | 32.0 | 29.6 | 26.3 |
| $\theta = 2\theta^*$ | All | 1827 | 1949 | 2120 | 2571 | 2173 | 2663 | 2467 | 3317 | 2222 | 2897 | 2658 | 3326 |
| for 50% | Correct sg | 3438 | 2158 | 1990 | 0 | 3943 | 1919 | 2040 | 0 | 3462 | 1637 | 1756 | 0 |
| | Other sg | 3331 | 4344 | 4357 | 5706 | 2538 | 4092 | 3955 | 4997 | 2734 | 3826 | 3670 | 4744 |
| | $\theta_{\hat{S}} \geq \theta^*$ | 3284 | 4247 | 4233 | 5385 | 2531 | 4046 | 3892 | 4867 | 2722 | 3781 | 3612 | 4636 |
| | None | 1404 | 1549 | 1533 | 1723 | 1346 | 1326 | 1538 | 1686 | 1582 | 1640 | 1916 | 1930 |
| | **Any rej.,** % | 86.0 | 84.5 | 84.7 | 82.8 | 86.5 | 86.7 | 84.6 | 83.1 | 84.2 | 83.6 | 80.8 | 80.7 |
| | $\theta^*$**-power,** % | 85.5 | 83.5 | 83.4 | 79.6 | 86.5 | 86.3 | 84.0 | 81.8 | 84.1 | 83.2 | 80.3 | 79.6 |

Table 4.9: Using with increasing sample sizes and varying $\theta^*$ to reflect 80% power with $K = 4$, $\ell = 3$, and $\alpha_F = 0.05$.

- The tuning parameter $\ell$ provides the maximum number of covariates allowed in a subgroup definition. Selecting the $\ell$ that corresponds to the true number of variables in the subgroup definition improves correct subgroup identification. Selecting a value greater than the true number results in decreased performance, but performance does not typically drop to zero.
- The metric $\theta^*$-power is driven by $\theta_S$ and $\zeta$. Larger effects in the subgroup increase $\theta^*$-power, which does not necessarily indicate identification of the correct group.
- When the subgroup effect is fixed, we are most successful identifying subgroups of prevalence $\zeta = 0.50$. We observe higher rejection rates compared to the full population when $\zeta \leq 0.50$, suggesting this is an area where it is worth searching for subgroups.
- SHAPES is able to handle the presence of covariates that are prognostic, with little increase in the type I as prognostic effects are increased. However, we only evaluate two scenarios so further research is needed to confirm this result.
- Changing the baseline response rate influences power. This influence may be different depending on subgroup definition.

Overall, the implementation of SHAPES requires consideration of several factors, similar to LR. Having evaluated the performance of LR and SHAPES, we now compare performance of the two methods.

- The main difference between LR and SHAPES is the decreased number of possible subgroups with SHAPES. This translates into better performance when the subgroup definition is part of the SHAPES collection, particularly with moderate $K$.
- Because SHAPES directly uses the p-value to select a candidate subgroup compared to the likelihood with LR, we typically identify lower p-values and critical values. This requires the selection of a higher $\alpha_F$ with SHAPES than with LR to ensure a balanced performance under heterogeneous and homogeneous effects.
- With LR, we sometimes get higher rates of correct subgroup identification, but performance suffers when adding in noisy covariates, getting $n_{leaves}$ incorrect or under prognostic effects. Suggesting LR is more sensitive to model misspecification than SHAPES.
- We observe similar rates of overall rejection for the two methods. With differences in some of our measures, like $\theta^*$-power more driven by simulation scenario differences.

- When the subgroup effect is fixed, both methods identify $\zeta = 0.50$ as a sweet spot where correct subgroup identification is highest. This result was not apparent with SIDES and ASD as those methods were evaluated on a reduced range of prevalences.

In many ways, performance is similar for the two methods. In some critical ways, it is different. While determining which method to implement requires consideration of these differences, in the remainder of this dissertation we focus on further evaluation of SHAPES. This is motivated by the inflation of type I error with LR when there are prognostic effects, which we expect to occur often, and the sensible restrictions of SHAPES providing some improvements in performance when the tuning parameter value is not the same as the truth.

# Chapter 5

# SHAPES AND THE DRUG DISCOVERY PROCESS

The usefulness of exploring for predictive subgroups depends on many factors, in particular the prevalence of heterogeneous effects in the collection of drugs under study and the availability and application of the subgroup defining covariates. In Chapter 4 our assumptions about the true state of nature were incorporated only indirectly through the use of $\ell$ and $\alpha_F$ and, for most results, we fixed $\alpha_F$ to be 0.05 and explored performance for specific values of $\ell$ for the null and several alternative scenarios. In this chapter we focus on performance when testing a collection of drugs where we assume some composition and prevalence of heterogeneous effects along with null and homogeneous effects. We believe this is the first time prior beliefs have been incorporated into performance evaluation. We begin by considering variation in the optimal values for $\alpha_F$ under various scenarios and then describe performance across phases when exploring for subgroups, with the purpose of describing situations where subgroup searching is and is not beneficial.

## *5.1 The General Set-up*

In planning a clinical trial, an investigator aims to maintain a specified type I error while enrolling some $n$ patients to ensure power to detect some overall alternative effect, with this effect possibly representing an average effect across multiple subgroups. In applying SHAPES to searching for subgroups, performance depends on: 1) the nature of heterogeneous and homogeneous effects in the collection of drugs under study and 2) the optimality criterion and 3) the tuning parameters ($\alpha$, $\alpha_F$ and $\ell$) and utility function.

For the first point, we now describe a framework to relate an investigator's prior belief of the likelihood of several hypotheses consisting of null, heterogeneous and homogeneous effects. We consider five probabilities that take on values between 0 and 1: $\pi_N$, $\pi_{S1}$, $\pi_{S2}$, $\pi_{F1}$, and $\pi_{F2}$. They represent, respectively, a null effect ($\pi_N = Pr(\theta = \theta_S = 0)$), a moderate subgroup effect ($\pi_{S1} = Pr(\theta_S = 1.5\theta^*)$), a strong subgroup effect ($\pi_{S2} = Pr(\theta_S = 2\theta^*)$), a weak full population effect ($\pi_{F1} = Pr(\theta = 0.75\theta^*)$) and the targeted full population effect ($\pi_{F2} = Pr(\theta = \theta^*)$). The marginal effect for $\pi_{S1}$ with $\zeta = 0.50$ is $0.75\theta^*$, which is the same as the marginal effect for $\pi_{F2}$, and the marginal effect for $\pi_{S2}$ with $\zeta = 0.50$ is $\theta^*$, which is the same as $\pi_{F2}$. We require the constraint that

either:

$$\pi_N + \pi_{S1} + \pi_{S2} + \pi_{F1} + \pi_{F2} = 1$$

or when the statistics do not depend on the prevalence of the null:

$$\pi_{S1} + \pi_{S2} + \pi_{F1} + \pi_{F2} = 1.$$

Additionally, we sometimes set $\pi_{S1} = \pi_{F1} = 0.00$ thereby reducing our collection of probabilities to only three. We also refer to this collection of probabilities as our prior distribution. This set-up can be thought of as a simplified Bayesian analysis with just a limited number of parameters in the prior. A true Bayesian analysis would allow non-zero probability to be assigned to a greater number of subgroups and corresponding effects. The collection of priors has two possible interpretations. First, under a simulation study with replications where the true effect is randomly selected, these probabilities reflect the true state of nature. Second, applying to a specific drug under investigation the probabilities represent the investigator's belief about the likelihood of each hypothesis being correct. These priors are not exhaustive but just one possible collection of interest. Additional alternatives, e.g. a harmful effect of treatment on the primary endpoint, could be included.

As the performance of SHAPES varies by subgroup definition, we consider three collections or *universes* of predictive subgroup effects. The universes are composed of definitions A, B, C and Z, as defined in Table 3.9. Universe I consists of only subgroups with definition A, i.e. a single variable among the K correctly identify the subgroup. Universe II consists of equal parts of the recoverable definitions A, B, and C. Universe III is equal parts of all four. The universes reflect increasing complexity of the mixture of predictive effects.

For the second point, we consider four metrics: the frequency of correct group identification (both subgroup and full population), $\theta^*$-power, any rejection and positive predictive value. The first three are defined in Section 2.5 and the latter is simply the proportion of trials where the null is rejected and there is some true positive effect on average for the selected population. To calculate these metrics under a prior distribution, we first determine the value of the metric on each one of the five effect scenarios associated with a prior probability. We then calculate the weighted linear combination of these metrics where the weights are the prior probabilities, which is a weighted average.

For the third point, the decisions we make depend on $\alpha_F$, the apportioning of $(\alpha - \alpha_F)$ equally in a cumulative sense to each depth and the *minimum p-value* utility function. We focus here on using a range of values of $\alpha_F$. We ultimately aim to examine how SHAPES performs based on the

metrics when applied to the universes as $\alpha_F$ varies.

### *5.2 Prior Collection 1 with Subgroups of* $\zeta = 0.50$

In this section we assume for $\pi_{S1}$ and $\pi_{S2}$ that $\zeta = 0.50$. This is likely false in the real world, but as discussed in Section 4.4 correct subgroup identification is highest when $\zeta = 0.50$ with roughly comparable performance between 0.40 and 0.60. We implement settings for SHAPES of $K = 4$ and $\alpha = 0.10$ while exploring values of $\alpha_F$ ranging from 0 to 0.10 by 0.0025 and $\ell = \{0, 1, 2, 3\}$ with the *minimum p-value* decision rule. The full population is denoted by either $\alpha_F = 0.10$ or $\ell = 0$.

We first consider rates of correct subgroup identification for values of $\alpha_F$ from 0.00 to 0.10 for the three universes with prior beliefs that are splits between $\pi_{S2}$ and $\pi_{F2}$, specifically:

$$(\pi_{S2}, \pi_{F2}) = \{(1.00, 0.00), (0.75, 0.25), (0.50, 0.50), (0.25, 0.75), (0.00, 1.00)\}.$$

These five priors describe a range of beliefs from only strong subgroups effects (where $\theta_S = 2\theta^*$) to only the MCID in the full population (where $\theta = \theta^*$).

Figure 5.1 shows the results. The columns in this figure represent the universes and the rows the values of $\ell$. For each plot, the x-axis shows $\alpha_F$ from 0.0 to 0.10 or from only testing for subgroups to only testing for the full population. The y-axis is the percent of correct group identification. Each line on the plot corresponds to specific values of a prior. Under most priors, the rate of correct group identification is increasing in $\alpha_F$ with a maximum at either $\alpha_F = 0.0975$ or 0.1000. When $\pi_{S2} \geq 50\%$ we often observe the maximum at $\alpha_F = 0.0975$ with a sharp decrease to 0.1000. This decrease disappears when $\pi_{S2} < 50\%$. The only prior with a decreasing trend in $\alpha_F$ is $\pi_{S2} = 1.00$, where correct identification is highest with $\alpha_F = 0.0000$ and we only explore for subgroups and never consider a full population analysis. Interestingly, however, this is not always true. For example, under universe I with $\ell = 1$ and $(\pi_{S2}, \pi_{F2}) = (1.00, 0.00)$, the maximum occurs at $\alpha_F = 0.0475$ with a rate of 38.4% (note the rate is essentially flat between $\alpha_F = 0$ and 0.0600). This is likely due to a combination of imprecision with the critical values which are derived for each value of $\alpha_F$ and how $(\alpha - \alpha_F)$ is shared among the subgroup depths.

For all universes and values of $\ell$ we observe an equivalence point where the lines cross. This point occurs at the value of $\alpha_F$ where the rate of correct group identification is the same for all priors. An investigator might chose this value if they were uncertain as to the prior and wanted to ensure consistent performance across all priors; this may be reasonable for $\ell = 1$ and universe I but with more exploration the equivalence point may not be reasonable as it relates to an approximate 20% rate of correct identification.

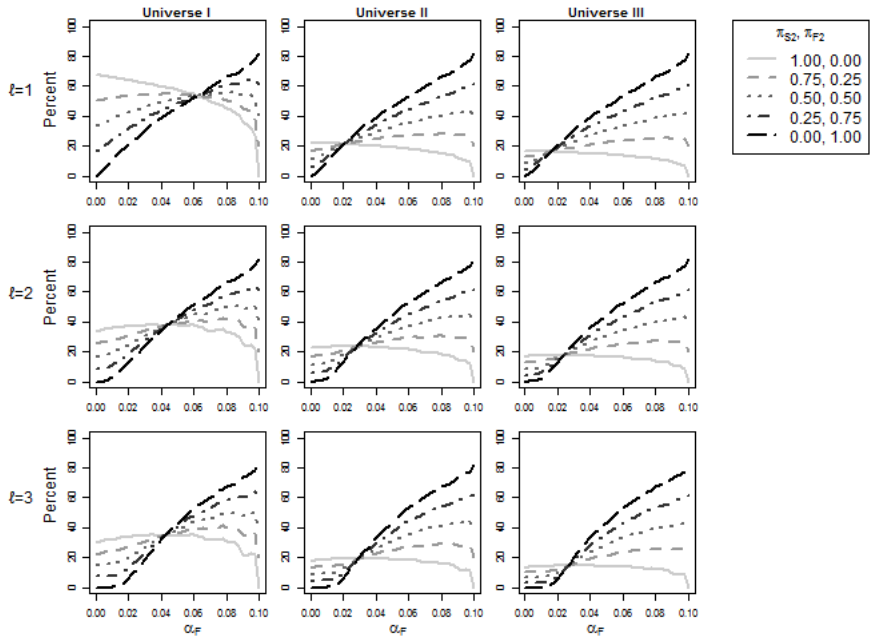



Figure 5.1: Correct group identification with varying $\alpha_F$ for the three universes with $\ell = \{1, 2, 3\}$ and $K = 4$.

We also see that the relationship between $\ell$ and the universe is complicated. With universe I and a prior with non-zero $\pi_{S2}$, there is a decrease in rates as $\ell$ increases. Here the match between $\ell$ and the truth is critical. For the other universes performance changes little as $\ell$ increases.

We next consider the *maximum* frequency of correct group identification, the composition thereof in terms of the full population and subgroups, and the corresponding value of $\alpha_F$, reported in Table 5.1. Here we use the same settings as in Figure 5.1 as this table essentially reports the maximum of each line from the figure. So the table represents the best case scenario. The 'Overall' column reports the frequency of correct group identification for various values of $(\pi_{S2}, \pi_{F2})$, which is a weighted average of the 'Sg' and 'Full' columns.

Performance depends on the prior. If there are only subgroups ($\pi_{S2} = 1$) we improve for all scenarios using $\ell \geq 1$ compared to $\ell = 0$, which will always have a rate of 0%. With this prior, rates (and hence improvements) range from 14.7% with universe III and $\ell = 3$ to 67.3% with universe I and $\ell = 1$. Under a prior with $\pi_{S2} = \pi_{F2} = 0.50$, with $\ell = 1$ and $\alpha_F = 0.800$ compared to $\ell = 0$ there is a 14.8% increase in correct identification from 40.9 to 55.7%. This increase is a combination of a 45.1% rate of correct subgroup identification and a 66.3% rate of correct full population identification. With

| | | | Universe I | | | | Universe II | | | | Universe III | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_{S2}$ | $\pi_{F2}$ | Overall | Sg | Full | $\alpha_F$ | Overall | Sg | Full | $\alpha_F$ | Overall | Sg | Full | $\alpha_F$ |
| 0 | 1.00 | 0.00 | 0.00 | 0.00 | – | 0.1000 | 0.00 | 0.00 | – | 0.1000 | 0.00 | 0.00 | – | 0.1000 |
| 0 | 0.75 | 0.25 | 20.45 | 0.00 | 81.78 | 0.1000 | 20.38 | 0.00 | 81.52 | 0.1000 | 20.42 | 0.00 | 81.68 | 0.1000 |
| 0 | 0.50 | 0.50 | 40.89 | 0.00 | 81.78 | 0.1000 | 40.76 | 0.00 | 81.52 | 0.1000 | 40.84 | 0.00 | 81.68 | 0.1000 |
| 0 | 0.25 | 0.75 | 61.34 | 0.00 | 81.78 | 0.1000 | 61.14 | 0.00 | 81.52 | 0.1000 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 0 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |
| 1 | 1.00 | 0.00 | 67.31 | 67.31 | – | 0.0000 | 22.44 | 22.44 | – | 0.0000 | 16.83 | 16.83 | – | 0.0000 |
| 1 | 0.75 | 0.25 | 54.72 | 59.34 | 40.86 | 0.0425 | 27.93 | 14.31 | 68.79 | 0.0850 | 25.17 | 10.73 | 68.49 | 0.0850 |
| 1 | 0.50 | 0.50 | 55.68 | 45.10 | 66.26 | 0.0800 | 43.25 | 8.97 | 77.52 | 0.0975 | 41.97 | 6.73 | 77.20 | 0.0975 |
| 1 | 0.25 | 0.75 | 65.22 | 26.92 | 77.99 | 0.0975 | 61.14 | 0.00 | 81.52 | 0.1000 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 1 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |
| 2 | 1.00 | 0.00 | 38.35 | 38.35 | – | 0.0475 | 23.55 | 23.55 | – | 0.0175 | 17.66 | 17.66 | – | 0.0175 |
| 2 | 0.75 | 0.25 | 42.66 | 35.21 | 65.01 | 0.0775 | 30.81 | 19.45 | 64.88 | 0.0775 | 27.11 | 14.59 | 64.67 | 0.0775 |
| 2 | 0.50 | 0.50 | 50.34 | 31.84 | 68.85 | 0.0850 | 44.77 | 11.87 | 77.67 | 0.0975 | 43.12 | 8.90 | 77.34 | 0.0975 |
| 2 | 0.25 | 0.75 | 64.26 | 22.23 | 78.27 | 0.0975 | 61.22 | 11.87 | 77.67 | 0.0975 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 2 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |
| 3 | 1.00 | 0.00 | 35.40 | 35.40 | – | 0.0350 | 19.54 | 19.54 | – | 0.0350 | 14.65 | 14.65 | – | 0.0350 |
| 3 | 0.75 | 0.25 | 41.23 | 33.10 | 65.64 | 0.0775 | 29.22 | 15.74 | 69.66 | 0.0850 | 26.23 | 11.80 | 69.53 | 0.0850 |
| 3 | 0.50 | 0.50 | 50.09 | 21.66 | 78.51 | 0.0975 | 44.51 | 11.18 | 77.83 | 0.0975 | 42.94 | 8.39 | 77.48 | 0.0975 |
| 3 | 0.25 | 0.75 | 64.30 | 21.66 | 78.51 | 0.0975 | 61.17 | 11.18 | 77.83 | 0.0975 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 3 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |

Table 5.1: Maximum rates of correct group identification for various prior distributions and universes of true drug effects.

universe II the improvement is 4.0%: an increase from 40.8 with $\ell = 0$ to 44.8% with $\ell = 2$ and $\alpha_F = 0.0975$, which consists of an 11.9% rate of correct subgroup identification and 77.7% rate of correct full population identification.

When the true subgroup generation mechanism is either more complicated (universe II or III) or the presence of homogeneous effects dominates ($\pi_{F2} \in \{0.75, 1.00\}$) we often select values of $\alpha_F$ that correspond to a rate of correct identification driven by rejection for the correct full population. This observation reflects the inherent cost of searching for subgroups which manifests as lower rates of correct subgroup identification. With the prior of $(\pi_{S2}, \pi_{F2}) = (0.25, 0.75)$ we sometimes improve by selecting $\alpha_F = 0.0975$ instead of $\alpha_F = 0.1000$. For example with universe I and $\ell = 2$, we see a slight improvement to 64.3 from 61.3%. There is a similar trend with universe II, but the differences are less than one percent.

Overall, we conclude from Figure 5.1 and Table 5.1 there is no single best value of $\alpha_F$ that always yields the maximum correct group identification. The performance of any particular $\alpha_F$ depends on the nature of the universe and the selected value of $\ell$. With universes II and III, we typically do best selecting $\alpha_F = 0.0975$ or 0.1000, implying the cost for searching for subgroups is too high to yield much benefit for correct identification. The only time the cost is relatively low is when we have $\ell = 1$ and universe I, where we do not consider many possible subgroups and the value of $\ell$ agrees with the true data generating mechanism. In other words, the costs of exploring for subgroups quickly add up in this setting of priors.

We next consider selection rates of a group with $\theta_{\hat{S}} \geq \theta^*$, shown in Table 5.2. When selecting the full population only ($\ell = 0$) we maintain the specified power of 80% because under either $\pi_{S2}$ or $\pi_{F2}$ the marginal effect is fixed at $\theta = \theta^*$. If $\pi_{S2} > 0$, we can achieve a higher rate of $\theta^*$-power implementing SHAPES with a specific $\alpha_F$. For example under universe I with $\pi_{S2} = 1$ and $\ell = 1$ we can identify a group with $\theta^*$ or better 5.5% more often setting $\alpha_F = 0.0600$ compared to $\ell = 0$. Under universes II and III with $\ell = 1$, maximum improvements are 3.6 and 2.8%, respectively, when setting $\alpha_F = 0.0775$. There are additional small gains for larger $\ell$, suggesting for these universes and for $\theta^*$-power getting $\ell$ correct is not important in contrast with correct group identification. Furthermore, as we decrease $\pi_{S2}$ and increase $\pi_{F2}$ we do better increasing $\alpha_F$ until under the $\pi_{F2} = 1$ prior, we do best setting $\alpha_F = 0.1000$, and thereby avoid any cost of exploration.

When we compare values of $\alpha_F$ that correspond to the best performance, we see that there is no optimal value for each metric. For correct identification unless an investigator's belief in only subgroups is strong, in most cases it is better to choose $\alpha_F$ close to $\alpha$ but for $\theta^*$-power a value in between 0 and $\alpha$ is better.

| | | | Universe I | | Universe II | | Universe III | |
|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_{S2}$ | $\pi_{F2}$ | Rate | $\alpha_F$ | Rate | $\alpha_F$ | Rate | $\alpha_F$ |
| 0 | 1.00 | 0.00 | 82.23 | 0.1000 | 81.72 | 0.1000 | 81.66 | 0.1000 |
| 0 | 0.75 | 0.25 | 82.12 | 0.1000 | 81.67 | 0.1000 | 81.66 | 0.1000 |
| 0 | 0.50 | 0.50 | 82.00 | 0.1000 | 81.62 | 0.1000 | 81.67 | 0.1000 |
| 0 | 0.25 | 0.75 | 81.89 | 0.1000 | 81.57 | 0.1000 | 81.67 | 0.1000 |
| 0 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 1 | 1.00 | 0.00 | 87.69 | 0.0600 | 85.27 | 0.0775 | 84.44 | 0.0775 |
| 1 | 0.75 | 0.25 | 85.47 | 0.0775 | 83.86 | 0.0775 | 83.28 | 0.0775 |
| 1 | 0.50 | 0.50 | 83.70 | 0.0775 | 82.53 | 0.0975 | 82.44 | 0.0975 |
| 1 | 0.25 | 0.75 | 82.48 | 0.0975 | 81.88 | 0.0975 | 81.94 | 0.0975 |
| 1 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 2 | 1.00 | 0.00 | 86.90 | 0.0425 | 85.64 | 0.0775 | 84.92 | 0.0775 |
| 2 | 0.75 | 0.25 | 85.10 | 0.0775 | 84.12 | 0.0775 | 83.63 | 0.0775 |
| 2 | 0.50 | 0.50 | 83.42 | 0.0775 | 82.76 | 0.0975 | 82.66 | 0.0975 |
| 2 | 0.25 | 0.75 | 82.49 | 0.0975 | 82.01 | 0.0975 | 82.08 | 0.0975 |
| 2 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 3 | 1.00 | 0.00 | 86.84 | 0.0775 | 85.67 | 0.0775 | 85.05 | 0.0775 |
| 3 | 0.75 | 0.25 | 85.19 | 0.0775 | 84.16 | 0.0775 | 83.75 | 0.0775 |
| 3 | 0.50 | 0.50 | 83.55 | 0.0775 | 82.83 | 0.0975 | 82.75 | 0.0975 |
| 3 | 0.25 | 0.75 | 82.46 | 0.0975 | 82.07 | 0.0975 | 82.14 | 0.0975 |
| 3 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |

Table 5.2: Maximum rates of selection of groups with $\theta_{\hat{S}} \geq \theta^*$ for various prior distributions and universes of true drug effects with $n = 200$ and $K = 4$.

### *5.3 Prior Collection 2 with Subgroups of* $\zeta = 0.50$

We next consider a universe where both the heterogeneous and homogeneous effects are each further divided into two possible effects. Using the previous prior set-up for both subgroups and the full population, we further split each probability in half. We consider performance for the following priors (rounded to the nearest thousandth):

$$\begin{aligned}(\pi_{S1}, \pi_{S2}, \pi_{F1}, \pi_{F2}) \quad = \quad & \{(0.500, 0.500, 0.000, 0.000), (0.375, 0.375, 0.125, 0.125), \\ & (0.250, 0.250, 0.250, 0.250), (0.125, 0.125, 0.375, 0.375), \\ & (0.000, 0.000, 0.500, 0.500)\}.\end{aligned}$$

The prevalence of both full population and subgroup effects remain the same as before, but sometimes the effect is weaker. Additionally for each prior the collection of marginal effects is the same: either

$0.75\theta^*$ or $\theta^*$ each occurring half of the time. This collection of priors may represent the situation where an investigator views the previous collection of priors as optimistic and this collection as realistic to account for smaller than expected effects. Compared to the previous figure and tables, we generally anticipate lower rates of correct group identification and $\theta^*$-power, given the addition of smaller marginal effects.

Figure 5.2 shows the rates of correct group identification for the range of $\alpha_F$ for the three universes and $\ell \in \{0, 1, 2, 3\}$. The plots are similar to those in Figure 5.1, with the exception of a shift downward for most rates. Again there are equivalence points.

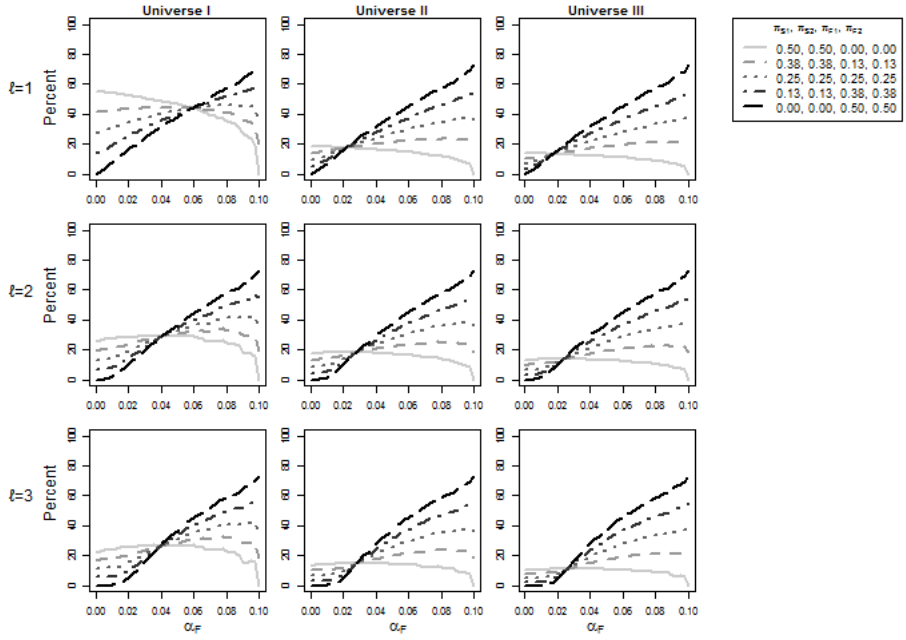



Figure 5.2: Correct group identification with varying $\alpha_F$ for the three universes with four values of the prior with $\ell = \{1, 2, 3\}$ and $K = 4$.

Table 5.4 shows the rates of maximal correct identification rates for this collection of priors. Under the prior where $\pi_{S1} = \pi_{S2} = 0.50$ with $\ell = 1$ we see the greatest improvement using SHAPES, with rates (corresponding $\alpha_F$ in parentheses): 55.2 (0.0000), 18.4 (0.0000), and 13.8% (0.0000) under universes I, II, and III, respectively, compared to a fixed rate of 0.0% with $\ell = 0$. Increasing to $\ell = 2$, the rates are 29.5 (0.0475), 18.3 (0.0225), and 13.8% (0.0225). Increasing to $\ell = 3$ the rates are 26.7 (0.0600), 14.9 (0.0350), and 11.1% (0.0350), demonstrating a large drop in performance for universe I but stable performance for universes II and III.

As we increase $\pi_{F1}$ and $\pi_{F2}$, for each universe and specific value of $\ell$ there is an increase in maximum correct group identification, and we concurrently select increasing values of $\alpha_F$. The only exception is with universe I and $\ell = 1$, where a u-shaped relationship exists between $\alpha_F$ and the rate. For the priors with $\pi_{S1} > 0$ and $\pi_{S2} > 0$, we often observe improvements over $\ell = 0$, but the differences are slight; mostly between 0 and 3%. When $\pi_{S1} = \pi_{S2} = 0$, we do best with the full population analysis.

When considering maximum $\theta^*$-power, under this prior distribution set-up, selecting the full population related to $\pi_{F1}$ correctly does not actually yield a group with $\theta \geq \theta^*$ and because of this the denominator for this statistic is changing for the various priors. This denominator is listed in the column '$\pi_{\theta^*}$' in Table 5.3 and ranges from 0.50 to 1.00. This table is answering the conditional question of when $\theta_S \geq \theta^*$, how often do we identify a group that has $\theta_{\hat{S}} \geq \theta^*$?

With a prior of only subgroup effects, $\pi_{S1} = \pi_{S2} = 0.50$, using SHAPES we achieve greater rates of $\theta^*$-power with a range from 59.4 to 67.2% for $\ell \in \{1, 2, 3\}$ compared to a range from 40.8 to 41.1% for $\ell = 0$. The corresponding values of $\alpha_F$ range from 0.0000 to 0.0300. While $\theta^*$-power increases as $\pi_{F1}$ and $\pi_{F2}$ increase, the improvement relative to an analysis with $\ell = 0$ decreases. We also select larger $\alpha_F$. Eventually, when $\pi_{F1} = \pi_{F2} = 0.50$, we do best with $\alpha_F = 0.10$.

When we compare the universes for the same prior and tuning parameter, we observe similar rates with the largest difference of 6.7% comparing universe III to I with $\ell = 2$ and the subgroups-only prior. Many differences are 3% or less. The selected values of $\alpha_F$ corresponding to the maximum tend to be less with Universe I and are always the same for Universe II and III.

### *5.4 Prior Collection 1 with Subgroups of* $\zeta = \{0.30, 0.50, 0.70\}$

In this section we expand the range of possible subgroup prevalences to $\zeta \in \{0.30, 0.50, 0.70\}$ while using prior collection 1. We use the same universe definitions as before, but for each possible subgroup definition we assume each of the prevalences one-third of the time.

Figure 5.3 displays results for the range of $\alpha_F$ for the universes and $\ell$. We observe a similar trend as with prior collection 2: overall rates of correct group identification decrease. Particularly with small values of $\alpha_F$ where we anticipate a greater presence of subgroup effects. Additionally the previously observed drop in rates from $\alpha_F = 0.0975$ to 0.1000 occurs less often, here this drop happens with universe I with $\pi_{S2} > 0.50$ and for the other universes with $\pi_{S2} = 1$.

Table 5.6 contains the maximum correct group identification rate with the corresponding $\alpha_F$ when $\zeta$ is variable. For the majority of priors and universes the best performance is when we select either $\alpha_F = 0.0975$ or 0.1000. This result is driven by the low rates of correct subgroup

| | | | | | | Universe I | | Universe II | | Universe III | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_{S1}$ | $\pi_{S2}$ | $\pi_{F1}$ | $\pi_{F2}$ | $\pi_{\theta^*}$ | Rate | $\alpha_F$ | Rate | $\alpha_F$ | Rate | $\alpha_F$ |
| 0 | 0.50 | 0.50 | 0.00 | 0.00 | 1.00 | 41.12 | 0.1000 | 40.86 | 0.1000 | 40.83 | 0.1000 |
| 0 | 0.38 | 0.38 | 0.13 | 0.13 | 0.84 | 46.92 | 0.1000 | 46.67 | 0.1000 | 46.67 | 0.1000 |
| 0 | 0.25 | 0.25 | 0.25 | 0.25 | 0.75 | 54.67 | 0.1000 | 54.41 | 0.1000 | 54.45 | 0.1000 |
| 0 | 0.13 | 0.13 | 0.38 | 0.38 | 0.63 | 65.51 | 0.1000 | 65.25 | 0.1000 | 65.34 | 0.1000 |
| 0 | 0.00 | 0.00 | 0.50 | 0.50 | 0.50 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 1 | 0.50 | 0.50 | 0.00 | 0.00 | 1.00 | 65.03 | 0.0000 | 62.95 | 0.0000 | 63.90 | 0.0000 |
| 1 | 0.38 | 0.38 | 0.13 | 0.13 | 0.84 | 65.92 | 0.0000 | 64.13 | 0.0000 | 65.00 | 0.0000 |
| 1 | 0.25 | 0.25 | 0.25 | 0.25 | 0.75 | 67.09 | 0.0000 | 65.70 | 0.0000 | 66.46 | 0.0000 |
| 1 | 0.13 | 0.13 | 0.38 | 0.38 | 0.63 | 71.06 | 0.0600 | 69.50 | 0.0775 | 69.56 | 0.0775 |
| 1 | 0.00 | 0.00 | 0.50 | 0.50 | 0.50 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 2 | 0.50 | 0.50 | 0.00 | 0.00 | 1.00 | 67.22 | 0.0050 | 62.44 | 0.0175 | 60.49 | 0.0225 |
| 2 | 0.38 | 0.38 | 0.13 | 0.13 | 0.84 | 67.84 | 0.0050 | 63.76 | 0.0225 | 62.19 | 0.0225 |
| 2 | 0.25 | 0.25 | 0.25 | 0.25 | 0.75 | 68.67 | 0.0050 | 65.83 | 0.0325 | 64.76 | 0.0325 |
| 2 | 0.13 | 0.13 | 0.38 | 0.38 | 0.63 | 71.30 | 0.0600 | 69.89 | 0.0775 | 69.71 | 0.0775 |
| 2 | 0.00 | 0.00 | 0.50 | 0.50 | 0.50 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 3 | 0.50 | 0.50 | 0.00 | 0.00 | 1.00 | 61.53 | 0.0250 | 61.39 | 0.0225 | 59.36 | 0.0300 |
| 3 | 0.38 | 0.38 | 0.13 | 0.13 | 0.84 | 63.36 | 0.0425 | 63.01 | 0.0300 | 61.37 | 0.0325 |
| 3 | 0.25 | 0.25 | 0.25 | 0.25 | 0.75 | 66.03 | 0.0425 | 65.51 | 0.0425 | 64.34 | 0.0425 |
| 3 | 0.13 | 0.13 | 0.38 | 0.38 | 0.63 | 70.76 | 0.0775 | 69.98 | 0.0775 | 69.81 | 0.0775 |
| 3 | 0.00 | 0.00 | 0.50 | 0.50 | 0.50 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |

Table 5.3: Maximum rates of selection of groups with $\theta_{\hat{S}} \geq \theta^*$ for various prior distributions and universes of true drug effects with $n = 200$ and $K = 4$.

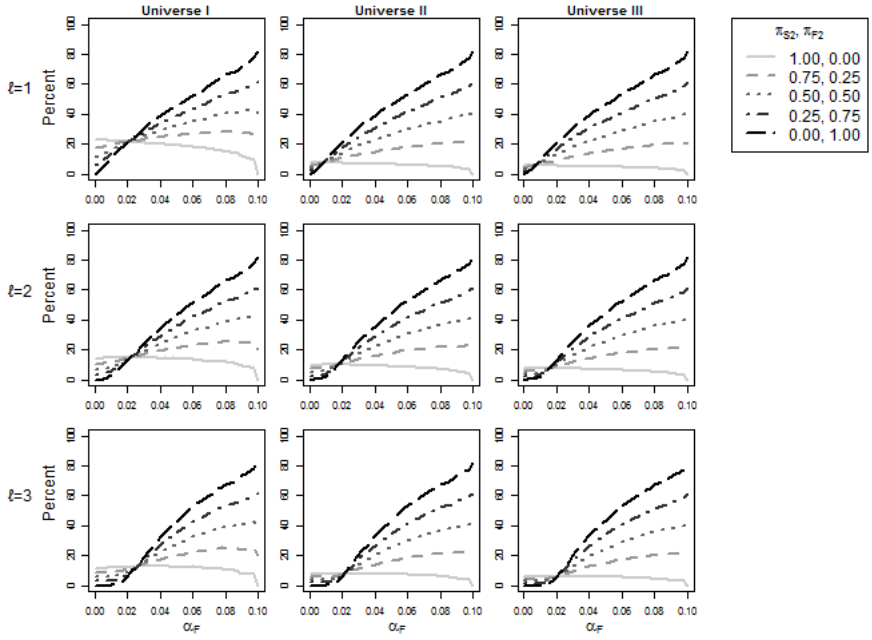



Figure 5.3: Correct group identification with $\zeta \in \{0.30, 0.50, 0.70\}$ and $\alpha_F$ for the three universes with $\ell = \{1, 2, 3\}$ and $K = 4$.

| | | | | | Universe I | | | | Universe II | | | | Universe III | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_{S1}$ | $\pi_{S2}$ | $\pi_{F1}$ | $\pi_{F2}$ | Overall | Sg | Full | $\alpha_F$ | Overall | Sg | Full | $\alpha_F$ | Overall | Sg | Full | $\alpha_F$ |
| 0 | 0.50 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | – | 0.1000 | 0.00 | 0.00 | – | 0.1000 | 0.00 | 0.00 | – | 0.1000 |
| 0 | 0.38 | 0.38 | 0.12 | 0.12 | 18.27 | 0.00 | 73.09 | 0.1000 | 18.20 | 0.00 | 72.78 | 0.1000 | 18.22 | 0.00 | 72.87 | 0.1000 |
| 0 | 0.25 | 0.25 | 0.25 | 0.25 | 36.55 | 0.00 | 73.09 | 0.1000 | 36.39 | 0.00 | 72.78 | 0.1000 | 36.43 | 0.00 | 72.87 | 0.1000 |
| 0 | 0.12 | 0.12 | 0.38 | 0.38 | 54.82 | 0.00 | 73.09 | 0.1000 | 54.59 | 0.00 | 72.78 | 0.1000 | 54.65 | 0.00 | 72.87 | 0.1000 |
| 0 | 0.00 | 0.00 | 0.50 | 0.50 | 73.09 | – | 73.09 | 0.1000 | 72.78 | – | 72.78 | 0.1000 | 72.87 | – | 72.87 | 0.1000 |
| 1 | 0.50 | 0.50 | 0.00 | 0.00 | 55.20 | 55.20 | – | 0.0000 | 18.40 | 18.40 | – | 0.0000 | 13.80 | 13.80 | – | 0.0000 |
| 1 | 0.38 | 0.38 | 0.12 | 0.12 | 44.59 | 48.26 | 33.59 | 0.0425 | 23.38 | 11.13 | 60.14 | 0.0850 | 21.27 | 5.97 | 67.20 | 0.0950 |
| 1 | 0.25 | 0.25 | 0.25 | 0.25 | 46.48 | 35.34 | 57.61 | 0.0800 | 37.83 | 6.53 | 69.14 | 0.0975 | 36.84 | 4.89 | 68.78 | 0.0975 |
| 1 | 0.12 | 0.12 | 0.38 | 0.38 | 57.09 | 19.58 | 69.59 | 0.0975 | 54.59 | 0.00 | 72.78 | 0.1000 | 54.65 | 0.00 | 72.87 | 0.1000 |
| 1 | 0.00 | 0.00 | 0.50 | 0.50 | 73.09 | – | 73.09 | 0.1000 | 72.78 | – | 72.78 | 0.1000 | 72.87 | – | 72.87 | 0.1000 |
| 2 | 0.50 | 0.50 | 0.00 | 0.00 | 29.48 | 29.48 | – | 0.0475 | 18.33 | 18.33 | – | 0.0225 | 13.75 | 13.75 | – | 0.0225 |
| 2 | 0.38 | 0.38 | 0.12 | 0.12 | 34.16 | 26.72 | 56.50 | 0.0775 | 25.24 | 14.83 | 56.44 | 0.0775 | 22.46 | 9.79 | 60.46 | 0.0850 |
| 2 | 0.25 | 0.25 | 0.25 | 0.25 | 42.85 | 15.80 | 69.89 | 0.0975 | 38.90 | 8.51 | 69.30 | 0.0975 | 37.65 | 6.38 | 68.92 | 0.0975 |
| 2 | 0.12 | 0.12 | 0.38 | 0.38 | 56.37 | 15.80 | 69.89 | 0.0975 | 54.59 | 0.00 | 72.78 | 0.1000 | 54.65 | 0.00 | 72.87 | 0.1000 |
| 2 | 0.00 | 0.00 | 0.50 | 0.50 | 73.09 | – | 73.09 | 0.1000 | 72.78 | – | 72.78 | 0.1000 | 72.87 | – | 72.87 | 0.1000 |
| 3 | 0.50 | 0.50 | 0.00 | 0.00 | 26.89 | 26.89 | – | 0.0600 | 14.86 | 14.86 | – | 0.0350 | 11.14 | 11.14 | – | 0.0350 |
| 3 | 0.38 | 0.38 | 0.12 | 0.12 | 32.91 | 24.79 | 57.27 | 0.0775 | 23.98 | 11.61 | 61.10 | 0.0850 | 21.77 | 8.71 | 60.97 | 0.0850 |
| 3 | 0.25 | 0.25 | 0.25 | 0.25 | 42.78 | 15.42 | 70.14 | 0.0975 | 38.72 | 8.00 | 69.44 | 0.0975 | 37.52 | 6.00 | 69.05 | 0.0975 |
| 3 | 0.12 | 0.12 | 0.38 | 0.38 | 56.46 | 15.42 | 70.14 | 0.0975 | 54.59 | 0.00 | 72.78 | 0.1000 | 54.65 | 0.00 | 72.87 | 0.1000 |
| 3 | 0.00 | 0.00 | 0.50 | 0.50 | 73.09 | – | 73.09 | 0.1000 | 72.78 | – | 72.78 | 0.1000 | 72.87 | – | 72.87 | 0.1000 |

Table 5.4: Maximum rates of correct group identification for various prior distributions and universes of true drug effects with $n = 200$ and $K = 4$.

identification, which range from 5.7 to 22.8%, and indicate the difficulty in selecting subgroups that have $\zeta \in \{0.30, 0, 70\}$. Overall rates of correct group identification range from 5.7 to 81.8% with the upper limit reached with a prior of $\pi_{F2} = 1.00$ and $\alpha_F = 0.10$. Compared to setting $\ell = 0$, SHAPES does better with priors of $\pi_{S2} \leq 0.50$. For priors with $\pi_{S2} < 0.50$ the full population analysis identifies the correct group (the full population in this case) most often.

Rates of $\theta^*$-power are listed in Table 5.5. With $\pi_{S2} = 1.00$, compared to using $\ell = 0$ by using SHAPES we see improvements between 1.7 and 5.9%, depending on $\ell$, the universe and selecting the appropriate $\alpha_F$. Improvements for SHAPES disappear for $\pi_{F2} \geq 0.50$ and for these priors we select $\alpha_F = 0.0975$ or 0.1000 with little rate difference between the two values. This suggests under this broader range of $\zeta$ and this measure that SHAPES is only worth implementing if $\pi_{S2} < 0.50$.

| | | | Universe I | | Universe II | | Universe III | |
|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_{S2}$ | $\pi_{F2}$ | Rate | $\alpha_F$ | Rate | $\alpha_F$ | Rate | $\alpha_F$ |
| 0 | 1.00 | 0.00 | 27.41 | 0.1000 | 27.24 | 0.1000 | 27.22 | 0.1000 |
| 0 | 0.75 | 0.25 | 41.00 | 0.1000 | 40.81 | 0.1000 | 40.83 | 0.1000 |
| 0 | 0.50 | 0.50 | 54.59 | 0.1000 | 54.38 | 0.1000 | 54.45 | 0.1000 |
| 0 | 0.25 | 0.75 | 68.19 | 0.1000 | 67.95 | 0.1000 | 68.06 | 0.1000 |
| 0 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 1 | 1.00 | 0.00 | 33.34 | 0.0225 | 29.99 | 0.0325 | 29.11 | 0.0600 |
| 1 | 0.75 | 0.25 | 44.01 | 0.0575 | 41.99 | 0.0775 | 41.64 | 0.0775 |
| 1 | 0.50 | 0.50 | 55.85 | 0.0775 | 54.73 | 0.0975 | 54.73 | 0.0975 |
| 1 | 0.25 | 0.75 | 68.46 | 0.0975 | 67.98 | 0.0975 | 68.09 | 0.0975 |
| 1 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 2 | 1.00 | 0.00 | 31.08 | 0.0425 | 30.17 | 0.0425 | 29.33 | 0.0600 |
| 2 | 0.75 | 0.25 | 42.86 | 0.0775 | 42.19 | 0.0775 | 41.84 | 0.0775 |
| 2 | 0.50 | 0.50 | 55.26 | 0.0775 | 54.82 | 0.0975 | 54.83 | 0.0975 |
| 2 | 0.25 | 0.75 | 68.41 | 0.0975 | 68.04 | 0.0975 | 68.16 | 0.0975 |
| 2 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |
| 3 | 1.00 | 0.00 | 30.25 | 0.0600 | 29.51 | 0.0600 | 28.95 | 0.0775 |
| 3 | 0.75 | 0.25 | 42.57 | 0.0775 | 41.92 | 0.0775 | 41.67 | 0.0775 |
| 3 | 0.50 | 0.50 | 55.22 | 0.0975 | 54.84 | 0.0975 | 54.85 | 0.0975 |
| 3 | 0.25 | 0.75 | 68.41 | 0.0975 | 68.08 | 0.0975 | 68.20 | 0.0975 |
| 3 | 0.00 | 1.00 | 81.78 | 0.1000 | 81.52 | 0.1000 | 81.68 | 0.1000 |

Table 5.5: Maximum rates of selection of groups with $\theta_{\hat{S}} \geq \theta^*$ for various prior distributions and universes of true drug effects with $n = 200$ and $K = 4$.

In this section, we observe that including a greater range of possible subgroup sizes decreases performance for all scenarios. The previous best performing scenario, universe I and $\ell = 1$, again does the best here but rates for both correct identification and $\theta^*$-power decrease to the extent of

| | | | Universe I | | | | Universe II | | | | Universe III | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_{S2}$ | $\pi_{F2}$ | Overall | Sg | Full | $\alpha_F$ | Overall | Sg | Full | $\alpha_F$ | Overall | Sg | Full | $\alpha_F$ |
| 0 | 1.00 | 0.00 | 0.00 | 0.00 | – | 0.1000 | 0.00 | 0.00 | – | 0.1000 | 0.00 | 0.00 | – | 0.1000 |
| 0 | 0.75 | 0.25 | 20.45 | 0.00 | 81.78 | 0.1000 | 20.38 | 0.00 | 81.52 | 0.1000 | 20.42 | 0.00 | 81.68 | 0.1000 |
| 0 | 0.50 | 0.50 | 40.89 | 0.00 | 81.78 | 0.1000 | 40.76 | 0.00 | 81.52 | 0.1000 | 40.84 | 0.00 | 81.68 | 0.1000 |
| 0 | 0.25 | 0.75 | 61.34 | 0.00 | 81.78 | 0.1000 | 61.14 | 0.00 | 81.52 | 0.1000 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 0 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |
| 1 | 1.00 | 0.00 | 22.81 | 22.81 | – | 0.0000 | 7.60 | 7.60 | – | 0.0000 | 5.70 | 5.70 | – | 0.0000 |
| 1 | 0.75 | 0.25 | 27.84 | 15.04 | 66.26 | 0.0800 | 21.62 | 2.99 | 77.52 | 0.0975 | 20.98 | 2.24 | 77.20 | 0.0975 |
| 1 | 0.50 | 0.50 | 43.48 | 8.97 | 77.99 | 0.0975 | 40.76 | 0.00 | 81.52 | 0.1000 | 40.84 | 0.00 | 81.68 | 0.1000 |
| 1 | 0.25 | 0.75 | 61.34 | 0.00 | 81.78 | 0.1000 | 61.14 | 0.00 | 81.52 | 0.1000 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 1 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |
| 2 | 1.00 | 0.00 | 14.69 | 14.69 | – | 0.0175 | 9.93 | 9.93 | – | 0.0175 | 7.45 | 7.45 | – | 0.0175 |
| 2 | 0.75 | 0.25 | 25.59 | 11.17 | 68.85 | 0.0850 | 22.48 | 4.08 | 77.67 | 0.0975 | 21.63 | 3.06 | 77.34 | 0.0975 |
| 2 | 0.50 | 0.50 | 42.88 | 7.49 | 78.27 | 0.0975 | 40.88 | 4.08 | 77.67 | 0.0975 | 40.84 | 0.00 | 81.68 | 0.1000 |
| 2 | 0.25 | 0.75 | 61.34 | 0.00 | 81.78 | 0.1000 | 61.14 | 0.00 | 81.52 | 0.1000 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 2 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |
| 3 | 1.00 | 0.00 | 12.84 | 12.84 | – | 0.0350 | 7.89 | 7.89 | – | 0.0250 | 5.92 | 5.92 | – | 0.0250 |
| 3 | 0.75 | 0.25 | 25.10 | 7.29 | 78.51 | 0.0975 | 22.37 | 3.88 | 77.83 | 0.0975 | 21.55 | 2.91 | 77.48 | 0.0975 |
| 3 | 0.50 | 0.50 | 42.90 | 7.29 | 78.51 | 0.0975 | 40.86 | 3.88 | 77.83 | 0.0975 | 40.84 | 0.00 | 81.68 | 0.1000 |
| 3 | 0.25 | 0.75 | 61.34 | 0.00 | 81.78 | 0.1000 | 61.14 | 0.00 | 81.52 | 0.1000 | 61.26 | 0.00 | 81.68 | 0.1000 |
| 3 | 0.00 | 1.00 | 81.78 | – | 81.78 | 0.1000 | 81.52 | – | 81.52 | 0.1000 | 81.68 | – | 81.68 | 0.1000 |

Table 5.6: Maximum rates of correct group identification with variable $\zeta$ for prior collection 1 and universes of true drug effects with $n = 200$ and $K = 4$.

making it unlikely an investigator would implement a subgroup search.

### *5.5 PPV and Any Rejection for Prior Collection 1 with* $\zeta = 0.50$

In describing performance we have focused on measures that do not depend on $\pi_N$, but we now consider measures that depend on $\pi_N$: positive predictive value (PPV) and rates of any rejection. Table 5.8 considers values of $\pi_N$ between 0.00 and 1.00 where the remaining probability is split between $\pi_{S2}$ and $\pi_{F2}$, specifically $\pi_{S2} = \pi_{F2} = 0.5(1 - \pi_N)$. For this section, we hold $\zeta = 0.50$.

We expect rates of any rejection to be insensitive to the value of $\alpha_F$ selected as SHAPES is tuned under the null hypothesis and the critical values are selected to keep the overall rejection rate at $\alpha$ under the null. Applying these values to alternatives results in similar rates of rejection regardless of the configuration of $\alpha_F$, typically less than a one percent difference, so in this table we do not report the corresponding $\alpha_F$ for any rejection. The rejection rate reported is that corresponding to the $\alpha_F$ that achieves the highest PPV. Overall, rejection rates are the same across the universes for a fixed $\pi_N$. Rates are sensitive to the value of $\pi_N$. The highest rates of any rejection, about 80% occur with $\pi_N = 0.00$ and lowest when $\pi_N = 1.00$ with rates around 10%. For values of $\pi_N$ between 0 and 1, rates increase steadily.

For this collection of priors, PPV spans from 0 to 100% and remains above 96% when $\pi_N \leq 0.60$. For most priors, the maximum PPV occurs with values of $\alpha_F \geq 0.0775$, although we observe that compared to a full population analysis with $\ell = 0$, maximum PPV is nearly the same. This suggests that similar to any rejection rates, PPV is insensitive to the value of $\alpha_F$.

Rates of regulatory approval for new drugs are quite low, as discussed in Chapter 1. With this consideration, we explore a mixture of heterogeneous and homogeneous effects by allowing the ratio of $\pi_{S2}$ and $\pi_{F2}$ to vary while holding $\pi_N = 0.80$ fixed. Results are shown in Table 5.7. What we observe is that we can slightly improve rates of any rejection and PPV, generally by less than one percent. Rates of any rejection range from 24.0 to 25.8%, with the upper end occurring for $\pi_{S2} = 0.20$ for various values of $\alpha_F$ between 0.0425 and 0.0775. The lower end of this range occurs for the full population analysis. Additionally, we observe a similar tight range of PPV for this scenario as well, with a range from 86.9 to 87.7%. These results suggest that slight improvements in PPV and any rejection are possible using SHAPES under a variety of priors.

### *5.6 Phase 2-3 Performance*

Having applied SHAPES to a collection of drug effects, we now evaluate the application of SHAPES to the drug discovery process. We perform a simulation study with 1,000,000 subjects to enroll in

| | | | | Universe I | | | Universe II | | | Universe III | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_N$ | $\pi_{S2}$ | $\pi_{F2}$ | Any | PPV | $\alpha_F$ | Any | PPV | $\alpha_F$ | Any | PPV | $\alpha_F$ |
| 0 | 0.80 | 0.20 | 0.00 | 24.05 | 86.98 | 0.1000 | 24.19 | 87.04 | 0.1000 | 24.26 | 87.08 | 0.1000 |
| 0 | 0.80 | 0.15 | 0.05 | 24.02 | 86.97 | 0.1000 | 24.18 | 87.04 | 0.1000 | 24.26 | 87.08 | 0.1000 |
| 0 | 0.80 | 0.10 | 0.10 | 24.00 | 86.96 | 0.1000 | 24.17 | 87.03 | 0.1000 | 24.26 | 87.08 | 0.1000 |
| 0 | 0.80 | 0.05 | 0.15 | 23.98 | 86.95 | 0.1000 | 24.16 | 87.03 | 0.1000 | 24.26 | 87.08 | 0.1000 |
| 0 | 0.80 | 0.00 | 0.20 | 23.96 | 86.94 | 0.1000 | 24.15 | 87.03 | 0.1000 | 24.26 | 87.08 | 0.1000 |
| 1 | 0.80 | 0.20 | 0.00 | 25.66 | 87.70 | 0.0600 | 25.20 | 87.50 | 0.0600 | 25.12 | 87.47 | 0.0775 |
| 1 | 0.80 | 0.15 | 0.05 | 25.18 | 87.49 | 0.0600 | 24.90 | 87.37 | 0.0775 | 24.88 | 87.36 | 0.0775 |
| 1 | 0.80 | 0.10 | 0.10 | 24.78 | 87.32 | 0.0850 | 24.66 | 87.26 | 0.0975 | 24.75 | 87.30 | 0.0975 |
| 1 | 0.80 | 0.05 | 0.15 | 24.45 | 87.17 | 0.0850 | 24.53 | 87.20 | 0.0975 | 24.65 | 87.26 | 0.0975 |
| 1 | 0.80 | 0.00 | 0.20 | 24.19 | 87.05 | 0.0950 | 24.39 | 87.14 | 0.0975 | 24.55 | 87.21 | 0.0975 |
| 2 | 0.80 | 0.20 | 0.00 | 25.61 | 87.67 | 0.0425 | 25.27 | 87.53 | 0.0775 | 25.17 | 87.49 | 0.0775 |
| 2 | 0.80 | 0.15 | 0.05 | 25.04 | 87.43 | 0.0425 | 24.97 | 87.40 | 0.0775 | 24.94 | 87.39 | 0.0950 |
| 2 | 0.80 | 0.10 | 0.10 | 24.74 | 87.30 | 0.0950 | 24.73 | 87.29 | 0.0950 | 24.78 | 87.31 | 0.0975 |
| 2 | 0.80 | 0.05 | 0.15 | 24.52 | 87.20 | 0.0950 | 24.54 | 87.21 | 0.0975 | 24.66 | 87.26 | 0.0975 |
| 2 | 0.80 | 0.00 | 0.20 | 24.29 | 87.09 | 0.0950 | 24.39 | 87.14 | 0.0975 | 24.54 | 87.21 | 0.0975 |
| 3 | 0.80 | 0.20 | 0.00 | 25.75 | 87.74 | 0.0600 | 25.37 | 87.57 | 0.0600 | 25.30 | 87.54 | 0.0775 |
| 3 | 0.80 | 0.15 | 0.05 | 25.31 | 87.55 | 0.0600 | 25.06 | 87.44 | 0.0775 | 25.02 | 87.42 | 0.0775 |
| 3 | 0.80 | 0.10 | 0.10 | 24.86 | 87.35 | 0.0600 | 24.74 | 87.30 | 0.0775 | 24.82 | 87.33 | 0.0975 |
| 3 | 0.80 | 0.05 | 0.15 | 24.51 | 87.19 | 0.0775 | 24.58 | 87.22 | 0.0975 | 24.69 | 87.28 | 0.0975 |
| 3 | 0.80 | 0.00 | 0.20 | 24.25 | 87.07 | 0.0975 | 24.43 | 87.15 | 0.0975 | 24.57 | 87.22 | 0.0975 |

Table 5.7: Positive predictive and frequency of any rejection for varying $\ell$ and prior beliefs with $\pi_N = 0.80$, $n = 200$ and $K = 4$.

| | | | | Universe I | | | Universe II | | | Universe III | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ell$ | $\pi_N$ | $\pi_{S2}$ | $\pi_{F2}$ | Any | PPV | $\alpha_F$ | Any | PPV | $\alpha_F$ | Any | PPV | $\alpha_F$ |
| 0 | 0.00 | 0.50 | 0.50 | 82.00 | 100.00 | 0.1000 | 81.62 | 100.00 | 0.1000 | 81.67 | 100.00 | 0.1000 |
| 0 | 0.20 | 0.40 | 0.40 | 67.50 | 99.67 | 0.1000 | 67.26 | 99.67 | 0.1000 | 67.32 | 99.67 | 0.1000 |
| 0 | 0.40 | 0.30 | 0.30 | 53.00 | 98.88 | 0.1000 | 52.89 | 98.88 | 0.1000 | 52.96 | 98.88 | 0.1000 |
| 0 | 0.60 | 0.20 | 0.20 | 38.50 | 96.61 | 0.1000 | 38.53 | 96.61 | 0.1000 | 38.61 | 96.62 | 0.1000 |
| 0 | 0.80 | 0.10 | 0.10 | 24.00 | 86.96 | 0.1000 | 24.17 | 87.03 | 0.1000 | 24.26 | 87.08 | 0.1000 |
| 0 | 0.90 | 0.05 | 0.05 | 16.75 | 67.41 | 0.1000 | 16.98 | 67.71 | 0.1000 | 17.08 | 67.83 | 0.1000 |
| 0 | 0.95 | 0.02 | 0.02 | 13.13 | 43.42 | 0.1000 | 13.39 | 43.92 | 0.1000 | 13.49 | 44.11 | 0.1000 |
| 0 | 1.00 | 0.00 | 0.00 | 9.50 | 0.00 | 0.1000 | 9.80 | 0.00 | 0.1000 | 9.90 | 0.00 | 0.1000 |
| 1 | 0.00 | 0.50 | 0.50 | 83.70 | 100.00 | 0.0000 | 82.54 | 100.00 | 0.0000 | 82.48 | 100.00 | 0.0000 |
| 1 | 0.20 | 0.40 | 0.40 | 68.92 | 99.67 | 0.0775 | 68.07 | 99.67 | 0.0975 | 68.05 | 99.67 | 0.0975 |
| 1 | 0.40 | 0.30 | 0.30 | 54.14 | 98.90 | 0.0775 | 53.60 | 98.89 | 0.0975 | 53.61 | 98.89 | 0.0975 |
| 1 | 0.60 | 0.20 | 0.20 | 39.37 | 96.69 | 0.0850 | 39.13 | 96.66 | 0.0975 | 39.18 | 96.67 | 0.0975 |
| 1 | 0.80 | 0.10 | 0.10 | 24.78 | 87.32 | 0.0850 | 24.66 | 87.26 | 0.0975 | 24.75 | 87.30 | 0.0975 |
| 1 | 0.90 | 0.05 | 0.05 | 17.49 | 68.34 | 0.0850 | 17.42 | 68.26 | 0.0975 | 17.54 | 68.40 | 0.0975 |
| 1 | 0.95 | 0.02 | 0.02 | 13.84 | 44.73 | 0.0850 | 13.81 | 44.67 | 0.0850 | 13.93 | 44.89 | 0.0975 |
| 1 | 1.00 | 0.00 | 0.00 | 10.19 | 0.00 | 0.0000 | 10.23 | 0.00 | 0.0000 | 10.37 | 0.00 | 0.0000 |
| 2 | 0.00 | 0.50 | 0.50 | 83.42 | 100.00 | 0.0000 | 82.77 | 100.00 | 0.0000 | 82.69 | 100.00 | 0.0000 |
| 2 | 0.20 | 0.40 | 0.40 | 68.72 | 99.67 | 0.0950 | 68.25 | 99.67 | 0.0975 | 68.22 | 99.67 | 0.0975 |
| 2 | 0.40 | 0.30 | 0.30 | 54.06 | 98.90 | 0.0950 | 53.73 | 98.90 | 0.0975 | 53.74 | 98.90 | 0.0975 |
| 2 | 0.60 | 0.20 | 0.20 | 39.40 | 96.69 | 0.0950 | 39.21 | 96.67 | 0.0975 | 39.26 | 96.68 | 0.0975 |
| 2 | 0.80 | 0.10 | 0.10 | 24.74 | 87.30 | 0.0950 | 24.73 | 87.29 | 0.0950 | 24.78 | 87.31 | 0.0975 |
| 2 | 0.90 | 0.05 | 0.05 | 17.43 | 68.27 | 0.0850 | 17.49 | 68.35 | 0.0950 | 17.57 | 68.44 | 0.0950 |
| 2 | 0.95 | 0.02 | 0.02 | 13.81 | 44.69 | 0.0425 | 13.88 | 44.80 | 0.0950 | 13.96 | 44.95 | 0.0950 |
| 2 | 1.00 | 0.00 | 0.00 | 10.28 | 0.00 | 0.0000 | 10.26 | 0.00 | 0.0000 | 10.36 | 0.00 | 0.0000 |
| 3 | 0.00 | 0.50 | 0.50 | 83.57 | 100.00 | 0.0000 | 82.84 | 100.00 | 0.0000 | 82.78 | 100.00 | 0.0000 |
| 3 | 0.20 | 0.40 | 0.40 | 68.89 | 99.67 | 0.0775 | 68.31 | 99.67 | 0.0975 | 68.29 | 99.67 | 0.0975 |
| 3 | 0.40 | 0.30 | 0.30 | 54.21 | 98.91 | 0.0775 | 53.78 | 98.90 | 0.0975 | 53.80 | 98.90 | 0.0975 |
| 3 | 0.60 | 0.20 | 0.20 | 39.52 | 96.70 | 0.0775 | 39.26 | 96.68 | 0.0975 | 39.31 | 96.68 | 0.0975 |
| 3 | 0.80 | 0.10 | 0.10 | 24.86 | 87.35 | 0.0600 | 24.74 | 87.30 | 0.0775 | 24.82 | 87.33 | 0.0975 |
| 3 | 0.90 | 0.05 | 0.05 | 17.68 | 68.59 | 0.0600 | 17.50 | 68.36 | 0.0825 | 17.57 | 68.45 | 0.0975 |
| 3 | 0.95 | 0.02 | 0.02 | 14.12 | 45.23 | 0.0475 | 13.93 | 44.90 | 0.0650 | 13.96 | 44.95 | 0.0950 |
| 3 | 1.00 | 0.00 | 0.00 | 10.61 | 0.00 | 0.0000 | 10.41 | 0.00 | 0.0000 | 10.36 | 0.00 | 0.0000 |

Table 5.8: Positive predictive value and frequency of any rejection for varying $\ell$ and prior beliefs with $n = 200$ and $K = 4$.

multiple clinical trials. The trials occur sequentially, in that we: 1) randomly sample a true drug effect with probabilities based on a prior distribution, 2) simulate a phase 2 trial that is analyzed with SHAPES, and 3) when indicated, simulate a phase 3 trial in the SHAPES-selected population, where the power corresponds to the true effect in this selected population. We repeat this sequence until all subjects have been used. For the prior distribution we use Universe I with $\pi_N = 0.80$ and for the remaining (1-$\pi_N$) probability we consider a range of splits between $\pi_{S2}$ and $\pi_{F2}$. For the operating characteristics, we implement a similar set-up as before by setting the effect the trials are design to detect to $\theta^* = 0.146$, with the phase 2 overall type I error to $\alpha_2 = 0.10$ (for a one-sided test) and a type II error of $\beta_2 = 0.20$ (or 80% power), which requires $n_2 = 200$. For the phase 3 trial we use the same effect, with $\alpha_3 = 0.025$ (one-sided) and $\beta_3 = 0.10$, which requires $n_3 = 450$. We set $\ell = 1$.

We repeat the simulations for each prior and value of $\alpha_F$ in the range from 0.0000 to 0.1000 by 0.0025. For each replication we determine the probability that a drug with a positive trial at the end of phase 2 or phase 3 has a positive treatment effect (positive predictive value) and the total number of drugs with a positive phase 3 trial. We are interested in determining whether SHAPES performs better than implementing a full population analysis. We anticipate that for each replication there will be a greater degree of variability in the results because of the random selection of the true drug effect and therefore to identify overall trends we will also estimate the linear relationship between $\alpha_F$ and the statistics of interest.

The binary covariates for each subject in a phase 2 trial are simulated to be random with $Pr(X_i = 1) = 0.5$ for $i \in \{1, 2..., K\}$ where $K = 4$. When a phase 2 trial is significant, the true effect in the SHAPES-selected population is calculated and the corresponding power for a phase 3 trial is determined. We then randomly sample a binary variable with the probability equal to this power to determine whether the phase 3 trial is significant or not. Implicit here is the assumption that we always have enough of the type of patients targeted for this trial, but in the real world there may be considerable additional effort to recruit for a phase 3 trial when a smaller population is targeted for enrollment.

Results are shown in Figure 5.4. The top row of plots contains the observed PPV at the end of phase 2, PPV at the end of phase 3, and total number of drugs approved at the end of phase 3 with each line corresponds to a certain prior ranging from $(\pi_{S2}, \pi_{F2}) = (0.20, 0.00)$ to $(\pi_{S2}, \pi_{F2}) = (0.00, 0.20)$. As expected, the variability due to the random selection of the true effect makes discerning trends difficult. To improve this the bottom row of plots contains the linear trends. Differences by prior are overall small. For all three measures, we observe an ordered trend where the

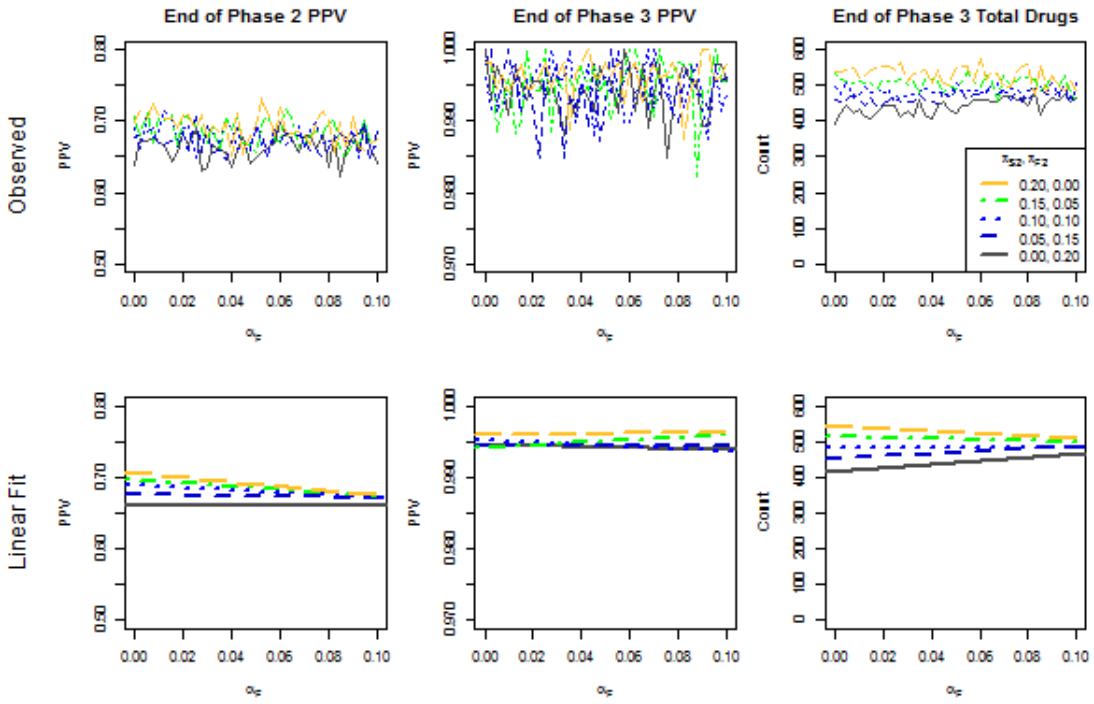



Figure 5.4: Observed and the linear fit of PPV at the end of the phases and the total number of drugs approved when using SHAPES.

prior corresponding to only heterogeneous effects has the highest rates. At the end of phase 2 for this prior, there is negative relationship between $\alpha_F$ and PPV, but it is not readily discernible at the end of phase 3. For all priors the PPV at the end of phase 3 is flat across $\alpha_F$ and approximately 99.5%. Around 500 total drugs are approved for each replication. We observe a slight increasing trend with the prior $(\pi_{S2}, \pi_{F2}) = (0.00, 0.20)$ and a slight negative trend with the prior $(\pi_{S2}, \pi_{F2}) = (0.20, 0.00)$. Selecting a higher $\alpha_F$ for the former (lower for the latter) results in about 50 additional drugs approved compared to a lower $\alpha_F$ (higher).

Overall these differences are quite subtle. These results suggest that in terms of PPV at the end of phase 3 and the total number of drugs approved getting $\alpha_F$ just right will not change performance much. Although, this may not hold for other measures or where the null is true less often.

## *5.7 Summary*

While ensuring external validity is difficult as the true state of nature is likely quite diverse, in this chapter we present simple scenarios that incorporate some exploration to determine whether there

is a sensible general approach. Generally we see that with universe I, $\ell = 1$ and $K = 4$, we can do well with correct group identification and other measures when selecting an optimal $\alpha_F$ with significant consideration given to the prior distribution of effects. For other combinations correct subgroup identification is lower. SHAPES struggles when we add to the collection of possible true subgroups or increase the number of candidate subgroups. For most results we focus on a subgroup prevalence of $\zeta = 0.50$. When we consider a greater collection of values performance deteriorates, suggesting $\zeta = 0.50$ may be the only reasonable subgroup prevalence to target for our setting of a fixed moderate to large subgroup effect. To fully evaluate SHAPES in the drug discovery process we implement a sequential two stage simulation study where the first stage is a phase 2 trial and the second stage is a phase 3 trial. We find that there is evidence for possible performance improvements under a high rate of heterogeneous effects, when correctly assuming a simple subgroup definition.

# Chapter 6

# LOGIC REGRESSION AND SHAPES APPLIED TO A CLINICAL TRIAL IN ACUTE MYELOID LEUKEMIA

In this chapter we apply LR and SHAPES to a clinical trial investigating treatments for acute myeloid leukemia (AML). Our primary aim is to describe some of the statistical and scientific issues when conducting an exploratory subgroup analysis with clinical trial data. Statistically, we examine how to generate the critical values for $\alpha_F$ and $n_{leaves}$ or $\ell$. Scientifically, we consider the evidence for baseline patient characteristics to explain variation in complete remission rate in AML under two treatments.

## *6.1 Scientific and Study Background*

AML is a cancer of the blood and bone marrow. The disease occurs when abnormal white blood cells proliferate in the marrow hindering the production of normal blood cells and creating abnormal versions of these cells to circulate. The general treatment strategy is to first destroy all observable traces of cancer via chemotherapy, referred to as the induction phase. Once patients complete this stage they move to the conduction phase where additional chemotherapy is given with the goal of destroying any lingering unobservable cancer.

In 1984 Memorial Sloan-Kettering Cancer Center conducted a randomized open label clinical trial comparing daunorubicin versus idarubicin to treat AML (for the original results see [3]). The drugs are both anthracyclines, a particular class of chemicals. Doses of 12 $mg/M^2$ (idarubicin) and 50 $mg/M^2$ (daunorubicin) were administered intravenously over the first three days of a five day round of chemotherapy. Both arms were given the drug cytosine arabinoside. Patients that did not have complete remission after the first round had a second round. The total sample size for the trial was $n = 130$ with equal treatment assignment by arm. The primary outcome of the trial was complete remission status after the induction phase defined as having either bone marrow of normal cellularity or not hypocellular with concurrent evidence of normal cells. During the course of the study 10 patients were identified as having issues, such as the original diagnosis of AML being incorrect or complete remission status indeterminate, resulting in a final sample size of $n = 120$.

### *6.2 Research Questions of Interest*

In developing the questions for this analysis, we seek to go beyond questions specific to disease area and to also investigate issues related to the implementation of LR and SHAPES. We ask:

1. When making a distributional assumption to determine critical values, how are the values different across various data generation approaches? What is the evidence that correlation should be considered?
2. What is the evidence for the presence of a predictive subgroup with respect to complete remission for AML patients treated with idarubicin?
3. How do decisions change as we vary the tuning parameters?

For the first question, we try three approaches to account for the correlation of the $K$ covariates when determining the critical values. If the covariates are independent, as in the previous simulation studies, this is not an issue. However, when dealing with multiple covariates from a similar source, for example blood tests, we expect strong correlation. We are interested in not only how the critical values change but also how they translate into a selected group.

For our second question, defining a predictive subgroup must be with respect to an effect on an outcome of interest (or in this case an estimated effect). For this analysis, we select complete remission as the outcome because it meets three criteria: 1) a binary response, 2) there is a possibility of large effects, and 3) it has clinical value. The first criterion is readily met. For the second criterion, Berman et al. [4] summarized three trials, including this one, comparing the same two treatments which estimated differences of at least 11% for complete response in favor of idarubicin. Finally as complete remission is required before a patient is cured, it has utility. Berman also reported a positive association between duration of remission and survival [4].

Finally, we are interested in how decisions change with the tuning parameters. Both LR and SHAPES select different groups depending on $\alpha_F$ and $n_{leaves}$ or $\ell$. Reporting how the selected group changes with the tuning parameter provides evidence for or against the robustness of an identified group and the ability of these methods to identify any group. The primary purpose is to understand how LR and SHAPES work. In actual practice the tuning parameters should be selected a priori to ensure proper calibration of the operating characteristics.

### *6.3 Data and the Covariates*

For all patients, available data includes baseline demographic, functional status, laboratory data and treatment arm assignment as well as follow-up data for complete remission, bone marrow transplan-

tation, and death (with some patients censored). For the covariates we include in our search, we focus on the blood tests values of white blood cell count, platelet count and hemoglobin along with gender. We do not include age in our analysis as it has been described as a strong prognostic factor with relation to earlier death for older patients [7] and would likely cause an overinflation of type I error with LR. We considered including the French-American-British classification system, which has seven subgroups of AML based on histology. While both LR and SHAPES could be modified to allow multiple groups, we felt there is not any obvious way to combine some of these groups and without combination the limited number of subjects limits our information.

In determining appropriate cut points to dichotomize the blood test values, we reviewed standard blood test cut points and also the distribution of covariate values. For white blood cell counts (WBC) we selected a cut-point of 4.0 $10^3/mm^3$, for platelet count a cutpoint of 50.0 $10^3/mm^3$ and for hemoglobin 9.0 $mg/dl$. One 54-year-old male assigned to the daunorubicin arm is missing laboratory values and therefore is excluded from our subgroup search. Table 6.1 reports the selected baseline variables for the evaluable patients by treatment group and overall. Males are 48.7% of the patients. While the distribution of hemoglobin appears similar comparing the treatment groups, both WBC and platelet counts are on average higher in the daunorubicin arm. This is driven in part by the higher maximum values in this arm, as a comparison of the quartiles shows greater similarity. For the dichotomized version of WBC, 30.5% of the daunorubicin arm are less than 4.0 compared to 30.0% of the idarubicin arm. For platelet count the percents by arm are 42.3 and 50.0% and for hemoglobin they are 32.2 and 43.3%. This suggests the covariates distribution are comparable for each arm but not the same.

Before moving onto LR and SHAPES we briefly consider all subgroups that are possible and their estimated effects. We highlight this to emphasize how many subgroups are possible in this situation and the relationship among them. The four dichotomized covariates form a collection of 65,536 possible candidate groups, including the full population and none. Figure 6.1 shows histograms for the observed prevalences and effects (in terms of risk difference) of each possible subgroup and a scatterplot of the same data. We observe that the majority of subgroups occur in between a prevalence of 0.40 and 0.60, potentially a positive as both LR and SHAPES do best in this range. Most of the subgroups have a difference of complete remission rates between 0.00 and 0.50 in favor of idarubicin. The scatterplot demonstrates the strong correlation in terms of prevalence and effect for the subgroups; with most values of $\zeta_D$ the actual effect is constrained. Only near $\zeta_D = 0.00$ do we observe a greater range of observed effects.

| Variable (Binary Label) | Statistic | Daunorubicin n=59 | Idarubicin n=60 | All n=119 |
|---|---|---|---|---|
| Male ($X_1$) | n (%) | 31 (52.5) | 27 (45.0) | 58 (48.7) |
| White Blood Cell Count ($X_2$) | mean (SD) | 45.35 (56.48) | 26.25 (33.03) | 35.86 (46.94) |
| $10^3/mm^3$ | $< 4.0$, n (%) | 18 (30.5) | 18 (30.0) | 36 (30.2) |
| | min, max | 0.7, 215.0 | 0.4, 139.0 | 0.4, 215.0 |
| | 50th (25th, 75th) %-ile | 17.7 (3.3, 75.4) | 11.0 (2.8, 40.2) | 13.8 (2.9, 52.4) |
| Platelet Count ($X_3$) | mean (SD) | 93.29 (88.71) | 66.32 (59.50) | 79.69 (79.30) |
| $10^3/mm^3$ | $< 50.0$, n (%) | 25 (42.3) | 30 (50.0) | 55 (46.2) |
| | min, max | 11.0, 457.0 | 11.0, 370.0 | 11.0, 457.0 |
| | 50th (25th, 75th) %-ile | 63.0 (36.0, 128.5) | 49.5 (31.5, 74.8) | 57.0 (33.5, 97.5) |
| Hemoglobin ($X_4$) | mean (SD) | 9.51 (1.41) | 9.16 (1.80) | 9.33 (1.62) |
| $g/dl$ | $< 9.0$, n (%) | 19 (32.2) | 26 (43.3) | 45 (37.8) |
| | min, max | 6.4, 13.9 | 2.8, 13.7 | 2.8, 13.9 |
| | 50th (25th, 75th) %-ile | 9.3 (8.6, 10.2) | 9.2 (8.0, 10.2) | 9.3 (8.3, 10.2) |

Table 6.1: Statistics for the baseline variables of interest by treatment group for the evaluable patients with one patient removed.

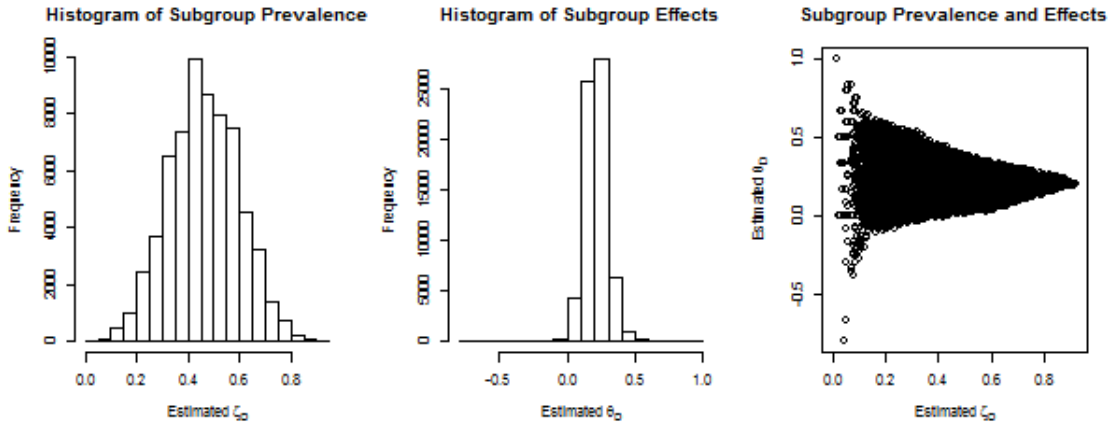



Figure 6.1: Prevalence and observed effect for the 65,536 possible subgroups generated from the four binary covariates.

### *6.4 Calculation of Critical Values*

While in our simulation studies we assume independent covariates, in this example we anticipate correlation among the blood test measurements. We implement the *distributional assumptions* approach to determine the critical values, which involves: defining a covariate distribution that models

the correlation, assuming under the null that the daunorubicin response rate reflects the complete remission rate for both arms, and replicating the trial 10,000 under the assumption that the covariates and outcome are independent. For the first step of this process we consider three covariate generation mechanisms.

We call our first covariate generation mechanism the *binomial* approach and we use the observed probabilities listed in Table 6.1 as the true probabilities. We also assume independence among the four baseline binary covariates and we assume for the $i^{th}$ patient the $K^{th}$ covariate for $K \in \{1, 2, 3, 4\}$ is distributed:

$$x_{Ki} \sim Bin(1, p_K) \text{ where } \vec{p} = (p_1, p_2, p_3, p_4) = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4) = (0.500, 0.302, 0.462, 0.378).$$

We are then able to sample from this distribution to observe covariate values in a simulated null trial.

We refer to the second mechanism as the multivariate normal or just *normal* approach. We model the three continuous covariates with a multivariate normal distribution and model gender as an independent binomial process as in the *binomial* approach. For $i \in \{1, ..., 129\}$, we assume the blood test levels are distributed as:

$$(x_{2i}, x_{3i}, x_{4i}) \sim N_3(\mu, \Sigma).$$

For the parameter values we could use the sample mean vector and covariance matrix. However, the statistics in Table 6.1 imply that white blood cell and platelet counts are skewed distributions and may not be well-approximated by the normal. Histograms confirm this, shown in the first column of Figure 6.2 with the cut points indicated by a vertical blue dashed line. Agreement with the empirical distribution as estimated by the histogram is important because it is our best guess of the truth and we take random samples from our selected distribution. If we consider the log-transformed histograms, shown in the second column, we observe a closer resemblance to a normal distribution, particularly for the white blood cell and platelet counts. The log-transformation also maps the positive blood test values to the entire real line, which agrees with the support of the normal distribution.

Under a log-transformation of the three variables, the empirical mean vector and covariance matrix the parameter values are:

$$\mu = \hat{\mu} = \begin{bmatrix} 2.503 & 4.057 & 2.217 \end{bmatrix}$$

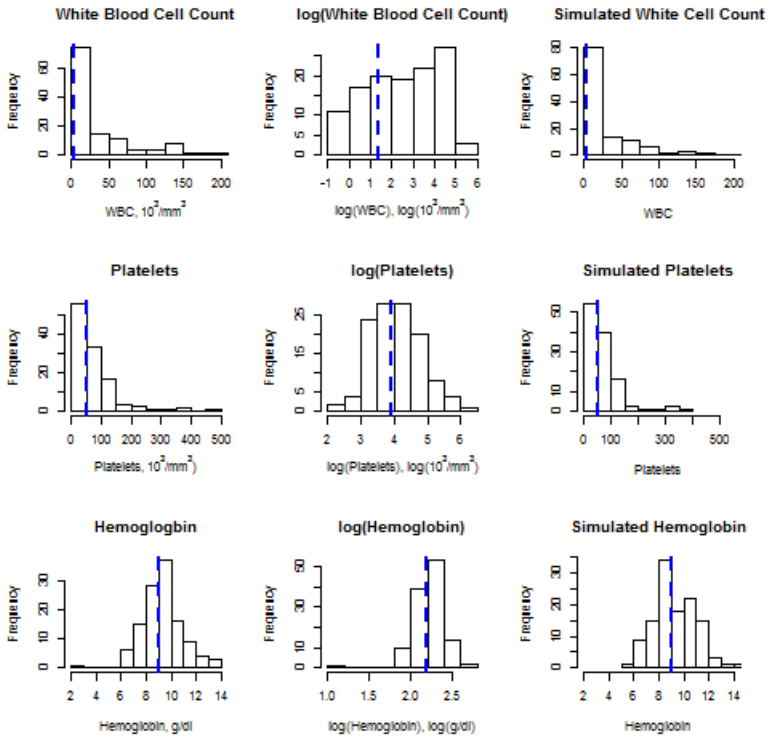



Figure 6.2: Histograms of the true distribution (column 1), the log-transformed true distribution (column 2) and one single random replication from the multivariate normal simulation (column 3) for the baseline covariates for the rows of white blood cell count, platelet count, and hemoglobin. The vertical blue, dotted lines indicate the cut-offs used in our analysis.

$$\Sigma = \hat{\Sigma} = \begin{bmatrix} 2.782 & -0.270 & -0.015 \\ -0.270 & 0.609 & 0.030 \\ -0.015 & 0.030 & 0.037 \end{bmatrix}.$$

The correlations are relatively weak. White blood cell count shows a negative relationship with both platelets and hemoglobin with these latter two having a weak positive correlation. To more closely model the expected true distribution we add more control over the range of possible values by implementing a truncated multivariate normal distribution and specify the maximum and minimum values of the simulated measurements. On the original scale of the data, for the maximum we use the observed maximum plus one standard deviation and for the minimum we select the minimum of either 0.01 or the observed minimum minus one standard deviation.

To assess whether this model is reasonable, the third column of Figure 6.2 shows a single random

sample of size 119 from this truncated multivariate normal, where the values are exponentiated and displayed by histograms. The similarity between columns 1 and 3 is apparent. We also considered using the untransformed hemoglobin, but simulated data resembles the observed distribution better with the transformation than without.

| W | n | % | Exp. % | Male | WBC $< 4.0$ | Plat. $< 50.0$ | Hgb. $< 9.0$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.8 | 2.6 | 1 | 1 | 1 | 1 |
| 2 | 4 | 3.1 | 4.3 | 1 | 1 | 1 | 0 |
| 3 | 3 | 2.3 | 3.1 | 1 | 1 | 0 | 1 |
| 4 | 6 | 4.7 | 5.1 | 1 | 1 | 0 | 0 |
| 5 | 7 | 5.4 | 6.1 | 1 | 0 | 1 | 1 |
| 6 | 11 | 8.5 | 10.0 | 1 | 0 | 1 | 0 |
| 7 | 9 | 7.0 | 7.1 | 1 | 0 | 0 | 1 |
| 8 | 17 | 13.2 | 11.7 | 1 | 0 | 0 | 0 |
| 9 | 6 | 4.7 | 2.6 | 0 | 1 | 1 | 1 |
| 10 | 3 | 2.3 | 4.3 | 0 | 1 | 1 | 0 |
| 11 | 1 | 0.8 | 3.1 | 0 | 1 | 0 | 1 |
| 12 | 12 | 9.3 | 5.1 | 0 | 1 | 0 | 0 |
| 13 | 12 | 9.3 | 6.1 | 0 | 0 | 1 | 1 |
| 14 | 11 | 8.5 | 10.0 | 0 | 0 | 1 | 0 |
| 15 | 6 | 4.7 | 7.1 | 0 | 0 | 0 | 1 |
| 16 | 10 | 7.8 | 11.7 | 0 | 0 | 0 | 0 |

Table 6.2: Frequency of the 16 baseline covariate combinations. A '1' indicates true and a '0' false. The 'Exp. %' columns reflects the expected percent if the covariates are independent.

The third approach we refer to as *bootstrap* involves a bootstrap procedure where we sample with replacement from the observed covariate vectors. Table 6.2 lists the 16 possible combinations of the four binary covariates, with '1' indicating true and '0' indicating false. There are observations of all possible combinations or points in Boolean space with frequencies ranging from 0.8 to 13.2%. For reference we also include the expected probabilities if the covariates are independent, which are the underlying probabilities of each point in the *binomial* approach. Similar to the *binomial* approach, we can index the points as W with $W \in \{1, ..., 16\}$ and then randomly sample a single W with each point occurring with probability equal to the observed frequency. This approach has the advantage of incorporating the observed correlation structure among the covariates.

Next, we combine the covariate generation mechanism along with an assumption about the remission rate to simulate trials under a null hypothesis of no difference by arm. The estimated complete remission rate in the daunorubicin arm is 59.3%. A single trial simulation uses LR and SHAPES to identify the best candidate subgroup based on each methods criterion and the significance of the candidate is tested through stratification. We perform 10,000 replications, record the minimum subgroup p-value and determine a critical value that preserves the overall experiment-wise type I error at level $\alpha$.

Figure 6.3 shows for a range of $\alpha_F$ the one-sided critical values on the p-value scale for a selected subgroup using LR for various $n_{leaves}$ and the three covariate generation mechanisms. Increasing the $n_{leaves}$ tuning parameter requires smaller critical values as the additional exploration of a larger $n_{leaves}$ creates more candidate subgroups and hence a smaller minimum p-value. Critical values are similar for the three generation mechanisms with no mechanism consistently providing either the highest or lowest values for all $n_{leaves}$. This is not unexpected as with the low empirical correlation the mechanisms should provide similar critical values. As $\alpha_F$ increases the critical values decrease with indications of discreteness, which suggests the p-values are not truly continuous in this case with a constrained sample size.

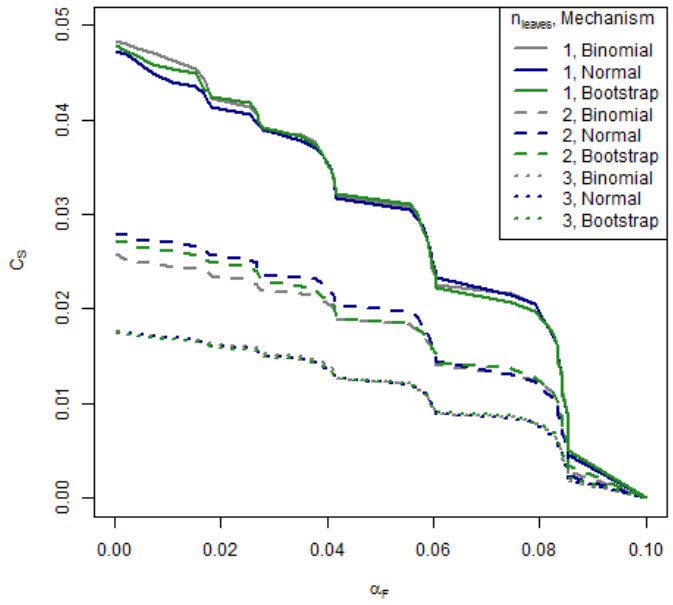



Figure 6.3: Subgroup critical values for LR on the p-value scale for various values of $\alpha_F$.

When calculating the critical values for SHAPES, we have the possibility of allotting a specific amount of $\alpha$ to each depth. We use the same approach in our simulations where we define vector of cumulative type I error where $\alpha$ is evenly split amongst the depths. For example with $\ell = 3$ and the vector of depths $(0, 1, 2, 3)$ we have:

$$\vec{\alpha} = (\alpha_F, \alpha_F + \frac{\alpha - \alpha_F}{3}, \alpha_F + \frac{2(\alpha - \alpha_F)}{3}, \alpha_F + \frac{3(\alpha - \alpha_F)}{3}).$$

Essentially, this vector indicates that we believe any of the possible subgroup depths are equally likely. The critical values are shown in Figure 6.4 with each plot representing a value of $\ell$. We again observe that the different covariate generation mechanisms (represented by different colors) provide similar critical values. The line style corresponds to the depth with a solid line indicating a depth of one, the dashed line a depth of two (for $\ell \in \{2, 3\}$) and the dotted line a depth of three (for $\ell = 3$). For the range of $\alpha$ the critical values maintain an ordered relationship with a depth of two having the lowest critical values, then a depth of one and finally a depth of two. Additionally, we again observe a step pattern in the critical values for changing $\alpha_F$ and the critical values are smaller than with LR because SHAPES directly uses the minimum p-value as the selection criterion.

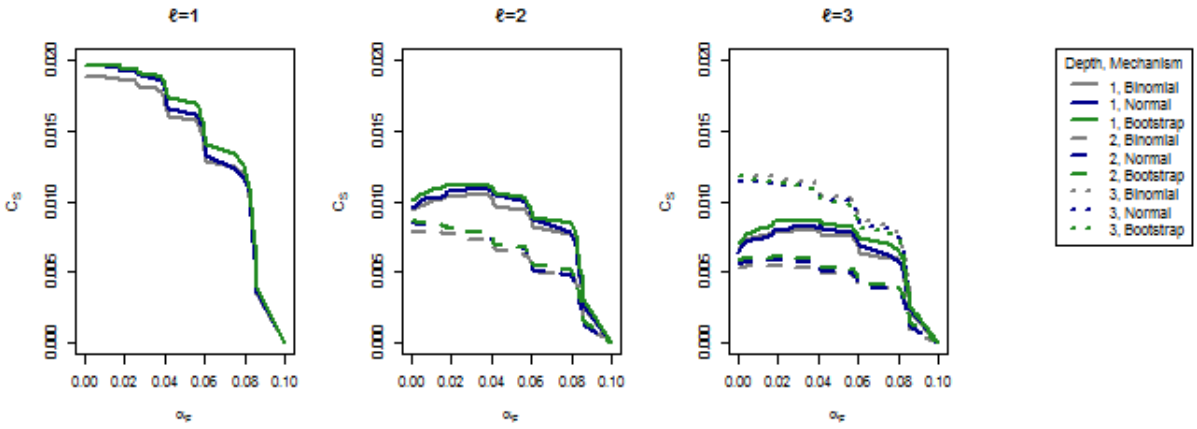



Figure 6.4: Subgroup critical values for SH on the p-value scale for various values of $\alpha_F$ and $\ell$.

### *6.5 Results from Logic Regression*

The results from LR are shown in Table 6.3. The unadjusted p-value suggest there is statistical significance for the full population and possibly the model with $n_{leaves} = 2$ where the predictive subgroup is defined by $X_1^C \vee X_2^C$. For the other values of $n_{leaves}$ the tests yield p-values greater than 0.10, so it is impossible for these subgroups to be selected. However for these models if $\alpha_F > 0.078$, the full population would be selected and all subgroups contain the $X_1^C$ term, suggesting this might be an important factor. Focusing then on $n_{leaves} = 2$, we apply the critical values obtained from the three covariate generation mechanisms along with the decision rule of *minimum p-value* to determine what our final group is. For values of $\alpha_F \leq 0.0375$ we select the subgroup and for greater we select

the full population with the *binomial* and *bootstrap* mechanisms. The transition point is similar but slightly greater with the *normal* at $\alpha_F \leq 0.0425$. The identified subgroup is females with a white blood cell count greater than or equal to 4.0.

| $n_{leaves}$ | Selected Subgroup | Subgroup $n$ | Estimated Difference | Unadjusted p-value |
|---|---|---|---|---|
| 0 | All | 119 | 0.21 | 0.0078 |
| 1 | $X_1^C$ | 61 | 0.13 | 0.1041 |
| 2 | $X_1^C \vee X_2^C$ | 105 | 0.23 | 0.0042 |
| 3 | $X_1^C \wedge (X_2 \vee X_3^C)$ | 38 | 0.13 | 0.1138 |

Table 6.3: Subgroups proposed by LR by $n_{leaves}$.

### *6.6 Results from SHAPES*

Compared to LR, SHAPES shows greater sensitivity to the tuning parameter and $\alpha_F$ as decisions change more often, as shown in Figure 6.5. This plot represents the decision made by SHAPES as $\alpha_F$ increases. Overall, we always reject some null hypothesis in favor of either the full population or a subgroup. For $\ell = 1$, we select either $X_4$ as the subgroup of interest or the full population. For $\ell = 2$ we also sometimes select $X_3 \vee X_4$ and for $\ell = 3$ we select either the full population or $X_2^C \vee (X_3 \vee X_4)$. The variable $X_4$ or hemoglobin $< 9.0$ consistently appears in all the accepted subgroups. Overall, the various covariate generation mechanism result in similar decisions with an exception that with $\ell = 3$ the *bootstrap* critical values select $X_4$.

Table 6.4 shows the $n$, estimated difference, and unadjusted p-values for the selected subgroups for SHAPES. Generally the unadjusted p-values and the estimated differences are larger than compared to the subgroups proposed by LR. The estimated differences in response rate are large and greater than in the full population effect. For example with $X_4$ there is a 44% higher complete response rate with daunorbucin, which is double the rate in the full evaluable population.

| Selected Subgroup | Subgroup $n$ | Estimated Difference | Unadjusted p-value |
|---|---|---|---|
| All | 119 | 0.21 | 0.0078 |
| $X_4$ | 45 | 0.44 | 0.0021 |
| $X_3 \vee X_4$ | 74 | 0.33 | 0.0017 |
| $X_2^C \vee (X_3 \vee X_4)$ | 101 | 0.27 | 0.0024 |

Table 6.4: Subgroups proposed by SHAPES.

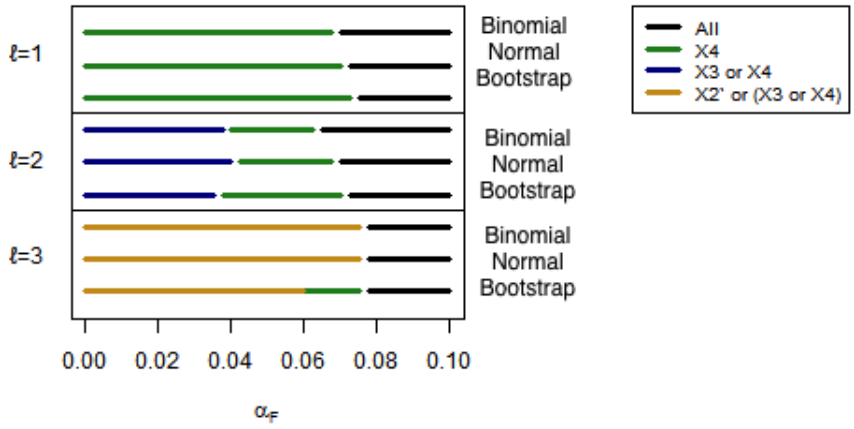



Figure 6.5: Subgroup critical values for SH on the p-value scale for various values of $\alpha_F$ and $\ell$.

### *6.7 Summary and Conclusions*

In applying LR and SHAPES to data from a clinical trial in AML we explore some of the practical aspects of the implementation of our methods. We find that various methods of generating baseline covariate distributions had little impact on the critical values or the ultimate decision of the analysis. Both methods are sensitive to the value of $\alpha_F$ and the tuning parameters. Interestingly, both methods select different candidate subgroups. With LR we mostly select the full population except when $n_{leaves} = 2$ and $\alpha_F$ is low and the subgroups selected include a term for females. With SHAPES we always reject the null hypothesis for either the full population or a subgroup. SHAPES proposes subgroup definitions that all contain the high hemoglobin group. Interpretation of these results is difficult given the significant amount of exploration.

# Chapter 7

# KEY FINDINGS, LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The problem of subgroup identification in clinical trials is of interest in a wide variety of disease areas, particularly oncology. Traditionally, an investigator: 1) limits the enrollment criteria to study patients that he or she presumes will benefit, while: 2) pre-specifying a limited number of subgroups for secondary analyses. However, the decisions about who to enroll and which subgroups to pre-specify are often based on limited observational data; interest is increasing in developing exploratory statistical methods for subgroup identification. Such methods must account for the amount of exploration conducted to maintain appropriate test calibration. Currently, there are few methods available and there is no generally accepted exploratory approach. In this dissertation, we develop two approaches that provide for an opportunity to identify a predictive subgroup while maintaining test calibration. In this chapter we describe some of the critical aspects of searching for and identifying predictive subgroups. For each aspect we summarize the key results of the preceding chapters in the context of our simulation settings. We focus on the most explored setting where $n = 200$ and the MCID is $\theta^*$. For each aspect we consider the limitations and suggest areas for future research.

## *7.1 Optimality Criterion*

Evaluating the performance of methods that search for subgroups requires the selection of a metric to allow for an evaluation of performance and a comparison with a single full population analysis. We propose several metrics in this dissertation that go beyond previous research, including rates of correct group identification, $\theta^*$-power, and rates of any rejection. Other optimality criteria are possible and the appropriate one to implement requires consideration of investigators and patients' priorities, including side effect profiles and the availability of alternative treatments. These priorities are likely to be different by disease area. While it is ideal to perfectly identify the correct group so that patients are only given a treatment that is safe, effective and efficacious for them, with limited data and lack of knowledge of the truth compromises are necessary. The metric of $\theta^*$-power indirectly measures the mix of patients in the selected subgroup, but large effects in a subgroup can drive this measure and thereby not provide a clear sense of how the selected subgroup relates to the

true subgroup. However this measure provides important informationl for a phase 3 trial. Future research could consider measures that capture performance within some tolerance. For example, we could ask do we capture at least 80% of the correct group or do we include no more than 10% of a group that does not benefit? These measures are akin to sensitivity and specificity.

### *7.2 Modifiable Aspects of the Methods*

Performance with respect to our optimality criterion depends on the decision rule and utility along with the selected values of $\alpha_F$ and $n_{leaves}$ or $\ell$. We focus on the use of a decision rule and utility that selects the minimum between the p-value from a subgroup search and the p-value from the full population. Each p-value is standardized by the corresponding critical value derived via simulations that assume knowledge of the true covariate distribution. The critical value is determined with respect to the parameter $\alpha_F$, which relates the strength of our belief in homogeneous effects. The collection of candidate subgroups is defined by the tuning parameters of $n_{leaves}$ or $\ell$ and the covariates under consideration. Using either greater values of the tuning parameters or more covariates will increase the number of candidate subgroups, particularly for LR as it allows all possible combinations whereas SHAPES considers a subset of all possible combinations. When using an unrestricted search for all binary covariates, as exemplified by LR, performance of correct subgroup identification deteriorates rapidly as the number of covariates or the complexity of subgroup definitions under consideration increases. If the tuning parameter matches exactly and the subgroup definition is one or two covariates, we can do well for most measures. One way to overcome the multiplicity induced by an unrestricted search is to restrict the number of subgroups under consideration to avoid power loss and/or subgroup misidentification, which we explore through SHAPES by only considering subgroups that are connected and convex or co-convex. With SHAPES, we also add the ability to allot some $\alpha$ to each depth, thereby allowing finer control over the complexity of the identified subgroup.

There are several limitations to consider regarding aspects of our methods and possibilities for future research. In our simulations we only consider 1:1 randomization but other ratios could be evaluated. Regarding our utility, it only considers p-values. Incorporating additional data may improve performance, for example the prevalence of a subgroup. We also select the most significant p-value, but there could be several significant p-values and selecting between them in a systematic way may improve performance. The critical value for a subgroup is derived assuming knowledge of the true covariate distribution. For the candidate subgroups, with both LR and SHAPES we order the collection of candidate subgroups so that if we consider two variable combinations we

also consider each of the one variable combinations. Relatedly, since selecting the correct tuning parameter improves performance, particularly for LR, methods such as cross validation could aid correct identification.

Overall, we consider both binary and continuous outcomes with most of our results about the former. A disease area where subgroup searching has the potential to show improvements in the drug discovery process is oncology, where the heterogeneity of cancer makes targeted therapies of particular interest. Often in this area, time-to-event outcomes like progression-free survival and overall survival are used as endpoints in phase 2 and 3 trials with either the log-rank test or the Cox proportional hazards model to assess significance. To modify our methods to use such an outcome is straightforward: the LogicReg package in R already includes the Cox model and SHAPES is easily modified to use the Cox model. However, designing simulation studies for this area may be more complex, and require consideration of several additional parameters related to the censoring distribution and the effects over time.

### *7.3 Characteristics of the Study Population*

Performance with respect to an optimality criterion also depends on characteristics of the population under study. To describe the population we use several parameters, including: predictive subgroup prevalence and the related covariate distribution, subgroup definitions, the magnitude and frequency of heterogeneous and homogeneous effects, prognostic effects, and the baseline prevalence or variance. With a binary endpoint, we find if we hold the subgroup effect fixed the best performance occurs with a subgroup prevalence of 50%, with performance deteriorating as we approach 0 or 100%. If we hold the marginal effect fixed and allow the subgroup effect to increase while decreasing prevalence we do well for small prevalences. The importance of this depends on whether investigators are more likely to think in terms of an effect for a subgroup or the marginal effect driven by a subgroup as we know under larger effects LR and SHAPES can recover the correct group more often, but it is not clear for which specific subgroup effect an investigator should plan. In this dissertation, we opt to describe the subgroup effect as a multiple of the homogeneous effect that a trial's power was calibrated with respect to, primarily to aid in interpretation, but under a different subgroup alternative or specification of the effect the performance will differ. Additionally, the ability to detect both heterogeneous or homogeneous effects depends on the baseline rate for binary endpoints or the variance for continuous endpoints. While these are nuisance parameters for our question they relate to the ability to calibrate methods for type I error control and power and we observe differences in performance by these parameters.

Overall, the scenarios we explore in this dissertation are a small subset of those possible. One aspect that will be different for a real world trial is the covariate distribution. In our simulations we mostly use independent covariates where non-informative covariates have a prevalence of 0.5. While this is unlikely, our results suggest selection of a subgroup with a prevalence around 0.5, which in the case of $K = 1$ corresponds to a covariate with a prevalence of 0.5. While it is unreasonable to always assume a subgroup prevalence of 0.5 when the identity of the subgroup is unknown. However, we find, that to have reasonable power to detect a subgroup we should have a prevalence around 0.5. Future research could focus on developing methods that restrict the collection of candidate subgroups to a specific range of prevalences. An additional restriction to the candidate subgroups that we did not explore is to only consider one value of a covariate; for example, if the covariate is HER2 status and we expect a drug to either work in all patients or HER2-positive patients we would not consider HER2-negative patients in the subgroup search.

Our simulation studies are limited to a mapping from a binary covariate space to an indicator for the subgroup, however the relationship between the covariates and a subgroup could be more nuanced. For example, a continuous covariate could have a linear relationship with the outcome, or in an extreme case the interaction of multiple non-linear continuous variables. The implication here is that the separation between the predictive subgroup and the rest of the population may not be as clean as our approach. In that case, what is the *correct* subgroup is not readily defined. One possible subgroup definition is those subjects with an effect greater than or equal to the MCID. Additionally, while we focus on the clean case of a quantitative interaction where a subgroup has a positive treatment effect and the complement of the subgroup does not, it is possible that there is a qualitative interaction, where some benefit and some worsen with treatment. This could be particularly true if we think not only in terms of the primary outcome but the overall risk to benefit profile. We observe under simple prognostic effects scenarios that control for type I error for LR was violated but not for SHAPES. Under more complicated situations for prognostic effects this may not hold. Further investigating this is important. Overall, future research on this topic could consider other data generation mechanisms. The many parameters involved with methods and data generation for subgroups suggest the answer to the question of when a subgroup search will be fruitful is specific to a particular drug at hand and investigator beliefs. The development of a flexible R package where parameters could be specified would allow investigators to use their expectations to explore the cost and benefits of subgroup searching.

Relatedly, the magnitude of the predictive effect is important to subgroup identification. We mostly limit simulations to focus on subgroup effects that were equal to or less than two times

the effect for which the study was powered. While considering larger effects would improve the performance of our methods, we feel the range we considered is reasonable. Nonetheless, it is important to determine which effects we can detect with a specific power.

We did not directly address how changing the population enrolled impacts our ability to identify a subgroup. Yet the inclusion and exclusion criteria define a population for a trial and the presence of a predictive subgroup is relative to those enrolled. If the criteria limit the trial to those only in the predictive subgroup, then necessarily the trial population has a homogeneous effect. Future research could consider what is the best population to enroll. For example, we can ask when is it better to first target a narrow indication where there is more evidence in favor of an effect and consider expanding the indication later and when is it better to achieve a broader indication? In summary, we could ask whether it is better to enroll a small group that has a greater probability of an effect or to enroll a larger group. Such research could be accomplished via priors as used in Chapter 5.

### *7.4 Number and Order of Covariates Under Consideration*

In using $K$ covariates to describe our population we restrict ourselves to the information available in those covariates to explain differences in outcomes. We primarily focus on four covariates as it provides a reasonable amount of exploration, but we also consider $K \in \{1, 2, 8\}$. In our simulations, we use true subgroup definitions with one or two covariates mostly and it follows by increasing $K$ we increase the noise and performance decreases. Implicitly, we had an assumption of a fixed ordered entry of covariates, which favors informative covariates and gives more optimistic results than if covariates are selected through some random process. As an example if $K = 1$ and we always consider $X_1$ as the first variable and it is the true subgroup definition, we will perform better than if we select one covariate from the collection $\{X_1, X_2, X_3, X_4\}$ with some probability for each. But an investigator when determining an analysis plan has a collection of many covariates he or she can choose from, and may not always select the most explanatory variables first. Accounting for this process in the search for predictive subgroups is then important because without any adjustment results are overly optimistic. Future research should consider simple probabilities for which variables are included in the subgroup search and describe the change in performance as the rates for inclusion of explanatory variables changes.

### *7.5 Presentation of Results*

An additional area that we develop is the presentation of results from predictive subgroup identification methods. To our knowledge, this dissertation is the first work to simultaneously present

decisions from simulation studies under null, heterogeneous and homogeneous effects to facilitate a comparison of performance across several possible truths. Our quilt plot provides an easy assessment of the frequency of rejection for subgroups of various prevalences and effects and provides summaries of performance within subregions to identify how often we identify groups within a range of prevalences and effects. With a binary covariate space, a quilt plot is straightforward to generate. Extending a quilt plot to a continuous covariate space requires the selection of thresholds, which the quilt plots can be animated to demonstrate how the distribution of subgroups changes for different thresholds.

### *7.6 Collection of Drugs and Phase II-III Simulations*

The extent to which this research has relevance for general clinical trial policy is not clear and one critical factor is the true frequency of heterogeneous and homogeneous effects. We model these effects by considering various priors on the possible effects in conjunction with several different collections of subgroup effects. The value of using SHAPES depends on the prior; with a greater frequency of subgroups performance improves. We also extend our results at phase 2 to conduct a phase 3 trial in the identified population and find under a high prevalence of subgroup effects we identify more drugs using SHAPES.

In describing the frequencies of heterogeneous effects we define a simple collection of subgroups, but in truth the collection could consist of a much greater variety of subgroup definitions. Which subgroups to include in a collection to be evaluated could be driven by the investigators' scientific expertise on a specific drug, but even then the truth is not known. Future research could explore additional definitions.

We use a single phase 2 trial to identify a subgroup and explore how the phase 2 decision extends to the final decision in a drug discovery process under a variety of scenarios. While we focus on moving from phase 2 to 3, our approach could be reconsidered as an adaptive enrichment design under appropriately calibrated operating characteristics, which leads to the question of the best time to conduct a subgroup search. It is possible that the additional data provided by a phase 3 trial could provide advantages, possibly even by pooling data across trials or by developing a candidate subgroup on phase 2 data that is later confirmed in a phase 3 trial. Determining the best time to conduct such a search is an important topic for future research. Additionally, the use of a phase 3 adaptive design could also be considered, where at some interim analysis before the end of enrollment a decision is made regarding which patients to enroll for the remainder of the trial.

### *7.7 Availability of Computer Programs*

Software to perform subgroup searches is lacking. ASD is not available in any program and SIDES is available as a program in Excel, which allows for retrospective implementation for an individual trial but prevents prospective adoption of these methods as they cannot be readily evaluated under a variety of scenarios. Appendix B contains a sample program in R for LR that can be used to assess operating characteristics for various simulation settings. Appendix C contains an equivalent version for the SHAPES method.

### *7.8 Prospective or Restrospective Application*

In our simulations we assume the use of LR or SHAPES is applied prospectively. This is consistent with simulations from ASD. SIDES considered both prospective and retrospective application of their method, where some simulations assume a negative trial result and rely on a scenario where the average treatment effect is zero but underlying this is a strong positive predictive subgroup effect and a strong negative effect in the complement of the subgroup. In this scenario, there are still some simulated trials that will reject for an effect in the full population as the type I error rate is non-zero. This set-up is quite specific and may not reflect an investigator's belief about the underlying truth when the result of a trial is not significant.

Any method for a predictive subgroup identification could be retrospectively applied to search for a predictive subgroup after the full population analysis is not significant, and investigators may be eager to implement such a search to salvage a negative trial. Evaluating the limitations and benefits of this application would be informative to either promote or discourage its use. Possible considerations include the type I error for this search and the corresponding experiment-wise inflation of type I error.

### *7.9 Summary of Suggestions for Future Research and Work*

We now summarize our recommendations for future research in this area discussed in the preceding sections. We list them in order based on what we believe will be most important for further developing and understanding methods for subgroup search and identification.

- To add additional restrictions to the subgroup definitions under consideration, including the restriction to a certain range of prevalences.
- The development of methods for a subgroup search using continuous covariates and covariate distributions with a variety of correlation structures.

- The implementation of simulations that use a time-to-event endpoint.
- To incorporate the use of cross validation in the selection of tuning parameters.
- The development of a package in R that would allow investigators to perform simulations under a variety of assumed population and tuning parameters to determine potential gains in performance for their specific scenario.
- To identify the optimal timing of a subgroup search in the context of the drug discovery process.
- To also explore these methods in the context of retrospective application.

These recommendations are meant to motivate further research regarding subgroups that is the most relevant for current clinical trials. Additional topics may be important as well.

### *7.10 Conclusion*

The area of subgroup exploration may improve the efficiency of the drug discovery process and to further inform research in personalized medicine. Yet, there are many parameters to consider when searching for a predictive subgroup, making the problem complex. In simulation studies, with simple scenarios and large effects we can capture the true subgroup, but how closely this aligns with actual clinical trials is not clear. In this dissertation we develop several novel metrics to evaluate performance in the subgroup context and propose two methods for subgroup search and identification. We also provide a framework for describing several alternatives of interest and evaluating the average performance based on an investigator's prior beliefs. There is still much work and exploration to be done for this topic.

# BIBLIOGRAPHY

[1] Luca Arcaini, Michele Merli, Francesco Passamonti, Raffaele Bruno, Ercole Brusamolino, Paolo Sacchi, Sara Rattotti, Ester Orlandi, Elisa Rumi, Virginia Ferretti, et al. Impact of treatment-related liver toxicity on the outcome of hcv-positive non-hodgkin's lymphomas. *American journal of hematology*, 85(1):46–50, 2010.

[2] Robert Bazell. *Her-2: the making of Herceptin, a revolutionary treatment for breast cancer*. Random House, 2011.

[3] Ellin Berman, Glenn Heller, J Santorsa, Susan McKenzie, Timothy Gee, Sanford Kempin, Subash Gulati, Michael Andreeff, Jonathan Kolitz, and Janice Gabrilove. Results of a randomized trial comparing idarubicin and cytosine arabinoside with daunorubicin and cytosine arabinoside in adult patients with newly diagnosed acute myelogenous leukemia. *Blood*, 77(8):1666–1674, 1991.

[4] Ellin Berman, Peter Wiernik, Ralph Vogler, Enrique Vélez-Gárcia, Alfred Bartolucci, and Fredrick S Whaley. Long-term follow-up of three randomized trials comparing idarubicin and daunorubicin as induction therapies for patients with untreated acute myeloid leukemia. *Cancer*, 80(S11):2181–2185, 1997.

[5] Werner Brannath, Emmanuel Zuber, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, and Amy Racine-Poon. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in medicine*, 28(10):1445–1463, 2009.

[6] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[7] Hartmut Döhner, Daniel J Weisdorf, and Clara D Bloomfield. Acute myeloid leukemia. *N Engl J Med*, 373:1136–1152, 2015.

[8] Oya Ekin, Peter L Hammer, and Alexander Kogan. On connected boolean functions. *Discrete Applied Mathematics*, 96:337–362, 1999.

[9] Robert H El-Maraghi and Elizabeth A Eisenhauer. Review of phase ii trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase iii. *Journal of Clinical Oncology*, 26(8):1346–1354, 2008.

[10] Jennifer L Fackler and Amy L McGuire. Paving the way to personalized genomic medicine: steps to successful implementation. *Current pharmacogenomics and personalized medicine*, 7(2):125, 2009.

[11] US Food, Drug Administration, et al. Draft guidance for industry: Enrichment strategies for clinical trials to support approval of human drugs and biological products. *Rockville, Maryland: FDA*, 2012.

[12] Boris Freidlin, Wenyu Jiang, and Richard Simon. The cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2):691–698, 2010.

[13] Boris Freidlin, Lisa M McShane, and Edward L Korn. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 2010.

[14] Boris Freidlin and Richard Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11(21):7872–7878, 2005.

[15] Lawrence M Friedman, Curt Furberg, David L DeMets, et al. *Fundamentals of clinical trials*, volume 4. Springer, 2010.

[16] Geoffrey T Gibney and Jonathan S Zager. Clinical development of dabrafenib in braf mutant melanoma and other malignancies. *Expert opinion on drug metabolism & toxicology*, 9(7):893–899, 2013.

[17] Nicolas Girard. Crizotinib in alk-positive lung cancer. *The Lancet Oncology*, 13(10):962–963, 2012.

[18] Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51, 2014.

[19] Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4):347–356, 2011.

[20] Said Abdullah Khelwatty, Sharadah Essapen, Izhar Bagwan, Margaret Green, Alan Michael Seddon, and Helmout Modjtahedi. Co-expression of her family members in patients with dukes c and d colon cancer and their impacts on patient prognosis and survival. *PloS one*, 9(3):e91139, 2014.

[21] Ilya Lipkovich and Alex Dmitrienko. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of biopharmaceutical statistics*, 24(1):130–153, 2014.

[22] Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601–2621, 2011.

[23] Baldur P Magnusson and Bruce W Turnbull. Group sequential enrichment design incorporating subgroup selection. *Statistics in medicine*, 32(16):2695–2714, 2013.

[24] Shigeo Masuda and Juan Carlos Izpisua Belmonte. Trametinib for patients with advanced melanoma. *The Lancet Oncology*, 13(10):e409, 2012.

[25] Grant A McArthur, Paul B Chapman, Caroline Robert, James Larkin, John B Haanen, Reinhard Dummer, Antoni Ribas, David Hogg, Omid Hamid, Paolo A Ascierto, et al. Safety and efficacy of vemurafenib in braf v600e and braf v600k mutation-positive melanoma (brim-3): extended follow-up of a phase 3, randomised, open-label study. *The lancet oncology*, 15(3):323–332, 2014.

[26] Cyrus R Mehta and Ping Gao. Population enrichment designs: case study of a large multinational trial. *Journal of biopharmaceutical statistics*, 21(4):831–845, 2011.

[27] Filippo Montemurro, Aleix Prat, Valentina Rossi, Giorgio Valabrega, Jeff Sperinde, Caterina Peraldo-Neia, Michela Donadio, Patricia Galván, Anna Sapino, Massimo Aglietta, et al. Potential biomarkers of long-term benefit from single-agent trastuzumab or lapatinib in her2-positive metastatic breast cancer. *Molecular oncology*, 8(1):20–26, 2014.

[28] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[29] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to braf (v600e) inhibition through feedback activation of egfr. *Nature*, 483(7388):100–103, 2012.

[30] Aleix Prat, Giampaolo Bianchini, Marlene Thomas, Anton Belousov, Maggie CU Cheang, Astrid Koehler, Patricia Gómez, Vladimir Semiglazov, Wolfgang Eiermann, Sergei Tjulandin, et al. Research-based pam50 subtype predictor identifies higher responses and improved survival outcomes in her2-positive breast cancer in the noah study. *Clinical Cancer Research*, 20(2):511–521, 2014.

[31] Michael Rosenblum and Mark J van der Laan. Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, 98(4):845–860, 2011.

[32] Ingo Ruczinski, Charles Kooperberg, and Michael LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[33] Manish R Sharma, Walter M Stadler, and Mark J Ratain. Randomized phase ii trials: a long-term investment with promising returns. *Journal of the National Cancer Institute*, 103(14):1093–1100, 2011.

[34] Xin Sun, Matthias Briel, Stephen D Walter, Gordon H Guyatt, et al. Is a subgroup effect believable? updating criteria to evaluate the credibility of subgroup analyses. *Bmj*, 340:c117, 2010.

[35] Le-Chi Ye, Tian-Shu Liu, Li Ren, Ye Wei, De-Xiang Zhu, Sheng-Yong Zai, Qing-Hai Ye, Yiyi Yu, Bo Xu, Xin-Yu Qin, et al. Randomized controlled trial of cetuximab plus chemotherapy for patients with kras wild-type unresectable colorectal liver-limited metastases. *Journal of Clinical Oncology*, pages JCO–2012, 2013.

# Appendix A

# DERIVATIONS

## *A.1 Subgroup Interaction Model*

For $i \in \{1, ..., n\}$, consider the linear model where $\epsilon_i \sim (0, \sigma^2)$ and $t_i$ is a binary indicator for treatment or control and $l_i$ is an indicator for subgroup membership:

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 (t_i * l_i) + \epsilon_i$$

The design matrix is:

$$X_{n\times 3} = \begin{pmatrix} \vec{1}_t & \vec{1}_t & \vec{L}_t \\ \\ \vec{1}_{(n-t)} & \vec{0}_{(n-t)} & \vec{0}_{(n-t)} \end{pmatrix}$$

Define the following sums:

$$\begin{aligned} n_t &= \sum_{i=1}^{n} I(t_i = 1) \\ n_\ell &= \sum_{i=1}^{n} (I(t_i = 1) * I(l_i = 1)) \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{y}_t &= \frac{1}{n_t} \sum_{i=1}^{n} (y_i * (I(t_i = 1)) \\ \bar{y}_\ell &= \frac{1}{n_\ell} \sum_{i=1}^{n} (y_i * I(t_i = 1) * I(l_i = 1)) \end{aligned}$$

The inverse of $X^T X$ is:

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n-n_t} & -\frac{1}{n-n_t} & 0 \\ \\ -\frac{1}{n-n_t} & \frac{n-n_\ell}{(n-n_t)(n_t-n_\ell)} & -\frac{1}{n_t-n_\ell} \\ \\ 0 & -\frac{1}{n_t-n_\ell} & \frac{n_t}{n_\ell(n_t-n_\ell)} \end{pmatrix}$$

Using the Gauss-Markov theorem the BLUE of $\beta = (\beta_0, \beta_1, \beta_2)$ is $\hat{\beta} = (X^T X)^{-1} X^T Y$, in this case:

$$\hat{\beta} = \begin{pmatrix} \frac{n}{n-n_t}\bar{y} - \frac{n_t}{n-n_t}\bar{y}_t \\ \frac{n_t}{n-n_\ell}\bar{y}_t - \frac{n_\ell}{n_t-n_\ell}\bar{y}_\ell - \frac{n}{n-n_t}\bar{y} \\ \frac{n_t}{n_t-n_\ell}(\bar{y}_\ell - \bar{y}_t) \end{pmatrix} = \begin{pmatrix} \frac{n}{n-n_t}\bar{y} - \frac{n_t}{n-n_t}\bar{y}_t \\ \bar{y}_\ell - \hat{\beta}_0 - \hat{\beta}_2 \\ \frac{n_t}{n_t-n_\ell}(\bar{y}_\ell - \bar{y}_t) \end{pmatrix}$$

Define $p_{\ell t} = \frac{n_\ell}{n_t}$. The Wald test for $H_0 : \beta_2 \leq \eta$ vs. $H_A : \beta_2 > \eta$ can be constructed as:

$$Z_{\beta_2} = \frac{\frac{1}{1-p_{\ell t}}(\bar{y}_\ell - \bar{y}_t) - \eta}{\sqrt{\frac{\hat{\sigma}_I^2}{n_t p_{\ell t}(1-p_{\ell t})}}} = \frac{\sqrt{n_t odds_{\ell t}}((\bar{y}_\ell - \bar{y}_t) - \eta)}{\hat{\sigma_I}}$$

### *A.2 Stratified Analysis*

Consider the set of indices S= $\{i : l_i = 1\}$. Note that we allow $l_i = 1$ when $t_i = 0$. The linear model with a main effect for treatment is:

$$y_i = \gamma_0 + \gamma_1 t_i + \epsilon_i, \text{ for } i \in S$$

Define

$$\begin{aligned} n_S &= \sum_{i=1}^{n} I(i \in S) \\ n_{St} &= \sum_{i=1}^{n} (I(i \in S) * I(t_i = 1)) \\ \bar{y}_S &= \frac{1}{n_S} \sum_{i=1}^{n} (I(i \in S) * y_i) \end{aligned}$$

With the following design matrix:

$$X_{n_S \times 2} = \begin{pmatrix} 1_{n_{St}} & 1_{n_{St}} \\ 1_{n_S - n_{St}} & 0_{n_S - n_{St}} \end{pmatrix}$$

And the inverse, setting $p_{St} = \frac{n_{St}}{n_S}$ :

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n_S - n_{St}} & \frac{-1}{n_S - n_{St}} \\ \frac{-1}{n_S - n_{St}} & \frac{n_S}{n_{St}(n_S - n_{St})} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_S - n_{St}} & \frac{-1}{n_S - n_{St}} \\ \frac{-1}{n_S - n_{St}} & \frac{1}{n_S p_{St}(1-p_{St})} \end{pmatrix}$$

With the following BLUE, $\hat{\gamma}$, of $\gamma = (\gamma_0, \gamma_1)$ as:

$$\hat{\gamma} = \begin{pmatrix} \frac{n_S}{n_S - n_{St}} \bar{y}_S - \frac{n_{St}}{n_S - n_{St}} \bar{y}_{St} \\ \\ \frac{n_S}{n_S - n_{St}} \bar{y}_{St} - \frac{n_S}{n_S - n_{St}} \bar{y}_S \end{pmatrix} = \begin{pmatrix} \frac{n_S}{n_S - n_{St}} \bar{y}_S - \frac{n_{St}}{n_S - n_{St}} \bar{y}_{St} \\ \\ \bar{y}_{St} - \hat{\gamma}_0 \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\gamma}_0 \\ \\ \frac{1}{1-p_{St}}(\bar{y}_{St} - \bar{y}_S) \end{pmatrix}$$

Developing the Wald test for $H_0 : \gamma_1 \leq \eta$ vs. $H_A : \gamma_1 > \eta$ is then:

$$Z^S_{\gamma_1} = \frac{\frac{1}{1-p_{St}}(\bar{y}_{St} - \bar{y}_S) - \eta}{\sqrt{\frac{\hat{\sigma}^2_S}{n_S p_{St}(1-p_{St})}}} = \frac{\sqrt{n_S odds_{St}}((\bar{y}_{St} - \bar{y}_S) - \eta)}{\hat{\sigma}_S}$$

### *A.3 Full Population Analysis*

If we consider the full population analysis, results from the previous section generalize by noting that $n_{St} = n_t, n_S = n, p_t = \frac{n_t}{n}, \bar{y}_S = \bar{y}$ and $\bar{y}_{St} = \bar{y}_t$. For the full population then:

$$\hat{\gamma} = \begin{pmatrix} \frac{1}{1-p_t} \bar{y} - \frac{p_t}{1-p_t} \bar{y}_t \\ \\ \frac{1}{1-p_t}(\bar{y}_t - \bar{y}) \end{pmatrix}$$

And the Wald test for the full study population is:

$$Z^F_{\gamma_1} = \frac{\frac{1}{1-p_t}(\bar{y}_t - \bar{y}) - \eta}{\sqrt{\frac{\hat{\sigma}^2_F}{n p_t (1-p_t)}}}$$

# Appendix B

## SAMPLE R CODE TO GENERATE DATA

Code in this appendix can be used to generate data that can be used by the code in Appendix C or D to run LR or SHAPES, respectively.

```
####  The function takes a vector of covariate probabilities (pVec), sample size (n),
####  a baseline rate (bl), an overall treatment effect (eta), a subgroup effect
####  (thetaS) and a subgroup definition (defS) and returns a list with the
####  covariates (X), the treatment assignment (Trt), and outcome (y)

genOne <- function(n, pVec, bl, eta, thetaS, defS){
  k <- length(pVec)
  ### Create the matrix of covariates based on p
  X <- matrix( rbinom( I(n*k),  1 ,  pVec), nrow=n, ncol=k, byrow=T)

  ### Test if n is even or odd for treatment assignment
  if( (n %% 2)==0){
    Trt <- c(rep(1, n/2), rep(0, n/2))
  }else{Trt <- c(1, rep(1, floor(n/2)), rep(0, floor(n/2)))}

  ### Now derive the subgroup truth
  if(defS==1){ # scene A
    pred <- X[,1]
  }else if(defS==2 & k > 1){  #scene B
    pred <- pmax(X[,1], X[,2])
  }else if(defS==3 & k > 1){  #scene C
    pred <- pmin(X[,1], X[,2])
  }else if(defS==26 & k > 3){
    pred <- pmax(pmin(X[,1], X[,2]), pmin(X[,3], X[,4]))
  }else{return(print('Issue with parameter settings'))}

  ### Generate the true parameter for each subject
  sumPar <- bl + eta + thetaS
  if(bl >=0 & eta >= 0 & thetaS >=0 & sumPar > 0 & sumPar < 1){
    outProb <- bl + I(eta * Trt) + I(thetaS * (Trt * pred) )
  }else{return(print('Issue with values or sum of effect parameters'))}

  y <- rbinom(n, 1, outProb)
  return(list(X, Trt, y))
}
```

./code/AppB.R

# Appendix C

# SAMPLE R CODE FOR LOGISTIC REGRESSION

Code this appendix runs a logic regression simulation study using the LR-A with stratification testing. First run the code in Appendix B, which will generate simulated trial data. Then run the code in each section in order. The last section contains the code with the simulation parameters, with the details in comments, and the line that will run a simulation. Note that *nullReps* is used to determine the number of null trials run to determine the critical value and *altReps* the number of trials under the alternative. The rate of any rejection and correct identification are returned.

## *C.1 Functions to Perform a Single LR-A with Stratification Analysis*

```
#### This function takes a list form of a dataset as produced
####  by the genOne function and applies LogicRegression to it
library(LogicReg)

# take a specific kind of product to calculate probabilities for each possible case
vecProd <- function(vec1, vec2p){
  return(prod((vec1 * vec2p) + ((1-vec2p) * (1-vec1))))
}

extract1 <- function(mat1){
         if(dim(mat1)[1]==1){
            val1 <- 0.001
         }else{
            val1 <- mat1[2,1]
         }
         return(val1)
}

#this is a function to create the truth table based on a pre-specified scenario
####  it takes as input the scene as in integer 1, 2, 3, 26 for A, B, C, Z
####  and K where K should be appropriately selected for the subgroup def
scenario <- function(scene, k){
    krec <- k
    if(k<3){k <- 3}

    varCt <- 0

  #### create a table with all possible combinations
  truthSub1 <- list()
```

```
30   for(i in 1:k){truthSub1[[i]] <- c(1,0)}
31   truthSub2 <- expand.grid(truthSub1)
32   truthtable <- truthSub2[k:1]
33   colnames(truthtable) <- paste('X', 1:k, sep='')
34
35   ######## scenario A ########################################
36   if(scene==1){
37     lpredF <- truthtable[,'X1']
38       varCt <- 2}
39   ########## scenario B #####################################
40   if(scene==2){
41     lpredF <- pmax(truthtable[,'X1'], truthtable[,'X2'])
42       varCt <- 3}
43
44   ########## scenario C #####################################
45   if(scene==3){
46     ### remember that pmax is saying or and pmin is saying and
47     lpredF <- pmin(truthtable[,'X1'], truthtable[,'X2'])
48       varCt <- 3}
49   ###################scenario Z ################################
50   if(scene==26){
51     lpredF <- pmax(pmin(truthtable[,'X1'], truthtable[,'X2']),
52         pmin(truthtable[,'X3'], truthtable[,'X4']))
53      varCt <- 4}
54
55   ##### here we are checking to make sure things set up correctly
56   repeat{
57     mod1 <- logreg(lpredF, truthtable[,1:k], type=2, ntrees=1, select=1)$model
58     lpred <- mod1$trees[[1]]
59     e1 <- eval.logreg(mod1, truthtable)
60       if( sum((e1-lpredF)^2)==0){
61         break
62       }
63   }
64
65    # i artificially added a variable in - now need to remove it
66    if(krec < k){
67
68      df <- as.data.frame(cbind(truthtable[, 1:krec], lpredF))
69      df <- df[duplicated(df),]
70      truthtable <- as.matrix(df[, 1:krec])
71      lpredF <- df[, (krec+1)]
72     }
73   return(list(truthtable, lpredF, NA, lpred, NA))
74 }
75
76
77 ######## VERSION of Logic Regression FOR K>2 ################
78 ######## Takes the trial data and the number of leaves as inputs
79 ######## Returns a truthtable vector, the fulldataset predictive
```

```
########    subgroup vector c(LpredMod2, rep(0)), the fulldataset
########    prognostic vector (fitAllProg)
LR1 <- function(trialList, leaves){

  covar <-trialList[[1]]
  Trt <- trialList[[2]]
  y <- trialList[[3]]
  k <- ncol(covar)

  #### create a table with all possible combinations
  truthSub1 <- list()
  for(i in 1:k){truthSub1[[i]] <- c(1,0)}
  truthSub2 <- expand.grid(truthSub1)
  truthsub <- truthSub2[k:1]

  #### set up code so it's ready
  txt <- cbind(covar[I(Trt==1), ], 1, y=y[I(Trt==1)])
  con <- cbind(covar[I(Trt==0), ], 0, y=y[I(Trt==0)])
  fullSub <- cbind(covar, Trt, y)

  # Let's fit the control data to assess for prognostic factors
  fitCon <- logreg(con[,'y'], con[,1:k], type=3, select=1, ntrees=1, nleaves=leaves)$model
  progCon <- fitCon$trees[[1]]

  ### fitTxt is the prognostic indicator for treatment data and fitAllProg is the same for all
       data
  fitTxt <- eval.logreg(progCon, txt[,1:k])
  fitAllprog <- eval.logreg(progCon, fullSub[,1:k])
  if(progCon$coef<0){fitTxt <- 1-fitTxt; fitAllprog <- 1 - fitAllprog}

  # now fitting logistic regression on txt data adjusting for prognostic factors,
  #     and pulling out what we care about
  fitLR1 <- logreg(txt[,'y'], txt[,1:k], sep=fitTxt, type=3, select=1, ntrees=1, nleaves=leaves)
  fit1pred<- eval.logreg(fitLR1$model, txt[,1:k])
  fit1predCon <- eval.logreg(fitLR1$model, con[, 1:k])

  truth1pred <- eval.logreg(fitLR1$model, truthsub)
  #check if the coefficient is negative we'll flip
  coefLR1 <- fitLR1$model$coef
  if(!is.na(coefLR1[3])){ if(coefLR1[3]<0){fit1pred <- 1-fit1pred; truth1pred <- 1-truth1pred;
        fit1predCon <- fit1predCon}}

  return(list(truth1pred, c(fit1pred, rep(0, nrow(con))), fitAllprog, c(fit1pred, fit1predCon)))
}



################################### Version of LR for K<=2 ##########################
######## Takes the trial data and the number of leaves as inputs
######## Returns a truthtable vector, the fulldataset predictive
```

```
129 ########    subgroup vector c(LpredMod2, rep(0)), the fulldataset
130 ########    prognostic vector (fitAllProg)
131
132 LR1small <- function(trialList, leaves){
133   covar <-trialList[[1]]
134   Trt <- trialList[[2]]
135   y <- trialList[[3]]
136   k <- ncol(covar)
137
138   #### create a table with all possible combinations
139   truthSub1 <- list()
140   for(i in 1:k){truthSub1[[i]] <- c(1,0)}
141   truthSub2 <- expand.grid(truthSub1)
142   truthsub <- truthSub2[k:1]
143
144   #### set up code so it's ready
145   txt <- cbind(covar[I(Trt==1), ], 1, y=y[I(Trt==1)])
146   con <- cbind(covar[I(Trt==0), ], 0, y=y[I(Trt==0)])
147   fullSub <- cbind(covar, Trt, y)
148
149
150    if(k==1){ truthsub <- matrix(truthsub, nrow=length(truthsub), ncol=k) }
151
152   # Let's fit the control data to assess for prognostic factors
153   if(k>2){
154     fitCon <- logreg(con[,'y'], con[,1:k], type=3, select=1, ntrees=1, nleaves=leaves)$model
155     progCon <- fitCon$trees[[1]]
156
157     ### fitTxt is the prognostic indicator for treatment data and fitAllProg is the same for all
          data
158     fitTxt <- eval.logreg(progCon, txt[,1:k])
159     fitAllprog <- eval.logreg(progCon, fullSub[,1:k])
160     if(progCon$coef<0){fitTxt <- 1-fitTxt; fitAllprog <- 1 - fitAllprog}
161   }else{
162
163     mod <- glm(con[,'y']~con[, 1:k], family="binomial")
164     summ <- summary(mod)$coef
165     co <- matrix(summ, nrow=(k+1), ncol=1)
166     co <- ifelse(summ[,4] <.20, co, 0)
167
168     des <- matrix( cbind(rep(1, nrow(txt)+nrow(con))), rbind(txt[, 1:k], con[, 1:k]), nrow=nrow(
          txt)+nrow(con), ncol=k+1)
169      # so this will be n by 1 matrix
170     prognostic <- (des %*% co)
171     progAll <- 1*(prognostic > 0)
172
173     progTxt <- progAll[1:nrow(txt)]
174     call <- list(progAll, progTxt)
175     fitTxt <- call[[2]]
176     fitAllprog <- call[[1]]
```

```
177   }
178
179   m1 <- glm(txt[,'y'] ~ txt[,1] + fitTxt, family=binomial)
180   oldRSS <- m1$deviance
181
182   if(extract1(summary(m1)$coef) >0){
183     truth1pred <- truthsub[,1]
184     fit1pred <- txt[,1]
185       fit1predcon <- con[,1]
186   }else{
187     truth1pred <- 1-truthsub[,1]
188     fit1pred <- 1-txt[,1]
189         fit1predcon <- 1-con[,1]
190
191   }
192
193   if(k>1){
194     m3 <- glm(txt[,'y'] ~ (txt[,2]) + fitTxt, family=binomial)
195       newRSS <- m3$deviance
196
197     if(newRSS< oldRSS & extract1(summary(m3)$coef)> 0 ){
198       truth1pred <- truthsub[,2]
199       fit1pred <- txt[,2]
200           fit1predcon <- con[,2]
201       oldRSS <- newRSS
202     }else if(newRSS < oldRSS){
203       truth1pred <- 1-truthsub[,2]
204       fit1pred <- 1-txt[,2]
205           fit1predcon <- 1-con[,2]
206       oldRSS <- newRSS
207     }
208   }
209
210
211   if(leaves==2){
212     ## so above we've programmed in the first 4 possibilities,
213     ##      now to add in the next 10 (in pairs)
214     mA <- glm(txt[,'y'] ~ I(txt[,1] * txt[,2]) +fitTxt, family=binomial)
215     newRSS <- mA$deviance
216     if(newRSS< oldRSS & extract1(summary(mA)$coef)> 0 ){
217       truth1pred <- truthsub[,1] * truthsub[,2]
218       fit1pred <- txt[,1] * txt[,2]
219           fit1predcon <- con[,1] * con[,2]
220       oldRSS <- newRSS
221     }else if(newRSS < oldRSS){
222       truth1pred <- 1-(truthsub[,1] * truthsub[,2])
223       fit1pred <- 1-(txt[,1] * txt[,2])
224           fit1predcon <- 1-(con[,1] * con[,2])
225       oldRSS <- newRSS
226     }
```

```
227
228     mB <- glm(txt[,'y'] ~ I(txt[,1] * (1-txt[,2])) +fitTxt, family=binomial)
229     newRSS <- mB$deviance
230
231     if(newRSS< oldRSS & extract1(summary(mB)$coef)> 0 ){
232       truth1pred <- truthsub[,1] * (1-truthsub[,2])
233       fit1pred <- txt[,1] * (1-txt[,2])
234           fit1predcon <- con[,1] * (1-con[,2])
235       oldRSS <- newRSS
236     }else if(newRSS < oldRSS){
237       truth1pred <- 1-(truthsub[,1] * (1-truthsub[,2]))
238       fit1pred <- 1-(txt[,1] * (1-txt[,2]))
239           fit1predcon <- 1-(con[,1] * (1-con[,2]))
240       oldRSS <- newRSS
241     }
242
243     mC <- glm(txt[,'y'] ~ I((1-txt[,1]) * txt[,2]) +fitTxt, family=binomial)
244     newRSS <- mC$deviance
245
246     if(newRSS< oldRSS & extract1(summary(mC)$coef)> 0 ){
247       truth1pred <- (1-truthsub[,1]) * truthsub[,2]
248       fit1pred <- (1-txt[,1]) * txt[,2]
249           fit1predcon <- (1-con[,1]) * con[,2]
250       oldRSS <- newRSS
251     }else if(newRSS < oldRSS){
252       truth1pred <- 1-((1-truthsub[,1]) * truthsub[,2])
253       fit1pred <- 1-((1-txt[,1]) * 1-txt[,2])
254           fit1predcon <- 1-((1-con[,1]) * 1-con[,2])
255       oldRSS <- newRSS
256     }
257
258     mD <- glm(txt[,'y'] ~ I((1-txt[,1]) * (1-txt[,2])) +fitTxt, family=binomial)
259     newRSS <- mD$deviance
260
261     if(newRSS< oldRSS & extract1(summary(mD)$coef)> 0 ){
262       truth1pred <- (1-truthsub[,1]) * (1-truthsub[,2])
263       fit1pred <- (1-txt[,1]) * (1-txt[,2])
264           fit1predcon <- (1-con[,1]) * (1-con[,2])
265       oldRSS <- newRSS
266     }else if(newRSS < oldRSS){
267       truth1pred <- 1- ((1-truthsub[,1]) * (1-truthsub[,2]))
268       fit1pred <- 1-((1-txt[,1]) * (1-txt[,2]))
269           fit1predcon <- 1-((1-con[,1]) * (1-con[,2]))
270       oldRSS <- newRSS
271     }
272   }
273   return(list(truth1pred, c(fit1pred, rep(0, nrow(con))),
274       as.vector(fitAllprog), c(fit1pred, fit1predcon)))
275 }
276
```

```
# this will give the full population p-value where we use test for txt effect
fulld <- function(trialList){
  m1 <- coef(summary(glm(trialList[[3]] ~ trialList[[2]], family="binomial")))[2,c(1,3)]

  return(c(m1[1], pnorm(m1[2], lower.tail=F)))
}

### The below function takes the trial data in list form and a vector for the
###   predictive subgroup and returns a p-value based on a stratified test
regress <- function(trialList, pred){

  covar <- trialList[[1]]
  Trt <- trialList[[2]]
  y <- trialList[[3]]
  fullSub <- cbind(covar, txt=Trt, y)

  # for binary data we can do this
    flag <- 0  # flag for small sample issues
  ### the last version is for the stratified analysis

     #stratified analysis
   sub1 <- fullSub[pred==1, ]
     ct <- sum(pred)

     if(ct>1 & sum(pred)>25){
       m1 <- glm(sub1[, 'y'] ~ 1, family=binomial )
       m2 <- glm(sub1[, 'y'] ~ sub1[, 'txt'], family=binomial )
      }else{flag <- 1}


     if((ct==0 | ct==1) | flag==1 ){
               return(1)
   }else{dim1 <- sum( !is.na(m2$coefficients) )

    if(dim1==1){co1 <- 0; t1 <- 1}else{co1 <- summary(m2)$coef[2,1];
            t1 <- summary(m2)$coef[2,3] }
      return(pnorm(t1, lower.tail=F))
    }
}
```

./code/AppC_1.R

### *C.2 Functions for a Simulation Study*

```

#### a simple function to calculate various stats of interest
# including overall error rate, the  predictive error    , PPV, whether trial has 0 error rate,
#     whether trial has 0 predictive error, whether trial has PPV above some threshold
trialStat <- function(decisionVec, lpredV, prob, effAll, effSub){
```

```
6       s1a <- sum(prob * (lpredV-decisionVec)^2)
7       m1a <- sum(prob * ((lpredV-decisionVec)==1))/sum(prob*lpredV)
8
9       ppv1a <- sum(prob * decisionVec * ((lpredV-decisionVec)==0))/sum(prob*decisionVec)
10      #mse1a <- sum(prob *  ((effTxt + effPred*lpredV) - (p1[1]))^2)
11       powA1a <- ifelse(s1a==0, 1, 0)
12       powC1a <- ifelse(m1a==0, 1, 0)
13       ppv1aThresh <- ifelse(ppv1a >.80, 1, 0)
14       dec1 <- ifelse(sum(decisionVec)==length(lpredV), 1, ifelse(sum(decisionVec)==0, 0 , 2))
15
16       trueEffectVec <- ((effAll*decisionVec)  +  (effSub * (decisionVec*lpredV) ))
17       wts <- (decisionVec*prob)  / (decisionVec %*% prob)
18       trueEffectSelect <- sum(wts * trueEffectVec)
19
20       c(ANY= 1*I(dec1==1 | dec1==2), CORRECT = 1*I(s1a==0) )
21  }
22
23  do.one <- function(n, p, scene, effPred, effTxt, bl, leaves, pTab=NA){
24    # number of covariates that we will consider
25    k <- length(p)
26
27    # here are we setting up a truth table to feed to logic regression
28    # this should then allows to ascertain truth trees.... PERHAPS NEED TO HAVE A CHECK HERE TO
          CONFIRM k is appropriate for the scenario?
29    set <- scenario(scene, k)     ### creation of truth table issue
30    truthtable <- set[[1]]
31    lpredV <- set[[2]]; if(effTxt>0){lpredV <- rep(1, 2^k)} ## the truth vector
32    lpred <- set[[4]] ## the truth in Boolean notation
33
34      # I will at some point need to build a different mean structure
35      prob <- apply(truthtable[, 1:k], 1, vecProd, vec2p=p)
36      ll <- nrow(truthtable)           # this should be in effect 2^k and is the number of cubes
37    # this is the truthtable subset that will get sent to the various functions
38    truthsub <- truthtable
39
40    trialList <- genOne(n, p, bl, effTxt, effPred, scene)
41
42    #### now moving on to the methods... remember to pass on fullSub and txt/con as appropriate
43    p1 <- fulld(trialList)
44
45    # the logic regression functions
46    if(k>=3){
47            p3 <- LR1(trialList, leaves)
48     }else{p3 <- LR1small(trialList, leaves)}
49
50    p10r <- regress(trialList, pred=p3[[4]])
51
52       if(is.na(pTab[1])){
53      return(c(p1, p10r))
54     }else{
```

```
    allCut <- pTab[1]
    metCut <- pTab[2]

    if(p1[2] < allCut){d1 <- rep(1, ll)}else{d1 <- rep(0, ll)}

    ckMain <- p1[2]/allCut
        ck10 <- p10r/metCut; if(is.na(ck10)){ck10 <- 5 }

        print(c(ckMain, ck10))

    if(ckMain < ck10 & ckMain <1){d10 <- rep(1, ll)}else if(ck10 <ckMain & ck10 <1){d10 <- p3
         [[1]]
      }else if(ck10 <ckMain & ck10 <1){d10 <- 1-p3[[1]]}else{d10 <- rep(0, ll)}

     print(d10)
     firstV <- c(trialStat(d10, lpredV, prob, effTxt, effPred))

    return(firstV)
  }
}


main <- function(n, p, scene, effPred, effTxt, bl=.2, leaves, alphaTot=.1, alphaFull=.05,
     nullReps=1000, altReps=1000){
  # number of covariates that we will consider
  k <- length(p)

  ### identify the critical value of interest here
    nullD <- t(replicate(nullReps, do.one(n, p, scene, 0, 0, bl, leaves, NA), simplify = "array")
         )
    print(nullD[,3]);
    subNull <- sort(subset(nullD, nullD[,2]>=alphaFull)[,3]); ct <- round(nullReps * alphaTot)
    print(subNull)
    tooSelect <- ct - (nullReps - length(subNull))
    critVal <- subNull[tooSelect]

  res1  <- t(replicate(altReps, do.one(n, p, scene,
            effPred, effTxt, bl, leaves, c(alphaTot, critVal)), simplify = "array"))

  return(c(ANY=I(sum(res1[,1])/altReps), CORRECT=I(sum(res1[,2])/altReps)))
 }
```

./code/AppC_2.R

### *C.3 Function to Run a Simulation Study with Parameter Value Specification*

```
set.seed(1)  ### set this to ensure reproducible results


```

```
4  n0 <- 200           # total sample size, >=200
5  pVec <- c(.5, .5, .5, .5)   #vector of probabilities for covariates (independent)
6  sce <- 1          # the scenario, A, B, C, Z = 1, 2, 3, 26
7                # and the length of pVec should be at least 1, 2, 2, 4
8                #   for each scenario
9  subEffect <- .2       # the subgroup treatment effect
10 allEffect <- 0        # the main treatment effect
11 baseline <- .3        # baseline response in control should be in (0, 1)
12 nleaves <- 1        # the number of leaves allowed
13 alpha <- .1         # type I error for all
14 alphaF <- .025        # type I error for the full population
15 nullReps <- 1000      # number of simulations to determine critical value
16 altReps <- 1000       # number of reps to simulate under alternative
17
18 #### This function will take the inputs and return any rejection rates and the correct group
       rates
19 main(n0, pVec, sce, subEffect, allEffect, baseline, nleaves, alpha, alphaF, nullReps, altReps)
```

./code/AppC_3.R

# Appendix D

# SAMPLE R CODE FOR SHAPES

Code this appendix runs a SHAPES simulation study. First run the code in Appendix B, which will generate simulated trial data. Then run the code in each section in order. The last section contains the code with the simulation parameters, with the details in comments, and the line that will run a simulation. Note that *nullReps* is used to determine the number of null trials run to determine the critical value and *altReps* the number of trials under the alternative. The rate of any rejection and correct identification are returned.

## *D.1 Functions to Perform SHAPES*

```
#### this function considers k total variables and up to \ell possible combinations thereof
#### to return the total unique SHAPES considered; generally \ell <=3 and k >= ell
comboCalc <- function(k, ell){
  tot <- 0
  for(i in 0:ell){tot <- tot + 2^i * choose(k, i)}
  return(tot)
}

# this is just a simple function to convert the truth table from a simple two vector form
# to the truth vector
convert <- function(varN, value, truthtable){
       ct <- length(varN)

       restr <- as.matrix(truthtable[, varN], nrows=nrow(truthtable), ncols= ct)
       true <- rep(1, nrow(truthtable))
       for(i in 1:ct){true <- true * (1*(restr[, i]==value[i]))}

      return(true)
}

scenario <- function(scene, k){
    krec <- k
    if(k<3){k <- 3}
    varCt <- 0
  #### create a table with all possible combinations
  truthSub1 <- list()
  for(i in 1:k){truthSub1[[i]] <- c(1,0)}
  truthSub2 <- expand.grid(truthSub1)
  truthtable <- truthSub2[k:1]
```

```
  colnames(truthtable) <- paste('X', 1:k, sep='')

  ######## scenario A ########################################
  if(scene==1){
    lpredF <- truthtable[,'X1']
    }
  ########## scenario B #######################################
  if(scene==2){
    lpredF <- pmax(truthtable[,'X1'], truthtable[,'X2'])
    }
  ########## scenario C #######################################
  if(scene==3){
    ### remember that pmax is saying or and pmin is saying and
    lpredF <- pmin(truthtable[,'X1'], truthtable[,'X2'])
    }
  ###################scenario Z ################################
  if(scene==26){
    lpredF <- pmax(pmin(truthtable[,'X1'], truthtable[,'X2']),
        pmin(truthtable[,'X3'], truthtable[,'X4']))
     }

  return(lpredF)
}

# THIS INPUTS ARE THE DATASET, A VARIABLE FOR THE DEPTH OF 1, 2, 3,
# AN INDICTATOR OF THE OUTCOME
# AN OVERALL ALPHA LEVEL (THAT AUTOMATICALLY GETS SPLIT INTO THE VARIOUS DEPTHS
# AND THE TRUTH TABLE (FROM THE SCENARIO FUNCTION)
# THE OUTPUT IS A Table with all the various subgroups and pvalues
shapesR <- function(trialList, ell, alphaTot, shapesP=NA){
  covar <-trialList[[1]]
  Trt <- trialList[[2]]
  y <- trialList[[3]]
  k <- ncol(covar)
  full <- cbind(covar, txt=Trt, y)

    # here the alpha levels are split evenly between the 'depths'
    # with the 0 depth accounting for the full population
    # i.e. the first entry in the vector
    n <- nrow(full)
    if(ell==3){alphaVec <- c(alphaTot/4, alphaTot/4, alphaTot/4, alphaTot/4)}
  if(ell==2){alphaVec <- c(alphaTot/3, alphaTot/3, alphaTot/3)}
    if(ell==1){alphaVec <- c(alphaTot/2, alphaTot/2)}


  #### create a table with all possible combinations
  truthSub1 <- list()
  for(i in 1:k){truthSub1[[i]] <- c(1,0)}
  truthSub2 <- expand.grid(truthSub1)
  truthtable <- truthSub2[k:1]
```

```
80  colnames(truthtable) <- paste('X', 1:k, sep='')
81  testCombo <- comboCalc(k, ell)
82
83
84  # set up the table to hold all of the zstatistics and subgroups analyzed
85  raw <- matrix(NA, nrow=testCombo-1, ncol=2*ell +12)
86  colnames(raw) <- c( paste('v', 1:ell, sep=''), paste('i', 1:ell, sep=''), 'zActual', 'zComp', '
        zInter', 'nActual', 'nComp', 'depth', 'crit', 'test1', 'test2', 'critB', 'testB1', 'testB2
        ')
87
88  index <- 1
89  for(ii in 1:ell){
90       possible <- t(combn(1:k, ii))
91
92       raw[index: (((2^ii) * nrow(possible)) + index - 1), 1:ii ] <-  matrix(rep(t(possible), 2^
             ii),byrow=T, ncol=ii)
93       raw[index: (((2^ii) * nrow(possible)) + index - 1), 'depth' ] <- ii # let's catalog the
             depth too 2*ell +6
94
95      if(ii==1){raw[index: (((2^ii) * nrow(possible)) + index - 1), (ell+1) ] <- c(rep(1, k), rep
            (0,k))}
96      if(ii==2){raw[index: (((2^ii) * nrow(possible)) + index - 1), (ell+1):(ell+2) ] <- cbind(c(
            rep(1, 2*nrow(possible)), rep(0, 2* nrow(possible))) , rep(c(rep(1, nrow(possible)),
            rep(0, nrow(possible))), 2))}
97
98      if(ii==3){raw[index: (((2^ii) * nrow(possible)) + index - 1), (ell+1):(ell+3) ] <-
99          cbind(  c(rep(1, 4*nrow(possible)), rep(0, 4*nrow(possible))), rep( c(rep(1, 2*nrow(
                possible) ), rep(0,2*nrow(possible) )), 2) ,  rep( c(rep(1, nrow(possible)), rep
                (0, nrow(possible))) , 4))}
100
101      index <- index+ ((2^ii) * nrow(possible))
102
103 }
104 if(k>1){raw <- raw[-((k+1):(2*k)),]} # here I am removing rows that correspond to the repeats
        of the single variables ,
105                             # where those groups with depth >1 have meaning
106
107 pen1 <- 2 * table(raw[,'depth'])
108 pen2 <- alphaVec[2:(ell+1)]/pen1 # this is then the appropriate p-value
109 critical <- c(alphaVec[1], pen2)
110
111 if(ell==3){
112 # let's fill in the critical values now
113 raw[, 'crit'] <- (pen2[1] * (raw[, 'depth']==1)) + (pen2[2] * (raw[, 'depth']==2)) + (pen2[3] *
         (raw[, 'depth']==3))}
114
115 if(ell==2){
116 # let's fill in the critical values now
117 raw[, 'crit'] <- (pen2[1] * (raw[, 'depth']==1)) + (pen2[2] * (raw[, 'depth']==2)) }
118
```

```
119  if(ell==1){
120  # let's fill in the critical values now
121  raw[, 'crit'] <- (pen2[1] * (raw[, 'depth']==1))}
122
123
124  for(iii in 1:nrow(raw)){
125    ck <- sum(is.na(raw[iii, 1:ell]))
126    if(ck==0 & ell==3){
127      d1 <- subset(full, full[, raw[iii, 1]]== raw[iii, ell+1] & full[, raw[iii, 2]]== raw[iii,
           ell+2] & full[, raw[iii, 3]]== raw[iii, ell+3])
128      d2 <- subset(full, !(full[, raw[iii, 1]]== raw[iii, ell+1] & full[, raw[iii, 2]]== raw[iii,
            ell+2] & full[, raw[iii, 3]]== raw[iii, ell+3]))
129    }else if((ck==1 & ell==3) | (ck==0 & ell==2)){
130      d1 <- subset(full, full[, raw[iii, 1]]== raw[iii, ell+1] & full[, raw[iii, 2]]== raw[iii,
           ell+2])
131      d2 <- subset(full, !(full[, raw[iii, 1]]== raw[iii, ell+1] & full[, raw[iii, 2]]== raw[iii,
            ell+2]))
132    }else{
133      d1 <- subset(full, full[, raw[iii, 1]]== raw[iii, ell+1] )
134      d2 <- subset(full, !(full[, raw[iii, 1]]== raw[iii, ell+1])) }
135
136    raw[iii, 'nActual'] <- nrow(d1)
137    raw[iii, 'nComp'] <- nrow(d2)
138
139
140    if(nrow(d1)>0 & sum(d1[,'txt'])!=0 & sum(d1[,'txt'])!= nrow(d1)){m1 <- summary(glm(d1[, 'y']~
           d1[,'txt'], family='binomial'))$coef; raw[iii, 2*ell+1] <- 1-pnorm(m1[2,1]/m1[2, 2])}
141    if(nrow(d2)>0 & sum(d2[,'txt'])!=0 & sum(d2[,'txt'])!= nrow(d2)){m2 <- summary(glm(d2[, 'y']~
           d2[,'txt'], family='binomial'))$coef; raw[iii, 2*ell+2] <- 1-pnorm(m2[2,1]/m2[2, 2])}
142  }
143
144  raw[,'test1'] <- raw[, 'zActual']/raw[,'crit']
145  raw[,'test2'] <- raw[, 'zComp']/raw[,'crit']
146
147    teeter <- raw
148  mF <- summary(glm(full[, 'y']~full[,'txt'], family='binomial'))$coef;  fullP <- 1-pnorm(mF[2,1]
       /mF[2, 2])
149
150   # here i am setting up a vector with minimum pvalue for the full population,
151   # the first subgroup (single var), the second etc.
152  pvalVec <- rep(NA, ell+1)
153  pvalVec[1] <- fullP
154  for(i in 1:max(raw[,'depth'])){
155        co1 <- subset(raw, raw[,'depth']==i)
156        pvalVec[(i+1)] <- min(c(co1[,'zActual'], co1[,'zComp']), na.rm=T)
157  }
158
159  min1 <- min(c(teeter[, 'test1'], teeter[, 'test2']), na.rm=T) # find the minimum p-value
160  ind1 <- which(teeter[, c('test1', 'test2')]==min1, arr.ind=T)
161
```

```
  standP <- fullP/alphaVec[1]

  # this is to generate alternative decicions

  if(!is.na(shapesP[1])){

            for(iiii in 2:(ell+1)){
              raw[, 'critB'] <- ifelse(raw[,'depth']!= (iiii-1), raw[, 'critB'], shapesP[iiii])
            }

        raw[, 'testB1'] <- raw[, 'zActual']/raw[,'critB']
        raw[, 'testB2'] <- raw[, 'zComp']/raw[,'critB']

        teeter <- raw
             min1 <- min(c(teeter[, 'testB1'], teeter[, 'testB2']), na.rm=T) # find the minimum p
                  -value
            ind1 <- which(teeter[, c('testB1', 'testB2')]==min1, arr.ind=T)


            standP <- fullP/shapesP[1]


              if(standP < 1){
                selected <-  c(rep(NA, ell*2), fullP, rep(NA,2), n, NA, 0, alphaVec[1], NA, NA,
                     shapesP[1], standP, NA)
              }else{selected <- teeter[ind1[1,1],]}


        names(selected) <- c( paste('v', 1:ell, sep=''), paste('i', 1:ell, sep=''), 'zActual', '
             zComp', 'zInter', 'nActual', 'nComp', 'depth',                'crit', 'test1', '
             test2', 'critB', 'testB1', 'testB2')


                testa <- ifelse(selected['testB1']<1 | selected['testB2']<1 ,1, 0 )
                testa <- ifelse(is.na(testa) & selected['testB1'] <1, 1, ifelse(is.na(testa) &
                     selected['testB1']>=1, 0, testa) )

            depth <- selected['depth']

          side <- ifelse(!is.na(selected['testB1']) & !is.na(selected['testB2']) & selected['
               testB1'] < selected['testB2'], 1 ,
              ifelse(!is.na(selected['testB1']) & !is.na(selected['testB2']) & selected['testB1']
                   >= selected['testB2'],2,
              ifelse(!is.na(selected['testB1']), 1, 2)))

            # here we cycle throught the possible depths to convert form the integer numbers and
                 values
            # to the actual estimated truth vector
        if(testa==0){dec2 <- rep(0, nrow(truthtable))
```

```
203         }else if(depth==0){dec2 <- rep(1, nrow(truthtable)) # some vector of length nrow(
                truthable)
204         }else if(depth==1 & side==1){ dec2 <- convert(selected['v1'], selected['i1'], truthtable)
205         }else if(depth==1 & side==2){ dec2 <- 1-convert(selected['v1'], selected['i1'],
                truthtable)
206         }else if(depth==2 & side==1){ dec2 <- convert(selected[c('v1', 'v2')], selected[c('i1', '
                i2')], truthtable)
207         }else if(depth==2 & side==2){ dec2 <- 1-convert(selected[c('v1', 'v2')], selected[c('i1',
                 'i2')], truthtable)
208         }else if(depth==3 & side==1){ dec2 <- convert(selected[c('v1', 'v2', 'v3')], selected[c('
                i1', 'i2', 'i3')], truthtable)
209         }else if(depth==3 & side==2){ dec2 <- 1-convert(selected[c('v1', 'v2', 'v3')], selected[c
                ('i1', 'i2', 'i3')], truthtable)
210           }else{dec2 <- NA}
211
212
213
214
215         return(dec2)
216    }else{
217       return(pvalVec)
218    }
219 }
```

./code/AppD_1.R

### *D.2 Functions for a Simulation Study*

```
 1 library(LogicReg)
 2 library(lmtest)
 3
 4 denom <- function(int){
 5   int2 <- 2^int
 6   ct <- 0
 7   for(q in 0:int2){
 8     ct <- ct + choose(int2, q)
 9   }
10   return(ct)
11 }
12
13 #### a simple function to calculate various stats of interest
14 # including overall error rate, the  predictive error    , PPV, whether trial has 0 error rate,
15 #     whether trial has 0 predictive error, whether trial has PPV above some threshold
16 trialStat <- function(decisionVec, lpredV, prob, effAll, effSub){
17      s1a <- sum(prob * (lpredV-decisionVec)^2)
18      m1a <- sum(prob * ((lpredV-decisionVec)==1))/sum(prob*lpredV)
19
20
21      ppv1a <- sum(prob * decisionVec * ((lpredV-decisionVec)==0))/sum(prob*decisionVec)
```

```
22      #mse1a <- sum(prob *  ((effTxt + effPred*lpredV) - (p1[1]))^2)
23       powA1a <- ifelse(s1a==0, 1, 0)
24       powC1a <- ifelse(m1a==0, 1, 0)
25       ppv1aThresh <- ifelse(ppv1a >.80, 1, 0)
26       dec1 <- ifelse(sum(decisionVec)==length(lpredV), 1, ifelse(sum(decisionVec)==0, 0 , 2))
27
28       trueEffectVec <-   ((effAll*decisionVec)  +  (effSub * (decisionVec*lpredV) ))
29         wts <- (decisionVec*prob)  / (decisionVec %*% prob)
30         trueEffectSelect <- sum(wts * trueEffectVec)
31
32       c(s1a, m1a, ppv1a, powA1a, powC1a, ppv1aThresh,  dec1==1, dec1==2, trueEffectSelect)
33  }
34
35  ### this is a function to calculate the proportion in each group
36  ### in the presence of correlated predictors
37  probCalc <- function(tt, corrS, mean1){
38    prob <- rep(NA, nrow(tt))
39    print(corrS); print('hi'); print(mean1)
40
41    for(y in 1:length(prob)){
42      vector1 <- rep(NA, ncol(tt))
43      vector2 <- rep(NA, ncol(tt))
44      info <- tt[y,]
45      for(yy in 1:length(info)){
46        if(info[yy]==1){vector1[yy]<- 0; vector2[yy]<- Inf}
47        else{vector1[yy] <- -Inf; vector2[yy] <- 0}
48      }
49      prob[y] <- pmvnorm(lower=vector1, upper=vector2, mean=2*(mean1-.5), sigma=corrS)
50    }
51    return(prob)
52  }
53
54
55  # this will be a function that takes a matrix of dimenions nullreps by ell+1
56  # and a cumulative vector of p-values of length ell+1 which increases up to alphaTOt
57  # it will return a vector of cut-off pvalues
58  shapesCO <- function(shapesP, prVec){
59          cycles <- length(prVec)
60          reps <- nrow(shapesP)
61          cutoffSH <- rep(NA, cycles)
62          cutoffSH[1] <- prVec[1]
63
64          leftOver <- reps - floor(quantile(1:reps, prVec))
65          remaining <- subset(shapesP, shapesP[,1]>=prVec[1])
66
67          for(i in 2:cycles){
68                ranker <- rank(1-remaining[,i])
69                subTab <- remaining[,i][ranker > leftOver[i]]
70                cutoffSH[i] <- max(subTab)
71                remaining <- subset(remaining, remaining[,i]>cutoffSH[i])
```

```
72          }
73          return(cutoffSH)
74  }
75
76  # n is the total sample size,
77  # p is a vector of length k that includes proportions of covariates
78  # scene is the specific scenario and should be a ltter
79  # random just allows for random sampling of the covariates, otherwise fixed attempts to ensure
        appropriate combinations
80  # correlate is whether the covariates are correlated
81  # force... ensures at least all possible combinations of covariates have at least one observation
        in each txt/control group
82  # sigma is a matrix of covariance of predictors, should be k by k
83  # effPred is the predictive effect, #effProg is the prognostic effect and effTxt is the overall
        treatment effect
84  # variance is the variance of the outcome, not needed in the case of binary
85  # binary indicates outcomes is binary
86  # selectLR is the logic regression fitting algorithm
87  # leaves is the number of leaves allowed in the model
88  # scene is the scenario under which the data is generated
89  # subVec is the subset of the covariates that are selected
90  # nullReps is the number of repetitions to generate null cut-offs
91  # altReps is the number of simulations
92  # alphaTot is the total alpha level for all comparisons
93  # alphaSub is the the alpha to spend on subgroup comparisons, note alphaSub <= alphaTot
94  # controlMeth is the method that is used to control for prognostic factors for 4 out of the 8
        methods, including whether prognostic factors are
95  #     exactly known
96  # ell is the depth for shapes
97
98  ####### should be able to use this 'do.one' function for both null and alternative data
        generation -
99  #######   remember to keep general enough for this task
100 do.one <- function(n, p, scene, effPred, effTxt, bl, nullReps=1000, altReps=1000, alphaTot=.1,
        alphaF=.05, ell, shapes_pTab=NA){
101   # number of covariates that we will consider
102   k <- length(p)
103
104   trialList <- genOne(n, p, bl, effTxt, effPred, scene)
105
106     if(is.na(shapes_pTab[1])){p10 <- shapesR(trialList, ell, alphaTot, NA)
107     }else{p10 <- shapesR(trialList, ell, alphaTot, shapes_pTab)
108
109
110     }
111     return(p10)
112 }
113
114
115
```

```
116
117
118 main <- function(n, p, scene, effPred, effTxt, bl, nullReps=1000, altReps=1000, alphaTot=.1,
        alphaF=.05, ell){
119
120   # number of covariates that we will consider
121   k <- length(p)
122   nullD <- replicate(nullReps, do.one(n, p, scene,
123             0, 0, bl,
124             nullReps, altReps, alphaTot, alphaF, ell), simplify = "array")
125
126     # this is basically code to generate the cut-offs for the various methods (null distribution
            assumptions)
127     alphaVec <- rep(NA, ell+1)
128     alphaVec[1] <- (alphaF)
129     alphaVec[2:(ell+1)] <- rep( (alpha- (alphaF))/ell, ell)
130     alphaVec <- cumsum(alphaVec)
131
132    pSH <- shapesCO(t(nullD), alphaVec)
133
134
135   # Now generate data under the specified alternative
136
137
138    res1  <- t(replicate(altReps, do.one(n, p, scene,
139             effPred, effTxt, bl,
140             nullReps, altReps, alphaTot, alphaF, ell,  pSH), simplify = "array"))
141
142    truthV <- rep(0, 2^k)
143    if(effPred > 0){truthV <- scenario(scene, k)}
144    if(effTxt > 0){truthV <- rep(1, length(truthV))}
145
146     vec1 <- rep(NA, altReps)
147     for(i in 1:nrow(res1)){
148       vec1[i] <- (sum( (res1[i,]-truthV)^2) ==0)
149     }
150
151     return( c(ANY=sum(apply(res1, 1, sum)>1)/altReps,  CORRECT=sum(vec1)/altReps))
152 }
```

./code/AppD_2.R

### *D.3 Function to Run a Simulation Study with Parameter Value Specification*

```
1 set.seed(1)  ### set this to reproduce results
2
3
4 n0 <- 200           # total sample size, >=200
5 pVec <- c(.5, .5, .5, .5)   #vector of probabilities for covariates (independent)
```

```
sce <- 1           # the scenario, A, B, C, Z = 1, 2, 3, 26
              # and the length of pVec should be at least 1, 2, 2, 4
              #   for each scenario
subEffect <- .3       # the subgroup treatment effect
allEffect <- 0        # the main treatment effect
baseline <- .3        # baseline response in control should be in (0, 1)
nullReps <- 1000      # number of simulations to determine critical value
altReps <- 1000       # number of reps to simulate under alternative
alpha <- .1         # type I error for all
alphaF <- .025        # type I error for the full population
ell <- 1            # the depth of the subgroup allowed

#### This function will take the inputs and return any rejection rates and the correct group
    rates
main(n0, pVec, sce, subEffect, allEffect, baseline, nullReps, altReps,  alpha, alphaF, ell)
```

./code/AppD_3.R