PARTITION TESTING FOR BROAD EFFICACY AND IN GENETIC SUBGROUPS

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Szu-Yu Tang, B.B.A., M.B.A.

Graduate Program in Statistics

The Ohio State University

2012

Dissertation Committee:
Professor Jason C. Hsu, Advisor
Professor Haikady N. Nagaraja
Professor Eloise E. Kaizar
Abstract

Testing a drug compound for efficacy increasingly involves simultaneously testing for efficacy in the entire patient population (broad efficacy), and efficacy in one or more pre-specified subgroups. Two scenarios leading to such testing are as follows. The first is that the mechanism of action of the compound might make it more beneficial to a subgroup of the patients. This is often true with cancer drugs, with recent examples such as crizotinib (Xalkori) for non-small cell lung cancer patients with ALK translocation, and vemurafenib (Zelboraf) for skin cancer patients with BRAF mutation. The second scenario is efficacy of a compound may be affected by polymorphism in genes in its metabolic and transport pathway. Hotly debated is whether polymorphism in CYP2C19 affects efficacy of clopidogrel (Plavix) and tamoxifen, for example.

When a drug is approved for use by the entire patient population, there may be concern that this efficacy is driven by extreme efficacy in a subgroup only. If a drug is approved for use in a subgroup, due to the practice of off-label use, some assessment of its efficacy for patients in the complementary subgroup is also desirable.

Our research shows Partition testing for broad efficacy and efficacy in pre-specified subgroups readily recognizes logical relationships among the parameters in the null hypotheses being tested, and readily inverts to simultaneous confidence bounds on
efficacy in broad population and in the subgroups. If these confidence bounds show efficacy in a subgroup is of real concern, then an appropriate statement can be included on the label.
Dedicated to my family and my Columbus Tzu Chi family
Acknowledgments

This dissertation is completed with support and help from a great number of individuals. First, I would like to express the deepest appreciation to my dissertation advisor, Professor Jason C. Hsu, for his constant support and encouragement. He has been a mentor, friend, and even a father for me. His guidance has made this a thoughtful and rewarding journey. His impact will constantly guide me in the right direction for my future career.

I would like to thank my committee members, Professor Haikady N. Nagaraja and Professor Eloise E. Kaizar, who provided valuable statistical advice. I especially appreciate Professor Kaizar’s help with the analysis for discrete data, clarification of statistical concepts, and revision of my dissertation. In addition, I would like to thanks Professor Jerry M. Bigham for his helpful editorial remarks.

A special thank for the financial support from the Ministry of Education in Taiwan and Department of Statistics at the OSU.

I would like to thank Hsun-Hsuan Huang and Yushi Liu, who as good friends, were always willing to help, accompany and give their best suggestions. Finally, I would like to thank my mother for her full support and her endless love.
Vita

December 11, 1975 .......................... Born - Taipei, Taiwan

1998 ................................. B.S. in Statistics,
   National Chengchi University (NCCU),
   Taipei, Taiwan

2000 ................................. M.B.A. in Finance,
   National Taipei University (NTPU),
   Taipei, Taiwan

2010-present .............................. Graduate Teaching Associate,
   The Ohio State University

Publications

Research Publications

Michelle L. Aspa, Angela L. Collene, Leigh E. Norris, Rachel M. Cole, Michael B.
Stout, Szu-Yu Tang, Jason C. Hsu and Martha A. Belury “Time-dependent effects
of safflower oil to improve glycemia, inflammation and blood lipids in obese, post-
menopausal women with type 2 diabetes: A randomized, double-masked, crossover

Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
</tbody>
</table>

## 1. Decision-theoretic Partition testing to control error rates

1.1 Different definitions of error rates                               | 1   |

1.2 Decision theoretic approach to multiple testing in Pharmaceuticals  | 4   |

1.2.1 Basic elements of statistical decision theory                  | 4   |

1.2.2 Decision theory for testing two null hypotheses                 | 6   |

1.3 Error rates in accelerated approval                               | 15  |

1.4 Partition testing principle for controlling error rates           | 17  |

1.4.1 Partition vs. Closed testing for controlling FWER              | 19  |

1.4.2 iPartition for controlling expected loss                        | 22  |

## 2. New and old formulations of genetic subgroups analysis

2.1 Potential of personalized medicine                                | 28  |

2.2 Motivating examples                                               | 29  |

2.3 Previous proposals                                                | 31  |

2.3.1 Previous formulations                                          | 31  |

2.3.2 Definition of parameters and setting of hypotheses             | 32  |

2.3.3 Methodology of genetic subgroup analysis                       | 33  |
3. Assessing broad efficacy and efficacy in two genetic subgroups ............ 61

3.1 Simultaneous testing for broad efficacy and efficacy in one pre-determined subgroup ................................................. 61
   3.1.1 Formulating null hypotheses ......................................................... 61
   3.1.2 Decision rule: Partition testing vs. Closed testing ............. 62
   3.1.3 Inference: consistency issue and off-label use .................. 64

3.2 Simultaneous testing for broad efficacy and efficacy in complementary subgroups .................................................. 64
   3.2.1 Formulating null hypotheses ......................................................... 65
   3.2.2 Decision rule: Partition testing vs. Closed testing ............. 68
   3.2.3 A logic and a distribution example ............................. 68
   3.2.4 Analysis of data from a motivating example .................. 72
   3.2.5 A simulation comparison of performance ............ 76

3.3 Simultaneous confidence bounds for broad efficacy and efficacy in complementary subgroups ................................................. 83
   3.3.1 Confidence set for assessing consistency in efficacy ........ 83
   3.3.2 Theorem of constructing a confidence set .................... 87
   3.3.3 Analysis of two artificial examples ..................... 90
   3.3.4 Analysis of data from a motivating example ................ 91

4. Simultaneous testing for broad efficacy and efficacy in three or more subgroups .................................................. 95

4.1 Simultaneous testing for broad efficacy and efficacy in three or more subgroups .................................................. 95

4.2 Testing efficacy in nested genetic subgroups ............ 98

5. Conclusion ................................................................. 100

5.1 Concluding remarks ............................................................. 100
5.2 Limitation ................................................................. 101
5.3 Future research ............................................................. 102
Appendices

A. maxT and minP tests ............................................ 104
B. Hochberg’s method ............................................... 105
C. Partitioning decision paths .................................. 107
D. Correlation for three subgroups ............................ 110

Bibliography ......................................................... 111
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Number of hypotheses in each category</td>
</tr>
<tr>
<td>1.2</td>
<td>Loss for testing one single null hypothesis</td>
</tr>
<tr>
<td>1.3</td>
<td>Loss Table for testing two null hypothesis simultanously. $L_i^I$ and $L_i^{II}$ indicate loss from Type I and Type II error for $i=1,2$.</td>
</tr>
<tr>
<td>1.4</td>
<td>$L(\theta_1, a_1; \theta_2, a_2) = I{V &gt; 0}; {L(\theta_1, a_1; \theta_2, a_2) = V} / V/R$</td>
</tr>
<tr>
<td>1.5</td>
<td>$R(\theta_i, \delta_i)$ for different definitions of loss assuming independent between two test statistics</td>
</tr>
<tr>
<td>1.6</td>
<td>Possible losses from policy makers (from clinicians, health economists and other stake-holders). T indicates True null hypothesis and F indicates false null hypothesis</td>
</tr>
<tr>
<td>2.1</td>
<td>Two-Stage Testing Procedure in Song and Chi (2007)</td>
</tr>
<tr>
<td>2.2</td>
<td>Flexible Testing Strategy in Alish and Huque (2009)</td>
</tr>
<tr>
<td>2.3</td>
<td>Flexible Testing Strategy in Alish and Huque (2009) illustrated by closed testing principle</td>
</tr>
<tr>
<td>2.4</td>
<td>A toy example showing difference between $\bar{\theta}<em>n$ and $\bar{\theta}</em>\gamma$</td>
</tr>
<tr>
<td>2.5</td>
<td>responders (R) and Non-responders (NR) for each genetic subgroup in the treatment and control arms and the combined group</td>
</tr>
<tr>
<td>2.6</td>
<td>Contingency tables for time to event data. D means patients died and ND means patients survived</td>
</tr>
</tbody>
</table>
3.1 Closed testing (C) and Partition testing (P) decision rules for rejecting $H_0$ and $H_{0+}$. A $\lor$ denotes a C or P null hypothesis that needs to be rejected.

3.2 Closed testing (C) and Partition testing (P) decision rules for rejecting $H_0$, $H_{0+}$, and $H_{0-}$. A $\lor$ denotes a C or P null hypothesis that needs to be rejected. An n/a denotes a P null hypothesis that need not be tested.

3.3 Analysis of Logic and Distribution examples, FWER=0.05. A Partition null hypothesis is rejected by H or P if there is at least one $\times$.

3.4 Analysis of Logic and Distribution examples (FWER=0.05), presented in terms of p-values. A Partition null hypothesis is rejected by H or P if there is at least one $\times$.

3.5 Number of incidence in clopidogrel (Plavix) example. (Number in parentheses is total number of patients in that group.)

3.6 Analysis of clopidogrel (Plavix) data, $\delta = 0$, FWER = 0.025. A Partition null hypothesis is rejected by H or P if there is at least one $\times$.

3.7 Analysis of clopidogrel (Plavix) data, $\delta = 0.016$, FWER = 0.05. A Partition null hypothesis is rejected by H or P if there is at least one $\times$.

3.8 null hypotheses constructed from closed testing and partition testing principle.

3.9 Confidence set is computed by first pivoting each acceptance region within its partitioning null hypothesis space to obtain a corresponding confidence region, and then taking the union of the confidence regions.

3.10 95% confidence bounds calculation for distribution and logic examples.

3.11 95% confidence bounds calculation for Plavix example, $\delta = 0$ and $\delta = 0.016$.

C.1 Partition testing (P) decision rules for rejecting $H_0$, $H_{0+}$ or $H_{0+}$. A $\lor$ denotes a null hypothesis that needs to be rejected.
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Risk function for testing one hypothesis</td>
<td>13</td>
</tr>
<tr>
<td>2.1 I: Alpha-split method, II: Two stage testing procedure (IIa: quantitative consistency, IIb: qualitative consistency), III: Flexible testing strategy</td>
<td>41</td>
</tr>
<tr>
<td>2.2 Row-uniform Design I</td>
<td>54</td>
</tr>
<tr>
<td>2.3 Row-uniform Design II</td>
<td>55</td>
</tr>
<tr>
<td>2.4 Row-uniform Design III</td>
<td>55</td>
</tr>
<tr>
<td>2.5 Row-column-uniform Design I</td>
<td>60</td>
</tr>
<tr>
<td>3.1 Correct and useful inference and its corresponding parameter region in the order $\bar{H}<em>0$, $H</em>{0+}$, $H_{0-}$. T or F means true or false null hypothesis for the parameter region. R or A means rejecting or accepting null hypothesis in its corresponding parameter region. The $g^+$ and $g^-$ values represent $\theta_{g^+}$ and $\theta_{g^-}$</td>
<td>78</td>
</tr>
<tr>
<td>3.2 $P{\text{correct and useful inference}}$ of Partition testing, and its difference with Hochberg’s method. Positive $g^+$ and $g^-$ value represent efficacy in the $g^+$ and $g^-$ subgroup respectively, while broad efficacy is represented by $(g^+, g^-)$ values to the upper-right of the $-45^\circ$ line.</td>
<td>79</td>
</tr>
<tr>
<td>3.3 $P{\text{correct and useful inference}}$ of Partition testing with maxT test, and its difference with closed testing with maxT test. Positive $g^+$ and $g^-$ value represent efficacy in the $g^+$ and $g^-$ subgroup respectively, while broad efficacy is represented by $(g^+, g^-)$ values to the upper-right of the $-45^\circ$ line.</td>
<td>84</td>
</tr>
</tbody>
</table>
3.4 Examples of confidence set. In each graph, the red dot is the point estimate of \((\theta_{g+}, \theta_{g-})\), and the dotted lines are the deduced confidence bounds on \(\theta_{g+}, \theta_{g-}, \hat{\theta}\).

3.5 Confidence Set for logic example and distribution example
Chapter 1: Decision-theoretic Partition testing to control error rates

Familwise Error rate (FWER) control is commonly used in drug efficacy clinical studies. However, it may not be appropriate for all scenarios. In this chapter, we first review several commonly used error rates. Then we use decision theory to connect multiple testing and control of error rates. Accelerated approval of oncology drugs will be used as an example to show none of these error rates is appropriate. At the end of this chapter, we show that the $iPartition$ principle can control different types of error rates in multiple testing.

1.1 Different definitions of error rates

In this section, we review several commonly used error rates: FWER, gFWER, E(V), FDR, and Fdr and connect them with three error rates listed in Tukey (1953).

Assume $k$ hypotheses are tested of which $k_0$ are true. Let $V$ indicate the number of incorrectly rejected true null hypotheses, and let $R$ indicate the total number of rejected null hypotheses. Notations are as listed in Table 1.1.
<table>
<thead>
<tr>
<th>True null hypotheses</th>
<th>False null hypotheses</th>
<th>All hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepted</td>
<td>Rejected</td>
<td>total</td>
</tr>
<tr>
<td>$U$</td>
<td>$V$</td>
<td>$k_0$</td>
</tr>
<tr>
<td>$T$</td>
<td>$S$</td>
<td>$k_1$</td>
</tr>
<tr>
<td>$W$</td>
<td>$R$</td>
<td>$k$</td>
</tr>
</tbody>
</table>

Table 1.1: Number of hypotheses in each category

- **FWER** (Family Wise Error Rate): FWER is the probability of at least one incorrect rejection.

  \[
  FWER = P(V > 0) = E[I(V > 0)]
  \]

- **gFWER** (generalized Family Wise Error Rate): gFWER is the probability of making more than $m$ ($> 0$) incorrect rejections.

  \[
  gFWER = P(V > m) = E[I(V > m)]
  \]

- **E(V)**: E(V) is the expected number of incorrect rejections.

  \[
  E(V) = E[\sum_i I(\text{Reject the true null hypotheses } H_{0i})]
  \]

- **FDR** (False Discovery Rate): FDR is the expected ratio of incorrect rejections among all rejections. It is an unconditional expectation of the random variable $V/R$.

  \[
  FDR = E\left[\frac{V}{R}\right]
  \]

  where $\frac{V}{R}$ is defined to be 0 when $R = 0$. 
• **Fdr**: Fdr is a conditional version of FDR. However, it is different from the conditional expectation of \( V/R \).

\[
Fdr = \frac{E(V)}{r} \neq \frac{E[V|R = r]}{r} = \frac{E[V|R = r]}{r}
\]

In the beginning, there was Tukey (1953). Page 4 of this document lists the following three error rates.

- error rate per comparison = \( \frac{\text{erroneous statements}}{\text{comparisons}} \)
- error rate per family = \( \frac{\text{erroneous statements}}{\text{families}} = E(V) \)
- error rate familywise = \( \frac{\text{erroneous families}}{\text{families}} = \text{FWER} \)

When any family-oriented error rate, or its "batch" version, is discussed in the 407 pages of notes of Tukey (1953), all three error rates pretty much always appear together. So it was not the case that Tukey thought error rate familywise, now popularly denoted as FWER (for familywise error rate), is especially more meaningful to control than any other error rate.

Tukey’s error rate per family is \( E(V) \), the average number of incorrect rejections, averaged over infinitely many (different) studies. As Tukey explains, in special circumstances, error rate per comparison and error rate per family reduce to error rate familywise. Specifically, when all the statements are completely positively dependent, so that the statements are all correct or incorrect simultaneously, then FWER is the same as error rate per comparison. On the other hand, if correctness of the statements are disjoint events, so that if one of the statements is incorrect than none of the other statements is incorrect ("strong negative dependence" in Tukey’s wording), then FWER is the same as error rate per family.
One can argue that in fact, FWER loses some details of the errors committed, because it does not track which or how many incorrect statements are made in a family. From the FWER point of view, one mistake counts the same as two or more mistakes since in both situation one has made "at least one incorrect rejection." An example for which FWER control is appropriate is in dose response studies with multiple endpoints. Here, FWER control guards against any incorrect inference on the efficacy of any endpoint at any dose level. However, with "complex clinical trials", we need error rates more detailed than FWER. We will discuss which error rates should be considered as appropriate in accelerated approval in section 1.3

1.2 Decision theoretic approach to multiple testing in Pharmaceuticals

Decision Theory formalized the calculation of risk from incorrect decisions. It is not necessarily applied to only Bayesian analysis. In this section, we connect the concept of decision theory with multiple testing in pharmaceutical analysis.

1.2.1 Basic elements of statistical decision theory

According to James O. Berger (1985) [2], decision theory is concerned with making a decision in the presence of uncertainty using all available information. The uncertainties can be considered as unknown quantities, denoted by $\theta$. The useful information can combine sample information with other aspects including prior information and knowledge of the consequence of any decision (the loss function). The basic elements defined in [2] are as follows.

- state of nature $\theta$: $\theta$ is the unknown quantity which affects the decision process.
  And $\Theta$ is the set of all possible states of nature.
• action $a$: An action $a$ is a decision. And $A$ is the set of all possible actions.

• loss function $L(\theta, a)$: A loss function is a key element of decision theory. A loss $L(\theta_1, a_1)$ will occur when taking action $a_1$ when $\theta_1$ is true state of nature. Loss is defined for all $(\theta, a) \in \Theta \times A$, all combinations of different actions and true states of nature, independent of sample data.

• random variable $\mathbf{X}$: A random variable $\mathbf{X} = (X_1, X_2, ..., X_n)$ is investigated to collect information about $\theta$. $X_i$ are variables from a common distribution $P_\theta(X \in A)$. The observation of $\mathbf{X}$ will be $\mathbf{x}$ and the set of possible outcomes is the sample space $\chi \in \mathbb{R}^n$.

• (nonrandomized) decision rule $\delta(x)$: $\delta(x) \in A$ is a function mapping sample information to actions ($\chi \to A$). If $X = x$ is observed, then $\delta(x) = a$ is the action that will be taken.

• risk function $R(\theta, \delta)$: $R(\theta, \delta)$ is the expected loss using the same decision rule $\delta(X)$ repeatedly by different sample data $X$. It is defined by $R(\theta, \delta) = E_\theta^X [L(\theta, \delta(X))]$, averaging over $X$ for fixed $\theta$. Since $\theta$ is unknown, $R(\theta, \delta)$ is a function of $\theta$.

Decision theory is not necessarily Bayesian. It applies to any decision making process involving actions, loss, and risk. A decision rule $\delta$ should be chosen to have small $R(\theta, \delta)$. In pharmaceuticals development, there are regulatory risk and sponsor risk. From a decision-theoretic point of view, multiple testing should control the worst regulatory risk that could happen in the parameter space at certain pre-specified level. That is, $\delta$ is constructed so that $\min \sup_{\theta \in \Theta} R(\theta, \delta) < \alpha$. 
However, if prior information is included in the decision making procedure, then it becomes Bayesian decision theory.

- **prior information $\pi(\theta)$**: Prior information of $\theta$ usually is the prior density of $\theta$ which is indicated by $\pi(\theta)$.

- **$\rho(\pi^*, a)$**: Bayesian expected loss of an action $a$ : $\rho(\pi^*, a)$ is an expected loss for an action $a$ averaging over different parameter spaces $\theta$. It is defined as $\rho(\pi^*, a) = E^\pi[L(\theta, a)]$. $\pi^*(\theta)$ is posterior distribution of $\theta$ after sample data involves.

  $\rho(\pi^*, a)$ is a function of an action $a$. The conditional Bayes decision principle is to choose $a^{\pi^*}$ which can minimize $\rho(\pi^*, a)$. Then the $a^{\pi^*}$ is called a Bayes action.

- **Bayes risk $r(\pi, \delta)$**: $r(\pi, \delta)$ is a risk of a decision rule $\delta$ with respect to a prior distribution $\pi$ on $\Theta$. It is another expected loss defined as $r(\pi, \delta) = E^\pi[R(\theta, \delta)]$ averaging over both $\theta$ and $X$.

  For Bayes risk $r(\pi, \delta)$, the Bayes risk principle is to choose $\delta^*$ that minimized the Bayes risk.

### 1.2.2 Decision theory for testing two null hypotheses

To show the connection between multiple testing in pharmaceutical development and decision theory, we illustrate decision theory in testing two null hypotheses in a clinical trial. The scenario could be (1) test primary endpoint and then secondary endpoint in a study, (2) study involves an interim analysis and a final analysis for the same endpoint, or (3) accelerated approval based on a surrogate endpoint and
final approval based on a clinical endpoint. We discuss the basic elements in these scenarios and consider reasonable definitions of error rates for each scenario in section 1.3.

The two hypotheses we will discuss in this section are: $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. Parameters $\theta_1$ and $\theta_2$ represent efficacy of a compound which could be a difference of means, an odds ratio, or a risk ratio between treatment and control groups. Now we consider basic elements of decision theory in testing two null hypotheses in a clinical study.

- states of nature $\theta_1$ and $\theta_2$:
  We want to test $\{H_{0i} : \theta_i \in \Theta_i, i = 1, 2\}$. For each unknown parameter $\theta_i$, either $\theta_i \in \Theta_i$ or $\theta_i \in \Theta_i^C$. For convenience of notation later, we write:
  
  $$H_{0i} = \begin{cases} 
  1, & \text{if } H_{0i} \text{ is false or } \theta_i > 0 \\
  0, & \text{if } H_{0i} \text{ is true or } \theta_i \leq 0 \text{ for } i = 1, 2
  \end{cases}$$

- action $a_1$ or $a_2$:
  For testing each single null hypothesis, we can only reject or accept (not reject) it. The two possible decisions are:
  
  $$a_i = \begin{cases} 
  1, & \text{if reject } H_{0i} \\
  0, & \text{if not reject } H_{0i} \text{ for } i = 1, 2
  \end{cases}$$

- loss function $L(\theta_i, a_i)$:
  $L(\theta_1, a_1)$ is defined for all $(\theta, a) \in \Theta \times A$ and it is not dependent on sample data. If only a single null hypothesis is tested, we can create a 2X2 table (Table 1.2) to state the loss of an incorrect decision.
\[ H_{01} = 0 \ (\theta_1 \leq 0) \]
\[ H_{01} = 1 \ (\theta_1 > 0) \]

<table>
<thead>
<tr>
<th>Reject ( H_{01} )</th>
<th>( L(0, 1) = 1 )</th>
<th>( L(1, 1) = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 = 1 )</td>
<td>Type I error loss</td>
<td>(no loss)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Do not reject ( H_{01} )</th>
<th>( L(0, 0) = 0 )</th>
<th>( L(1, 0) = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 = 0 )</td>
<td>(no loss)</td>
<td>Type II error loss</td>
</tr>
</tbody>
</table>

Table 1.2: Loss for testing one single null hypothesis.

In the Table 1.2, we only consider whether an incorrect decision happens, so

\[
L(\theta_i, a_i) = \begin{cases} 
1, & \text{if an incorrect decision is made} \\
0, & \text{if correct decision is made for } i = 1, 2
\end{cases}
\]

In traditional measurement of incorrect decisions in pharmaceuticals, only loss from Type I error (regulatory error) is considered. However, decision theory does not only consider loss from regulatory error (Type I error) but also sponsors’ loss (Type II error loss). Therefore, possible loss should be decided from all the stakeholders (clinicians, health economists, policy makers). In addition, different weights of loss can be assigned regarding severity of incorrect decisions. For example, \( L(0, 1) = 2/3 \) and \( L(1, 0) = 1/3 \) shows that the loss from regulatory error is twice as big as loss from sponsor error.

Now we test two null hypotheses simultaneously and consider possible loss. There are four possible combinations of parameter spaces (either True or False for \( H_{01} \) and \( H_{02} \)) and four possible combinations for decisions (either Reject or Not Reject for \( H_{01} \) and \( H_{02} \)). We can create a 4 by 4 table to consider different loss \( L(\theta_1, a_1; \theta_2, a_2) \).
<table>
<thead>
<tr>
<th>$a_1 = 1$ and $a_2 = 0$ (Reject $H_{o1}$ but not reject $H_{o2}$)</th>
<th>$H_{o1}=0 \cap H_{o2}=1$</th>
<th>$H_{o1}=0 \cap H_{o2}=0$</th>
<th>$H_{o1}=1 \cap H_{o2}=0$</th>
<th>$H_{o1}=1 \cap H_{o2}=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L(0, 1; 1, 0) = f(L_1^I, L_2^{II})$</td>
<td>$L(0, 1; 0, 0) = L_1^I$</td>
<td>0</td>
<td>$L(1, 1; 1, 0) = L_2^{II}$</td>
<td></td>
</tr>
<tr>
<td>$a_1 = 1$ and $a_2 = 1$ (Reject $H_{o1}$ and reject $H_{o2}$)</td>
<td>$L(0, 1; 1, 1) = L_1^I$</td>
<td>$L(0, 1; 0, 1) = f(L_1^I, L_2^I)$</td>
<td>$L(1, 1; 0, 1) = L_2^I$</td>
<td>0</td>
</tr>
<tr>
<td>$a_1 = 0$ and $a_2 = 1$ (not reject $H_{o1}$ but reject $H_{o2}$)</td>
<td>0</td>
<td>$L(0, 0; 0, 1) = L_2^I$</td>
<td>$L(1, 0; 0, 1) = f(L_1^{II}, L_2^I)$</td>
<td>$L(1, 0; 1, 1) = L_1^{II}$</td>
</tr>
<tr>
<td>$a_1 = 0$ and $a_2 = 0$ (not reject $H_{o1}$ and not reject $H_{o2}$)</td>
<td>$L(0, 0; 1, 0) = L_2^{II}$</td>
<td>0</td>
<td>$L(1, 0; 0, 0) = L_1^{II}$</td>
<td>$L(1, 0; 1, 0) = f(L_1^{II}, L_2^{II})$</td>
</tr>
</tbody>
</table>

Table 1.3: Loss Table for testing two null hypothesis simulataneously. $L_i^I$ and $L_i^{II}$ indicate loss from Type I and Type II error for $i=1,2$.

Since the two null hypotheses are tested simultaneously, the loss function $L(\theta_1, a_1; \theta_2, a_2)$ should depend on two states of nature and two decisions. From Table 1.3, the loss function could contain loss from the first test ($L_j^I, j = I, II$) or the second test ($L_j^I, j = I, II$) or both. And, the loss could be from Type I error ($L_i^I, i = 1, 2$) or Type II error ($L_i^{II}, i = 1, 2$). If $L(\theta_1, a_1; \theta_2, a_2)$ contains losses from both tests, the loss is a function of two losses, $f(L_1^I, L_2^I)$, regarding big or small of each loss. For example, $L(0, 1; 0, 1) = L_1^I + L_2^I$ assigns equal severity for making a type I error in both null hypotheses. $L(0, 1; 1, 0) = 2/3 * L_1^I + 1/3 * L_2^{II}$ considers loss from Type I error to be twice as big as loss from Type II error. Furthermore, the loss function can be defined in various ways by combining the opinions from all the stake-holders (clinicians, health economists, policy makers). It is not necessarily the sum of two losses. However, if the loss
function $L(\theta_1, a_1; \theta_2, a_2)$ is sum of losses, it would be easier to compute the risk (expectation of loss).

Here we discuss three different definitions of loss function in order to connect to the controlling error rate in section 1.1. Let $V$ indicate the number of incorrect rejections when null hypotheses are true. $R$ is the total number of rejections. The numerical losses for three different losses are expressed in Table 1.4.

1. $L(\theta_i, a_i) = I\{V > 0\}$: "0-1" loss occurs when at least one true null hypothesis is rejected. Controlling this loss in the long term is equal to control FWER.

2. $L(\theta_i, a_i) = V$: loss is the number of Type I errors. Controlling this loss in the long term is equal to controlling $E(V)$.

3. $L(\theta_i, a_i) = V/R$: loss is the proportion of type I errors among all rejections. Controlling this loss in the long term is equal to controlling FDR.

• random variable $X_1$ and $X_2$:
  $X_i$ is the measurements on endpoints $\theta_1$ and $\theta_2$ from the clinical trial. They could be binary to indicate response or non-response, or they could be continuous. $X_1$ and $X_2$ could also be measurements for the same endpoint at different time points, if the clinical study has an interim analysis and a final analysis. Or they could be measurements for different endpoints, such as primary and secondary endpoints, or surrogate and clinical endpoints.

• decision rule $\delta(x_i, A_i)$:
  A decision rule specifies an action given sample data. In a clinical trial, $\delta(x_i, A_i)$
is a testing method to decide whether to reject or not reject the null hypothesis based on an observed data point $X_i=x_i$. For example, the rejection region for $H_i$ from a certain testing method could be $C=\{T_i > c_i\}$ where $T_i$ is the test statistic computed from $x_i$ and $c_i$ is the critical value for $H_i$. The decision rule for rejecting $H_i$, $a_i=1$, is then:

$$\delta(x_i, a_i = 1) = \begin{cases} 
1, & \text{if } T_i > c_i \\
0, & \text{if } T_i \leq c_i \text{ for } i = 1, 2 
\end{cases}$$

- risk function $R(\theta_i, \delta_i)$:

Controlling the error rate for one testing procedure in a clinical trial is like controlling the loss from a decision rule $\delta(x_i, A_i)$, $L(\theta_i, \delta(X_i))$, in the long term. By Kolmogorov’s version of the Law of Large Numbers (without an identical distribution assumption), risk can be controlled across different studies for different compounds for different diseases in the long run, if we control each study at a
certain level of loss.

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} L(\theta_i, \delta(X_i)) \to E[L(\theta_i, \delta(X_i))] = R(\theta_i, \delta)
\]

Take the testing of one null hypothesis as an example. The risk function can be computed based on "0-1" loss in Table 1.2 when the null hypotheses is true (i.e. \( \theta_1 \in \Theta_1 \)):

\[
R_{\theta_1 \in \Theta_1}(\theta_1, \delta) = E_{\theta_1 \in \Theta_1}^X [L(\theta_1, \delta(X))]
\]

\[= 1 \cdot P_{\theta_1 \in \Theta_1}(\text{reject } H_1) + 0 \cdot P_{\theta_1 \in \Theta_1}(\text{not reject } H_1)
\]

\[= 1 \cdot P(\text{Type I error}) = \alpha(\theta)
\]

On the other hand, the risk when the null hypotheses is false (i.e. \( \theta_1 \in \Theta_1^C \)) is type II error rate.

\[
R_{\theta_1 \in \Theta_1^C}(\theta_1, \delta) = E_{\theta_1 \in \Theta_1^C}^X [L(\theta_1, \delta(X))]
\]

\[= 0 \cdot P_{\theta_1 \in \Theta_1^C}(\text{reject } H_1) + 1 \cdot P_{\theta_1 \in \Theta_1^C}(\text{not reject } H_1)
\]

\[= 1 \cdot P(\text{Type II error}) = \beta(\theta)
\]

When only one hypothesis is tested, the risk function is \( \alpha(\theta) \) and \( \beta(\theta) \) which are function of type I and type II error rates. It is a continuous function of \( \theta \) and it has a jump at \( \theta_1 = \theta_1^0 = 0 \) (Figure 1.1). The traditional frequentist approach is to control the type I error rate \( R_{\theta_1 \in \Theta_1}(\theta_1, \delta) = \alpha(\theta) \leq 5\% \), while minimizing type II error rate \( R_{\theta_1 \in \Theta_1^C}(\theta_1, \delta) = \beta(\theta) \) (equivalent to maximizing power function).
Figure 1.1: Risk function for testing one hypothesis

When two hypotheses are tested, the risk function becomes a smooth surface in three dimensional space and it has a scar at the boundary of each disjoint parameter subspace. Therefore, how to evaluate a decision gets more complicated. However, we can still show the risk function for each parameter subspace by using the loss listed in Table 1.3. A simplified formulation assumes the test statistics for $\theta_1$ and $\theta_2$ are independent which is not always true in the clinical
studies.

\[ R_1 = R_{(\Theta_1 \cap \Theta_2^c)}(\theta_i, \delta_i) = E_{X(\Theta_1 \cap \Theta_2^c)}^X[L(\theta_i, \delta(X_i))] \]

\[ = f(L_1^I, L_2^II) \ast P_{(\Theta_1 \cap \Theta_2^c)}(\text{reject } H_1 \text{ and not reject } H_2) \]

\[ + L_1^I \ast P_{(\Theta_1 \cap \Theta_2^c)}(\text{reject } H_1 \text{ and reject } H_2) \]

\[ + L_2^II \ast P_{(\Theta_1 \cap \Theta_2^c)}(\text{not reject } H_1 \text{ and not reject } H_2) \]

\[ = f(L_1^I, L_2^II) \ast \alpha_1(\theta_1)\beta_2(\theta_2) + L_1^I \ast \alpha_1(\theta_1)(1 - \beta_2(\theta_2)) + L_2^II \ast (1 - \alpha_1(\theta_1))\beta_2(\theta_2) \]

Similarly,

\[ R_2 = R_{(\Theta_1 \cap \Theta_2)}(\theta_i, \delta_i) \]

\[ = L_1^I \ast \alpha_1(\theta_1)(1 - \alpha_2(\theta_2)) + f(L_1^I, L_2^I) \ast \alpha_1(\theta_1)\alpha_2(\theta_2) + L_2^I \ast (1 - \alpha_1(\theta_1))\alpha_2(\theta_2) \]

\[ R_3 = R_{(\Theta_1^c \cap \Theta_2)}(\theta_i, \delta_i) \]

\[ = L_2^I \ast (1 - \beta_1(\theta_1))\alpha_2(\theta_2) + f(L_1^II, L_2^I) \ast \beta_1(\theta_1)\alpha_2(\theta_2) + L_1^II \ast \beta_1(\theta_1)(1 - \alpha_2(\theta_2)) \]

\[ R_4 = R_{(\Theta_1^c \cap \Theta_2^c)}(\theta_i, \delta_i) \]

\[ = L_2^II \ast (1 - \beta_1(\theta_1))\beta_2(\theta_2) + L_1^II \ast \beta_1(\theta_1)(1 - \beta_2(\theta_2)) + f(L_1^II, L_2^I) \ast \beta_1(\theta_1)\beta_2(\theta_2) \]

Furthermore, we apply the three different definitions of loss in Table 1.4 to compute the risk function to show FWER, E(V) and FDR over different parameter subspaces by assuming independence between two test statistics. The result is shown in Table 1.5.
Table 1.5: \( R(\theta_i, \delta_i) \) for different definitions of loss assuming independent between two test statistics.

### 1.3 Error rates in accelerated approval

Even though different types of error rates have been discussed, FWER controlling is commonly used to guard against making any false claims on the efficacy of any endpoint in clinical studies. FWER treats 1 incorrect rejection and 2 incorrect rejections as equally bad. This may be reasonable when *any* incorrect rejection is unacceptable.

Two different scenarios are discussed and considered: [1] taking two looks at a clinical endpoint, with potential full approval from data at the interim, and [2] testing primary and secondary endpoints simultaneously. For [1] FWER is fine of course, but for [2] FWER does not take into account the different consequences of incorrect rejections. The consequence of incorrect inference on the primary endpoint(s) is more serious than incorrect inference on secondary endpoints.

In this section, we will especially discuss the appropriate error rate for accelerated approval. Accelerated approval is one of the approaches developed by the U.S. Food
and Drug Administration (FDA) to speed up the availability of drugs to cure serious
diseases. Accelerated approval regulation allows earlier approval of drugs based on
a surrogate endpoint to substitute a clinically meaningful outcome. A surrogate
endpoint is a marker - a laboratory measurement, or physical sign that can represent
the primary clinical endpoint. After conditional approval, drug companies still need
to conduct a confirmatory trial to prove the efficacy in the clinically meaningful
outcome. Then, the FDA grants full approval for the drug. Otherwise, the FDA
might withdraw the conditional approval of the drug from market.

• **Controlling FWER is inappropriate**

  The FDA accelerated approval draft guidance (May 2012) suggests to split *alpha*
between surrogate endpoint and the clinically meaningful endpoint to control
the false positive rate (Type I error). It is not easy to interpret if the requirement
is FWER. In the case of accelerated approval, since full approval can be denied,
the possibility of an incorrect conditional approval is intentionally allowed. An
incorrect full approval is worse than an incorrect conditional approval. We
suggest consideration of error rates more insightful than FWER, of error rates
that acknowledge 2 incorrect decisions are worse than 1 incorrect decision to
start.

• **Controlling FDR is inappropriate**

  Error rate control in discovery studies often controls the false discovery rate
(FDR), which is the (unconditional) proportion of incorrect rejections, across
infinitely many families of hypotheses tested. Controlling FDR is not appropriate in the accelerated approval setting since the rate of incorrect full approval can be inflated by efficacy in the surrogate endpoint.

We can actually simply calculate that the error rate is inflated as twice $\alpha$. If the effect of surrogate endpoint is highly efficacious such that the probability of rejection is close to 1, then the clinical endpoint can be tested with a Type I error rate higher than $\alpha$ while FDR is still controlled at $\alpha$.

$$\alpha \approx \frac{1}{2} P\{V = 0\} + \frac{1}{2} P\{V = 1\} = \frac{1}{2} P\{\text{Reject } H_{0C}\}.$$ 

If $\alpha$ is 5%, then $H_{0C}$ is actually tested at 10%.

We recommend a decision-theoretic approach (not the same as a Bayesian approach), involving all the stake-holders (clinicians, health economists, policy makers) in the determination of the Losses from incorrect decisions (for example, in Table 1.6), which will then enable statisticians to calculate Risks (in essence, more appropriate “error rates”). Remember the Losses do not depend on data but depend on the parameters ($H_{0S}$ and $H_{0C}$ is true or false) and actions (reject or not reject null hypothesis). This is our recommendation for future consideration.

1.4 Partition testing principle for controlling error rates

A principle of multiple testing which is conditional in nature is the Partitioning Principle of Stefansson, Kim, and Hsu (1988), which was further refined by Finner and Strassburger (2002). Holm’s step-down method and Hochberg’s (1988) step-up method, each comparing p-values against a data-dependent threshold, can be thought of as partition testing, obtained by applying the Bonferroni’s inequality and
| accelerated approval but no full approval (Reject $H_{0s}$ but not reject $H_{0c}$) | labelling error | conditional approval error | (surrogacy + conditional approval errors) | No Type I error (sponsor loss) |
| accelerated approval + full approval (Reject $H_{0s}$ and reject $H_{0c}$) | labelling error | conditional approval + full approval errors | (surrogacy + conditional approval) + full approval errors | No error |
| No accelerated, no full approval | | potential sponsor loss | | |
| No accelerated, but full approval | | | usually not happen | |

Table 1.6: Possible losses from policy makers (from clinicians, health economists and other stake-holders). T indicates True null hypothesis and F indicates false null hypothesis.

a modification of Simes’ (1986) inequality, respectively, to component partitioning tests (see Huang and Hsu, 2005). The step-down version of Dunnett’s (1955) method in Marcus, Peritz, and Gabriel (1976) can be thought of as partition testing taking into account the multivariate $t$ joint distribution of the test statistics (see Stefansson, Kim, and Hsu, 1988).
1.4.1 Partition vs. Closed testing for controlling FWER

In testing multiple hypotheses, statistical methods must account for the possibility that different combinations of the null hypotheses may be true. For controlling FWER, two standard principles for constructing multiple tests are the Closed testing principle and the Partition testing principle. The concise descriptions of these principles in this section are slightly different from those given in earlier literature, to highlight their difference in the genetic subgroup testing problem.

One advantage Partition testing has over Closed testing is that, when decision-making involves paths, Partition testing formulates the null hypotheses in a way that ensures statistical inferences will respect the decision paths. Consider an efficacy study with multiple doses and primary-secondary endpoints, for example. Then, at a given dose, efficacy of a secondary endpoint is not tested unless efficacy of the primary endpoint can be established. Thus, decision-making follows a path within each dose. Liu and Hsu (2009) showed how Partition testing naturally formulates null hypotheses to ensure decision-making follows such paths. Another example of how Partition testing does this, in the genetic subgroup testing setting, is given in section 4.2. Note that, in the Gatekeeping literature (e.g., Dmitrienko, Tamhane, and Wiens 2008, Dmitrienko and Tamhane 2011), decision paths are referred to as logical relationships among null hypotheses. Typical Gatekeeping methods, which are Closed tests, test all possible intersections of the null hypotheses regardless of a study’s decision paths, and then impose restrictions on how the intersection hypotheses can be tested to ensure statistical inferences respect decision paths.

Another advantage of Partition testing is that it readily recognizes logical relationships among the parameters in the null hypotheses being tested. Not to be confused
with logical relationships among null hypotheses, we show in sections 3.2 and 4.1 the importance of such cognition, and that the popular Hochberg’s method (Hochberg 1980) is not cognizant of logical relationships among parameters.

A third advantage Partition testing has over Closed testing in general, and the Hochberg method in particular, is that by constructing only the null hypotheses that need to be tested, it makes clearer the extent to which the joint distribution of the test statistics can be taken into account (in calculating critical values).

Let \( \theta_i, i = 1, \ldots, k \), be the parameters of interest and let \( \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \). Consider testing the \( k \) null hypotheses

\[
H_{0i}: \theta_i \in \Theta_i, \ i = 1, \ldots, k. \tag{1.1}
\]

To test (1.1), Partition testing proceeds as follows.

P1: Partition \( \Theta_1 \times \cdots \times \Theta_k \). How to partition depends on whether there are Decision Paths between the hypotheses of interest. If there are no Decision Paths, then it can be done as follows. For each \( I \subseteq \{1, \ldots, k\} \), form

\[
\Theta_I^* = \{ \boldsymbol{\theta} | \theta_i \in \Theta_i \text{ for all } i \in I \text{ and } \theta_j \notin \Theta_j \text{ for all } j \notin I \}. \tag{1.2}
\]

Think of \( \Theta_I^* \) is the parameter sub-space in which \( H_{0i}, i \in I \), are true and \( H_{0j}, j \in I^c \), are false. (Here the superscript \( ^c \) denotes the complement of a set.) So, if \( I \neq J \), then \( \Theta_I^* \cap \Theta_J^* = \emptyset \). And the true parameter value belongs to exactly one \( \Theta_I^* \).

P2: For each \( \Theta_i^* \neq \emptyset, I \neq \emptyset \), test \( H_{0I}^* : \boldsymbol{\theta} \in \Theta_I^* \) at level-\( \alpha \). (The hypothesis \( H_{0\emptyset}^* \) need not be tested since it corresponds to all \( H_{0i}, i = 1, \ldots, k \), being false.)
P3: For each \( i \), reject \( H_{0i} \) and infer \( \theta_i \notin \Theta_i \) if and only if all \( H_{0i}^* \) with \( \Theta_i^* \cap \Theta_i \neq \emptyset \) are rejected. This is a logical conclusion because all \( H_{0i}^* \) that implies \( H_{0i} \) could be true have to be rejected in order to reject \( H_{0i} \).

The number of non-empty \( \Theta_i^* \) is \( 2^k \), or less, less if the null hypotheses \( H_{0i}, i = 1, \ldots, k \), are logically related in terms of parameters. For an example different from the one in section 3.1, consider testing the pairwise equality of three means, a setting considered in Shaffer (1986) and Holland and Copenhaver (1987), \( H_{01} : \theta_1 = \mu_1 - \mu_2 = 0, H_{02} : \theta_2 = \mu_2 - \mu_3 = 0, H_{03} : \theta_3 = \mu_3 - \mu_1 = 0 \). Then, \( \Theta_{\{1,2\}}^* = \emptyset \) because it is impossible for \( \mu_1 = \mu_2, \mu_2 = \mu_3 \), but \( \mu_3 \neq \mu_1 \). Therefore, the only non-empty \( \Theta_i^* \) are \( \Theta_{\{1\}}, \Theta_{\{2\}}, \Theta_{\{3\}}, \Theta_{\{1,2\}}, \Theta_{\emptyset} \).

Since the null hypotheses \( H_{0i} \)'s are disjoint, at most one \( H_{0i}^* \) can be true. Therefore, even though there is no multiplicity adjustment in testing the up to \( 2^k - 1 \) hypotheses \( H_{0i}^* \), Partition testing controls FWER strongly.

The Closed testing approach of Marcus, Peritz, and Gabriel (1976) is as follows regardless of whether there is a Decision Path between the hypotheses of interest or not.[20]

C1: For each \( I \subseteq \{1, \ldots, k\} \), form \( \Theta_I = \{\theta|\theta_i \in \Theta_i \text{ for all } i \in I\} \).

C2: Test each \( H_{0I} : \theta \in \Theta_I \) at level-\( \alpha \).

C3: For each \( i \), reject \( H_{0i} \) and infer \( \theta_i \notin \Theta_i \) if and only if all \( H_{0I} \) with \( i \in I \) are rejected.

Closed Testing guarantees strong control of FWER because \( \Theta_i^* \subset \Theta_i \), so a level-\( \alpha \) test for \( H_{0I} \) is also a level-\( \alpha \) test for \( H_{0I}^* \).
Compared to Partition testing, the parameter sub-space $\Theta_I$ tested by Closed testing lacks the specification “and $\theta_j \notin \Theta_j$ for all $j \notin I$” that the parameter sub-space $\Theta_{\theta_I}$ tested by Partition testing has. This difference is what lets Partition testing recognize logical relationships among parameters in the null hypotheses $H_{0i}, i = 1, \ldots, k$.

### 1.4.2 iPartition for controlling expected loss

In section 1.4.1, the Partition testing principle has been described specifically for controlling FWER. However, we have discussed that different error rates should be considered in different scenarios, and error rate control actually is identical to control of expected loss in decision theory. Here we further extend the concept of the Partitioning principle to a general version which is named iPartition to control expected loss for different types of error rates. This has been discussed in Xu and Hsu (2007) and the dissertation of Yi Liu (2009).

The iPartition Principle states:

- **P1.** Partition the parameter space $\Theta_1 \times \cdots \times \Theta_k$ into disjoint sub-spaces. For each $I \subseteq \{1, \ldots, k\}$, form

  $$\Theta_I^* = \{\theta_i \in \Theta_i \text{ for all } i \in I \cap \theta_j \notin \Theta_j \text{ for all } j \notin I\}.$$ 

  $\Theta_I^*$ is the parameter sub-space in which $H_{0i}, i \in I$, are true and $H_{0j}, j \in I^c$, are false.

- **iP2.** In each $\Theta_I^*$, test $\{H_{0i} : \theta_i \in \Theta_i, i \in I\}$ at level $\alpha$ controlling,

  $$\sup_{\theta \in \Theta_I^*} E_{\theta} [L(V, S)] \leq \alpha \quad (1.3)$$ 

  Define $H_{0i}$ to be **I-rejected** if it is rejected in $\Theta_I^*$. 

22
iP3. Reject \( H_{0i} \) and infer \( \theta_i \notin \Theta_i \) if and only if \( H_{0i} \) is \( I \)-rejected for all \( I, \ I \subseteq K \).

The key difference in \textit{iPartition} compared with classical Partitioning is that \textit{iPartition} must test each individual \( H_{0i}, \ i \in K \) in each \( \Theta_i^* \). Here \( i \) in \textit{iPartition} can be referred as "\( I \)-reject" each "individual" in each sub-space.

The advantage of \textit{iPartition} is that \( I \)-reject can show detailed rejections for each \( H_{0i} \) in each sub-space. Only exactly one \( \Theta_i^* \) contains the true parameter \( \theta^* \). It then controls expected loss by controlling expected loss within each \( \Theta_i^* \).

The generalization happens in [iP2.]. Error rates can be generalized to expected loss for defining the loss function as the number of incorrect rejections (\( L(V) \)) or the number of incorrect rejections and the number of correct rejections (\( L(V, S) \)), defined in section 1.2.2.

We will show the proof of controlling \( E[L(V)] \) and \( E[L(V, S)] \) in detail in the following sections.

\textbf{iPartition for controlling \( E[L(V)] \)}

When the loss function only considers the mistake of incorrect rejections, it could be 0-1 loss counting "1" for at least \( (m+1) \) incorrect rejections or counting the number of incorrect rejections. For example,

- \( \setlength{\parskip}{1ex} \notag L(V) = I\{V > 0\} \) is used to control FWER. (Classical Partitioning principle is just a special version of \textit{iPartition}.)

- \( \setlength{\parskip}{1ex} \notag L(V) = I\{V > m\} \) is used to control gFWER.

- \( \setlength{\parskip}{1ex} \notag L(V) = V = \sum I\{\text{incorrect rejection for } H_{0i}\} \) is used to control \( E(V) \).
Theorem 1.4.1 Let $L_I(V)$ be the loss from $I$-rejects in $\Theta_i^*$. Controlling $E[L_I(V)]$ for testing each partition hypothesis $\Theta_i^*, I \subseteq K$ following iPartition principle guarantees the overall control of $E[L(V)]$ under assumption A1.

Assumption A1: In [iP2], automatically $I$-reject all $H_{0i}, \ i \notin I$.

Proof: Suppose the true $\theta^* \in \Theta_i^*(V)$. The loss $L_I^*(V)$ could be 0-1 loss by indicating 1 for at least $(m+1)$ $I$-rejects in $\Theta_i^*$, or it could be the number of $I$-rejects in $\Theta_i^*$. $L(V)$ is the same definition of loss for incorrect rejections for $H_i, i \in I^*$. Then $L \leq L_I^*(V)$, since the rejection of any true $H_{0i}, i \in I^*$ implies it has to be $I$-rejected for all $I \subseteq K$ which includes $I^*$.

$$
\sup_{\theta \in \Theta_i^*} E_{\theta}[L(V)] = \sup_{\theta \in \Theta_i^*} E_{\theta} \text{(loss from incorrect rejections for } H_i, i \in I^*) \\
\leq \sup_{\theta \in \Theta_i^*} E_{\theta} \text{(loss from } I^*\text{-rejects in } \Theta_i^*) \\
= \sup_{\theta \in \Theta_i^*} E_{\theta}[L_I^*(V)] \leq \alpha
$$

Consider, for example $k = 3, \Theta_i = \{\theta_i \leq 0\}$. Then there are $2^3 = 8 \Theta_i^*$’s. Each $H_{0i}: \theta_i \leq 0$ is tested eight times, once in each $\Theta_i^*$, and is ultimately rejected only if it is rejected in each $\Theta_i^*$. Within each $\Theta_i^*$, $H_{0i}$ for $\theta_i > 0$ are automatically rejected, while $H_{0i}$ for $\theta_i \leq 0$ are tested with multiplicity adjustments in such a way that $E[L(V)]$ is controlled at level-\(\alpha\) within $\Theta_i^*$.

For $I = \{2, 3\}$, traditional partition testing controlling FWER, or equivalently gFWER with $m = 0$, would test $H_{02}$ and $H_{03}$ within $\Theta_{\{2, 3\}}^*$ so that the probability of rejecting at least one is no more than $\alpha$.

To see that the iPartition controls gFWER, suppose, for example, $\theta^* = (1, -1, -1)$. Then, with $m = 1$, $H_{01}$ is automatically rejected while $H_{02}$ and $H_{03}$ are tested so that
the probability of rejecting both is no more than $\alpha$. Since $H_{02}$ and $H_{03}$ will be rejected if they are $I$-rejected in $\Theta_i^*$ for all $I \subseteq K$, the probability that they will both be rejected is no more than the probability that both are $I$-rejected for $I = \{2,3\}$, and that probability is no more than $\alpha$ by [IP2] of the iPartition principle.

To illustrate the proof, let $k = 3, m = 1$, gFWER is equal to zero in $\theta^* \in \Theta_1^*$, or $\theta^* \in \Theta_2^*$, or $\theta^* \in \Theta_3^*$, since the chance of at least two incorrect rejections in these sub-spaces is zero. Suppose $\theta^* \in \Theta_2^*$,

$$
\sup_{\theta \in \Theta_2^*} gFWER = \sup_{\theta \in \Theta_2^*} E_{\theta}(I\{V > 1\})
= \sup_{\theta \in \Theta_2^*} P_{\theta}\{\text{reject } H_{02} \text{ and reject } H_{03}\}
= P\{I\text{-reject } H_{02} \text{ in } \Theta_2^*, \Theta_3^*, \Theta_{13}^* \text{ and }
I\text{-reject } H_{03} \text{ in } \Theta_{13}^*, \Theta_{23}^*, \Theta_{123}^* \}
\leq P\{I\text{-reject } H_{02} \text{ and } H_{03} \text{ in } \Theta_2^* \} \leq \alpha.
$$

To see how the iPartition controls $E(V)$, suppose the true $\theta^* \in \Theta_2^*$. Let $V_{23}$ be the numbers of true null hypotheses $I$-rejected in $\Theta_2^*$. $V_{23}$ could be 0, 1, 2. Then $V \leq V_{23}$, since the incorrect rejections of any true $H_{0i}, i \in \{2,3\}$ implies it has to be $I$-rejected for all $I \subseteq K$ which includes $I = \{2,3\}$.

$$
\sup_{\theta \in \Theta_2^*} E_{\theta}(V) = \sup_{\theta \in \Theta_2^*} E_{\theta}(I\{\text{reject } H_{02}\} + I\{\text{reject } H_{03}\})
\leq \sup_{\theta \in \Theta_2^*} E_{\theta}(I\{\text{I-reject } H_{02} \text{ in } \Theta_2^*\} + I\{\text{I-reject } H_{03} \text{ in } \Theta_2^*\})
= \sup_{\theta \in \Theta_2^*} E_{\theta}(V_{23}) \leq \alpha
$$
iPartition for controlling $E[L(V,S)]$

Loss function can be defined as the mistakes from incorrect rejections and correct rejections. For example, FDR control is controlling the loss of proportion of incorrect rejections among all rejections.

- $L(V,S) = \frac{V}{V+S} I(V > 0)$ where $S$ is numbers of correct rejections.

To obtain the information of $S$ in $\Theta_i^*$, modified iPartition tests not only hypotheses $H_{0i}, i \in I$ that are true, but also hypotheses that are false $H_{0i}, i \in K \setminus I$.

**Theorem 1.4.2** Let $L_I(V,S)$ be the loss from $I$-rejects in $\Theta_i^*$. Controlling $E[L_I(V,S)]$ for testing each partition hypothesis $\Theta_i^*, I \subseteq K$ following iPartition principle guarantees the overall control of $E[L(V,S)]$ under assumption A2.

**Assumption A2:** If $H_{0i}$ is $J$-rejected for any $i \notin J$, then it is automatically $I$-rejected for every $I \in i$. That is, if $H_{0i}$ is $J$-rejected for any $i \notin J$, then we assume it is finally rejected.

**Proof:** Suppose the true $\theta^* \in \Theta_i^{**}$, that means $H_{0i}, i \in I^*$ are true, and $H_{0i}, i \in K \setminus I^*$ are false. Let $V^*(\leq |I^*|)$ be the number of true null hypotheses being rejected, and $S^*(\leq k - |I^*|)$ be the number of false null hypotheses being rejected, then

a. $V^* \leq V_I$, since the rejection of any true $H_{0i}, i \in I^*$ implies it has to be $I^*$-rejected.

b. $S_I^* \leq S^*$ since if the false null hypothesis $H_{0i}, i \in K \setminus I^*$ is $I^*$-rejected then by assumption A2 ($i \notin I^*$), it is $I$-rejected for every $I \ni i$, which leads to the rejection of $H_{0i}$.
Combining a and b, we have $\frac{V^*}{S^*} \leq \frac{V_{I^*}}{S_{I^*}}$, which leads to

\[
sup_{\theta \in \Theta^*_{\lambda}} FDR_{\theta} = sup_{\theta \in \Theta^*_{\lambda}} E_{\theta} \left( \frac{V^*}{V^* + S^*} I(V^* > 0) \right) \\
\leq sup_{\theta \in \Theta^*_{\lambda}} E_{\theta} \left( \frac{V_{I^*}}{V_{I^*} + S_{I^*}} I(V_{I^*} > 0) \right) \leq \alpha.
\]

We use the same example from the previous section to illustrate controlling the FWER. Suppose the true $\theta^* \in \Theta^*_{23}$, that means $H_{02}, H_{03}$ are true, and $H_{01}$ is false. Let $V^* (\leq 2)$ be the number of true null hypotheses being rejected, and $S^* (\leq 1)$ be the number of false null hypotheses being rejected. Then

a. $V^* \leq V_{\{2,3\}}$ since the rejection of either $H_{02}$ or $H_{03}$ implies they need to be $\{2,3\}$-rejected.

b. $S_{\{2,3\}} \leq S^*$. If $H_{03}$ is $\{2,3\}$-rejected ($S_{\{2,3\}} = 1$), by assumption A2, it implies the rejection of $H_{03}$, i.e. $S^* = 1$.

We have $\frac{V^*}{S^*} \leq \frac{V_{\{2,3\}}}{S_{\{2,3\}}}$, and thus FDR is controlled.
Chapter 2: New and old formulations of genetic subgroups analysis

2.1 Potential of personalized medicine

Traditionally, drug companies pursue marketing approval for the broadest possible patient population for each compound. However, when a compound is insufficiently efficacious when measured over a broad patient population, if it is sufficiently efficacious for a subgroup of the patients, then its sponsor may seek approval to market to that subgroup. There are two possible paths to the approval process:

1. Conduct a clinical study aiming to prove broad efficacy of the compound. If the study fails to show broad efficacy, but a retrospective analysis indicates efficacy in a subgroup, then conduct another (independent) study to validate efficacy in that subgroup.

2. If there is the potential that the compound will be more efficacious in some subgroup(s), then plan and conduct a study testing for efficacy in the broad patient population, and/or efficacy in a pre-specified subgroup(s), adjusting for multiplicity in the analysis as appropriate.
The purpose of this chapter is to explain inadequacies in previous proposals, and propose a new problem formulation.

The efficacy of a drug is typically defined in terms of its average effect, usually compared to a control group, possibly measured relative to baseline measurements. (Note that responder rates can be considered averages also, averages of indicator variable of responders.)

In personalized medicine and targeted therapy, a compound may be tested for efficacy averaged over the entire patient population, and in one or more subgroups. The scenarios we discuss below are such that statistical testing adjusted for multiplicity of groups can retain sufficient power to establish efficacy, while still controlling the familywise error rate (FWER).

2.2 Motivating examples

Cancer drugs may target specific subgroups. Herceptin targets breast cancer patients with HER-2/neu over expression, for example. Another recent example is vemurafenib, which targets breast cancer patients with a mutation in the BRAF gene. A potential new example comes from PARP inhibitors for breast cancer.

PARP inhibitors for triple-negative breast cancer patients inhibit the PARP1 enzyme. The PARP1 enzyme and the BRCA genes work in concert to repair DNA damage, enabling survival of cancer tumors. In theory then, when PARP inhibitors inhibit the PARP1 enzyme, tumors in patients with mutation in the BRCA genes may be less able to repair DNA damage and die. Thus, there is the potential that patients with BRCA mutation may benefit more from PARP inhibitors than patients without such mutation.
A Phase III trial for PARP inhibitors taking the second path can thus aim to prove broad efficacy only, or aim to prove efficacy in the BRCA-mutated subgroup only, or simultaneously test for broad efficacy and efficacy in the BRCA-mutated subgroup adjusting for multiplicity.

Sanofi conducted a Phase III trial for *iniparib*, a PARP inhibitor, pursuing broad indication without pre-planning for any subgroup testing. The results from this study with 519 women from 109 U.S. sites failed to meet a pre-specified goal for the co-primary endpoints of overall survival and progression-free survival. It seems plausible that future Phase III trials of PARP inhibitors may stratify on patients with BRCA mutations, either testing for broad efficacy and efficacy in the BRCA-mutated subgroup simultaneously, or testing for efficacy in the BRCA-mutated subgroup only.

Mutations in genes involved in metabolic and transport pathways may also affect the efficacy of drug compounds. Examples of such genes include genes in the cytochrome P450 family such as 2C19 and 2D6, and the ABCB1 transporter gene.

The cytochrome P450 family of enzymes (abbreviated as CYP) is responsible for metabolizing more than half of the drugs on the market to their active forms. A P450 gene can be highly polymorphic. For example, CYP2D6 has more than 90 alleles. A patient with variants of a P450 gene involved in the metabolic pathway of a drug might derive differential benefits from that drug.

However, which variant of a P450 gene may allow patients to receive maximum benefit from a drug can be difficult to anticipate. If dosing turns out to be inappropriate for wild type patients, then loss-of-function carriers or gain-of-function carriers may do better than the wild type. In an analysis of how polymorphism in CYP2C19
affected the efficacy of clopidogrel (Plavix) on patients in the CURE study, point estimate of hazard ratio in the primary endpoint shown in Table 1 of Paré et al. (2010) was actually lower for poor and ultra metabolizers than for patients with wild type CYP2C19.

2.3 Previous proposals

Traditionally, broad efficacy needs to be established before exploring the efficacy in a genetic subgroup. However, this conventional thinking has been changed in subgroup analysis of personalized medicine. The sponsor might want to show the efficacy in the pre-determined subgroup even if the broad efficacy cannot be demonstrated. Several important references in genetic subgroup analysis will be discussed throughout this section by discussing their problem formulation, parameter definition and hypotheses setting, subgroup analysis methodology, consistency requirement and power computation separately.

2.3.1 Previous formulations

Most references formulated the problem in genetic subgroup analysis only considering the drug efficacy in overall patients population and in one pre-determined subgroup ignoring the efficacy in the complementary subgroup. However, there still exists some concerns about the efficacy in the complementary subgroup. Therefore, some papers started to discuss the requirement of consistency in genetic subgroup analysis.

Wang, O’Neil and Hung (2007) [32] from the Center for Drug Evaluation and Research (CDER) of the Food and Drug Administration (FDA) proposed an alpha-split method to test broad efficacy and efficacy in a pre-determined genetic subgroup
simultaneously and took correlation between parameters into account to increase power. Therefore, a pre-specified genetic subgroup can be tested regardless of showing the efficacy in the overall patient population. They discussed a genetic subgroup analysis on a fixed study design and an adaptive study design. Spiessens and Debois (2010) [26] applied group sequential method for interim analysis to genetic subgroup analysis so that the standard software is available to compute significance levels.

The importance of consistency within subgroups due to ethical or regulatory concerns was discussed in both Song and Chi(2007) [25] and Alish and Huque (2009) [1]. For example, imperfect classifier devices cause ethical issues and an off-label use problem would be a regulatory concern. For both papers, the consistency within subgroups is required before testing the efficacy in a pre-determined subgroup. Song and Chi (2007) proposed a statistical methodology based on the closed testing principle to strongly control FWER and considered the correlation between test statistics. Alish and Huque (2009) proposed a flexible strategy for subgroup analysis which can be viewed as an extension of a fallback procedure. Both methods provided the flexibility to test subgroups based on the satisfaction of consistency requirements even when broad efficacy is not established.

2.3.2 Definition of parameters and setting of hypotheses

Among most references ([32], [1], [25],[26]) in subgroup analysis, the parameters’ definitions and hypotheses settings are very similar. The two interesting null hypotheses were formulated as $\bar{H}_0 : \Delta \leq 0$ and $H_{0+} : \Delta_{g+} \leq 0$ where $\Delta = \frac{\mu^{Rx} - \mu^{C}}{\sigma}$ and $\Delta_{g+} = \frac{\mu^{Rx}_{g+} - \mu^{C}_{g+}}{\sigma_{g+}}$ are the standardized treatment effects. $\mu^{Rx}$ and $\mu^{C}$ are the true mean responses for treatment and control groups. The authors usually assumed
the sample sizes are the same in the treatment and control arms; furthermore, they
assumed equal sample sizes for the $g^+$ subgroup in each treatment and control arm.
$\bar{Z}$, $Z_+$ and $Z_-$ are the test statistics for testing broad efficacy, and efficacy in $g^+$ and
$g^-$ subgroups, respectively. With these assumptions, $\bar{Z}$ is a weighted average which
is weighted by the sample proportion in each subgroup.

$$\bar{Z} = \sqrt{f} Z_+ + \sqrt{1-f} Z_-$$

A bivariate normal with zero mean vector and correlation $\sqrt{f}$ is assumed for the
distribution of test statistics for broad efficacy and efficacy in $g^+$ subgroup.

$$\left( \begin{array}{c} \bar{Z} \\ Z_+ \end{array} \right) \sim N \left( \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} 1 & \sqrt{f} \\ \sqrt{f} & 1 \end{array} \right) \right)$$

$f$ is the sample size in the $g^+$ subgroup divided by the total sample size. The average
weighted by the sample proportion may cause a paradox, especially in enriched de-
signs. The concern of paradox for average definition will be discussed in section 2.4.1
and 2.4.3.

2.3.3 Methodology of genetic subgroup analysis

A common used conservative Bonferroni approach in subgroup analyses is to allo-
cate Type I error rate, $\alpha$, for testing broad efficacy and efficacy in a pre-determined
subgroup simultaneously requiring the sum of $\alpha$ at a fixed level. More efficient meth-
ods were proposed to improve the Bonferroni approach by taking the correlation
between test statistics into account. In addition, consistency was required in some
literatures to enhance the chance to test for a pre-determined subgroup even if broad
efficacy is insufficiently significant.
Alpha-split method: correlation incorporated

Wang, O’Neil and Hung (2007) proposed an alpha-split method to test $H_0$ and $H_{0+}$ simultaneously. FWER is controlled by:

$$\alpha = P\{\tilde{Z} > z_\bar{\alpha} \text{ or } Z_+ > z_{\alpha_+}|H_0\}$$

$$= 1 - P\{\tilde{Z} \leq z_\bar{\alpha} \text{ and } Z_+ \leq z_{\alpha_+}|H_0\}, \quad (2.1)$$

where $H_0 = \bar{H}_0 \cap H_{0+} = \{\Delta = 0 \cap \Delta_{g+} = 0\}$. $\bar{\alpha}$ and $\alpha_+$ are the $\alpha$ allocated to $\bar{H}_0$ and $H_{0+}$ respectively. Since correlation $\sqrt{J}$ is taken into account, it is possible that $\bar{\alpha} + \alpha_+ > \alpha$. This alpha-split method can be applied to a fixed study design and or an adaptive study design.

- **Fixed study design with no adaptation:** For a fixed study without any interim look, $\bar{H}_0$ and $H_{0+}$ are tested by the alpha-split method in (2.1) to control FWER.

- **Adaptive Study Design:** An adaptive design provides the flexibility to take an interim look at Stage I and decides whether to include $g^-$ patients after Stage I.

**Stage I:** Taking an interim analysis at time $t$, a futility assessment for the $g^-$ subgroup is conducted to decide whether the study should continue recruiting $g^-$ patients.

$$H_0^I = H_{0-}$$

if $Z_- \leq b_t$ or $Z_- > b_t^{safety}$, then exclude $g^-$ patients,

where $b_t$ is an efficacy boundary and $b_t^{safety}$ is a safety margin
In Stage I, only $tN$ patients are recruited. If $g^-$ patients trigger any safety concern or the treatment effect is small, then $g^-$ patients will be excluded in Stage II. The safety assessment is generally empirically based.

**Stage II:** Final analysis at the end of study

$$H_0^{II} = \begin{cases} H_{0+} & \text{if } Z_- \leq b_t \text{ or } Z_- > b_t^{safety} \\ \bar{H}_0 \text{ or } \{H_0 \text{ and } H_{0+}\} & \text{o.w.} \end{cases}$$

In Stage II, the total sample size is kept as the originally planned $N$. If only $g^+$ patients are included in Stage II, the test statistic becomes $Z_{++}^{AD} = \sqrt{f_1} Z_+^I + \sqrt{1 - f_1} Z_+^{II}$ and $f_1$ is the sample size proportion in stage I over the whole study period for the $g^+$ subgroup only. At the end of the study, $H_{0+}$ and/or $\bar{H}_0$ are tested as if there was no interim analysis.

Spiessens and Debois (2010) showed the similarity between the alpha-split method and group sequential methods by taking $Z_+$ as the test statistic at the interim analysis and $\bar{Z}$ as the test statistic at the final analysis. Therefore, the standard software for group sequential designs like ADDPLAN and EAST can be used to compute the significance levels.

In this alpha-split method, correlation $\sqrt{f}$ is taken into account and $f$ is the proportion of $g^+$ patients in the study. Therefore, the $\alpha_+$ increases at a fixed $\bar{\alpha}$ level when the study enrolls more $g^+$ patients. This is different from our methodology in which our correlation is related to population prevalence $\gamma$ in the $g^+$ subgroup so that enrolling more $g^+$ patients should not change the significance level of $\alpha_+$ for testing the $g^+$ subgroup. Increasing $\alpha_+$ due to a sample size increase is not reasonable, since the increased sample size has contributed to the reduced variance of the efficacy in the $g^+$ subgroup. This will be discussed later in section 2.4.1.
Hochberg's step-up method is recommended by Wang, O’Neil and Hung (2007) for the multiplicity adjustment. Therefore, we will compare our methodology with Hochberg’s step-up method in our illustrative examples.

**Alpha-split method: consistency and correlation incorporated**

An interaction test is sometimes required for testing consistency within genetic subgroups. One decides to test broad efficacy ($\bar{H}_0$) only if there is no significant evidence of interaction, or to test efficacy in each genetic subgroups separately ($H_{0+}$ and $H_{0-}$) if an interaction exists. Wang, O’Neil and Hung (2007) stated that an interaction test can not strongly control FWER since failure to reject interaction does not mean there is no interaction.

The consistency issue was considered in both Song and Chi (2007) and Alos and Huque (2009) by requiring a certain level of consistency, say $\alpha^*_1$, between the pre-determined subgroup and the overall population before testing the efficacy in the pre-determined subgroup. In other words, when consistency requirements are met and a subgroup claim is the most desired, their methods provide more alpha for subgroup testing. Therefore, subgroup efficacy can be claimed without claiming broad efficacy in advance. The more general definition of consistency will be discussed in section 2.3.4.

- **Two-Stage Testing Procedure:**

Song and Chi(2007) proposed a two-stage testing procedure based on the closed testing principle to strongly control FWER and considered the correlation between parameters. In their methodology, a conditional error function $A(\bar{z})$ is
proposed to consider certain efficacy consistency requirements in subgroup analysis.

The two-stage testing procedure is based on the closed testing principle. As we discussed in 1.4.1, testing for $\bar{H}_0$ and $H_{0+}$ is actually testing $\bar{H}^C_+, \bar{H}^C_0$ and $H^C_{0+}$ by the closed testing principle.

<table>
<thead>
<tr>
<th>Stage I</th>
<th>$\alpha_1 &lt; \alpha &lt; \alpha^*_1 \leq 1$</th>
</tr>
</thead>
</table>
| $\bar{H}^C_0 = \bar{H}_0 \cap H_{0+}$ | \begin{align*}
\text{if } \bar{p} \leq \alpha_1 & \quad \text{Reject } \bar{H}^C_+ \\
\text{if } \bar{p} > \alpha^*_1 & \quad \text{Do not Reject } \bar{H}^C_+ \\
\text{if } \alpha_1 < \bar{p} \leq \alpha^*_1 & \quad \Rightarrow \text{Test } p_+ \\
\{ & \\
\text{if } p_+ \leq \alpha_2 & \quad \text{Reject } \bar{H}^C_+ \\
\text{if } p_+ > \alpha_2 & \quad \text{Do not Reject } \bar{H}^C_+
\end{align*} |

<table>
<thead>
<tr>
<th>Stage II</th>
<th>$\bar{H}^C_0 = \bar{H}_0$</th>
<th>If $\bar{p} \leq \alpha$, reject $\bar{H}^C_+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0+} = H_{0+}$</td>
<td>If $p_+ \leq \alpha$, reject $H^C_+$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Two-Stage Testing Procedure in Song and Chi (2007).

The two-stage testing procedure is described in Table 2.1. $\bar{p}, p_+, p_-$ are the nominal one-sided p-values from $\bar{Z}, Z_+, Z_-$ respectively. To strongly control FWER, the Type I error rate in three closed testing hypotheses needs to be controlled at $\alpha$. For testing $\bar{H}^C_{0+}$, the error rate is controlled by (2.2):

$$\alpha = \alpha_1 + \int_{z_{\alpha_1}}^{z_{\alpha^*_1}} P(Z_+ > z_{\alpha_2}|\bar{Z} = \bar{z}) \phi(\bar{z}) d\bar{z} \quad (2.2)$$

$$= \alpha_1 + \int_{z_{\alpha^*_1}}^{z_{\alpha_1}} A(\bar{z}) \phi(\bar{z}) d\bar{z} \quad (2.3)$$

The authors noted that $\alpha^*_1$ is chosen to represent a degree of efficacy consistency in the overall population that is required for the subgroup test.
• **Flexible Testing Strategy:**

Alosh and Huque (2009) proposed a flexible strategy for subgroup analysis. Their testing procedure can be viewed as an extension of a fallback procedure. It means that a subgroup can be tested even when broad efficacy can not be established. The consistency within subgroups is required to test for the pre-determined subgroup. Their methodology also took the correlation between test statistics into account.

Table 2.2 describes the Flexible Testing Strategy. \(\bar{H}_0\) is tested at \(\alpha_1^*\) level and stop if \(\bar{p} > \alpha_1^*\). Similar to Song and Chi (2007), \(\alpha_1^*\) is chosen to represent a degree of efficacy consistency in the overall population required for the subgroup test. If the conditional requirement is met, \(\alpha_1 < \bar{p} < \alpha_1^*\), then test \(H_{0+}\) at \(\alpha_2\) level, which is computed by taking the correlation of the statistics into account. If \(\bar{H}_0\) is rejected at \(\alpha_1\) level, \(\bar{p} < \alpha_1\), then test \(H_{0+}\) at the full \(\alpha\) level. The last step has the flavor of a fallback procedure.

The flexible testing strategy is actually very similar to the two-stage testing procedure and can be performed as a closed testing procedure (Table 2.3). The only difference between the flexible testing strategy and the two-stage testing procedure is \(\star\) in Table 2.3. The flexible testing strategy tests \(H_{0}^C\) at an \(\alpha_1\) level instead of an \(\alpha\) level in the two-stage testing procedure.

**Examples**

An example and plots in Figure 2.1 are used to illustrate the difference between the methods discussed in section 2.3.3.
Step 1: \( \alpha_1 < \alpha < \alpha_1^* \leq 1 \)

\begin{align*}
\text{if } \bar{p} > \alpha_1^* & \quad \text{Do not Reject } \bar{H}_0 \quad \text{(stop)} \\
\bar{H}_0 \{ \text{if } \alpha_1 < \bar{p} < \alpha_1^* \} & \quad \text{Do not Reject } \bar{H}_0 \quad \Rightarrow \\
\text{if } \bar{p} \leq \alpha_1 & \quad \text{Reject } \bar{H}_0 \quad \Rightarrow
\end{align*}

Step 2

\[ \bar{H}_+ \{ \text{if } p_+ < \alpha_2 \} \quad \text{Reject } H_{0+} \]

Table 2.2: Flexible Testing Strategy in Alish and Huque (2009).

\[ \alpha_1 < \alpha < \alpha_1^* \leq 1 \]

\[ \bar{H}_{0+}^C = \bar{H}_0 \cap H_{0+} \]

\[ \begin{array}{c}
\text{if } \bar{p} \leq \alpha_1 & \quad \text{Reject } \bar{H}_+^C \\
\text{if } \bar{p} > \alpha_1^* & \quad \text{Do not Reject } \bar{H}_+^C \\
\{ \text{if } \alpha_1 < \bar{p} < \alpha_1^* \} & \quad \Rightarrow \text{Test } p_+ \\
\{ \text{if } p_+ \leq \alpha_2 & \quad \text{Reject } \bar{H}_+^C \\
\text{if } p_+ > \alpha_2 & \quad \text{Do not Reject } \bar{H}_+^C
\end{array} \]

\[ H_{0+}^C = H_0 \]

\[ \bar{H}_{0+}^C = H_{0+} \]

If \( \bar{p} \leq \alpha_1 \), reject \( H_{0+}^C \) (\(*\))

If \( p_+ \leq \alpha \), reject \( H_+^C \)

Table 2.3: Flexible Testing Strategy in Alish and Huque (2009) illustrated by closed testing principle.

Let \( \alpha = 0.025 \) and set \( \alpha_1 = 0.02 \) to test \( \bar{H}_0 \). For the consistency requirement, set \( \alpha_1^* = 0.1 \) to represent a degree of efficacy consistency between the overall population and the pre-determined subgroup.

Plot I in Figure 2.1 illustrates the alpha-split method without considering consistency and Plots II and III display the two-stage testing procedure and flexible testing strategy, respectively. In Figure 2.1,

1. The \(*\) and shaded triangle area in plot I is the region where the consistency requirement is not met. Therefore, both the two-stage testing procedure and
the flexible testing strategy do not test this region to increase $\alpha_2$ for testing $H_{0+}$, which is the dark shaded region in plot I.

2. The two-stage testing procedure and the flexible testing strategy are basically the same. The only difference is that $\star$ and the dark shaded region in plot II shows rejecting both $H_0$ and $H_{0+}$ while only $H_{0+}$ is rejected in this region by the flexible testing strategy.

3. This type of consistency only ensures consistency between the overall population and the pre-determined subgroup. However, efficacy between subgroups could be very inconsistent. Different types of consistency requirements will be discussed in section 2.3.4.

2.3.4 Consistency requirements

Three types of consistency requirements are discussed in Song and Chi (2007). The example in section 2.3.3 and plots in Figure 2.1 are continuously used to illustrate different types of consistency requirements.

- Consistency of the overall test

A non-inferiority for broad efficacy is required before the subgroup analysis is conducted. This is the type of consistency requirement that we discussed in the previous section (2.3.3), which is plot II and III in Figure 2.1. The consistency requirement is,

$$\bar{Z} > \bar{z}_{\alpha^*}$$

- Consistency of the quantitative interaction between the two subgroups

The test statistic for quantitative interaction is defined as the difference of
Figure 2.1: I: Alpha-split method, II: Two stage testing procedure (IIa: quantitative consistency, IIb: qualitative consistency), III: Flexible testing strategy

weighted efficacy between two subgroups.

\[ Z_d = \sqrt{1 - \bar{f}Z_+} - \sqrt{\bar{f}Z_-} \]
The requirement for quantitative interaction means that a test for the predetermined subgroup can be conducted only when there is no significant quantitative interaction between subgroups. That is, if

\[ Z_d \leq z_{\alpha_d} \text{ where } \alpha_d \in [0,1] \]

Let the quantitative consistency requirement be set by \( Z_d \leq z_{0.025} \). Plot IIa in Figure 2.1 demonstrates the quantitative consistency requirement in the two-stage testing procedure. When only \( H_{0+} \) is rejected, we know the quantitative difference of efficacy between \( g^+ \) and \( g^- \) is controlled at a certain level.

- Consistency of the qualitative interaction between the two subgroups

Qualitative interaction means that the efficacy in \( g^- \) is no worse than a pre-specified boundary when the test for \( g^+ \) is conducted. That is,

\[ Z_- \geq z_{\alpha_-} \text{ where } \alpha_- \in [0,1] \iff Z_d \leq \frac{\sqrt{1 - f\bar{z}} - z_{\alpha_-}}{\sqrt{f}} \]

Let the qualitative consistency requirement be set by \( Z_- \geq 0 \). Plot IIb in Figure 2.1 demonstrates the qualitative consistency requirement in the two-stage testing procedure. When only \( H_{0+} \) is rejected, we know the efficacy in \( g^- \) subgroup is positive.

Different from discussing the consistency requirements, we propose to resolve the consistency issue by providing a simultaneous \((1 - \alpha)\) confidence set for the average, \( g^+ \) and \( g^- \) subgroups. Here are the reasons why a confidence set is favored:

1. No matter which type of consistency requirement is of concern, the inconsistency between subgroups could happen in the * region of plot IIa and IIb in Figure 2.1. That is, when both \( \tilde{H}_0 \) and \( H_+ \) are rejected, the efficacy of \( g^- \) subgroup
could be much worse, so that the drug should not be approved for the overall patient population even if \( \bar{H}_0 \) is rejected.

2. No matter if qualitative or quantitative interactions are concerned, interactions not a good way to define consistency between subgroups. Consistency between subgroups should be an absolute efficacy between treatment and control arms, not a relative efficacy between subgroups. That is, it is possible that there is a significant treatment effect in either subgroup regardless of rejecting or not rejecting the null hypotheses of no quantitative or qualitative interactions.

3. A simultaneous \((1 - \alpha)\) confidence set for the average, \(g^+\) and \(g^-\) subgroups is a simple, direct and transparent way to solve the consistency issue. We will have further discussion in section 3.3.

2.3.5 Power computation

The conventional definition of power is computed among all references. However, the conventional definition only considers marginal power for broad efficacy and efficacy in the \(g^+\) subgroup, like equations (2.4) and (2.5).

\[
\text{power for broad efficacy} = P(\bar{\rho} \leq \alpha_1|\Delta) \tag{2.4}
\]

\[
\text{power for efficacy in the } g^+ \text{ subgroup} = P(p_+ \leq \alpha_2|\Delta_{g^+}) \tag{2.5}
\]

We will propose another version of power in section 3.2.5 to give a probability of correct and useful inference. Furthermore, this version of power is easy to convert to probability of a successful trial, which is useful information for sponsors to evaluate the chance for approving their products for marketing.
2.4 New formulation

The efficacy of a drug is typically defined in terms of its average effect, usually compared to a control group, possibly measured relative to baseline measurements.

Consider a single biomarker which classifies patients into a \( g^+ \) subgroup and its complementary subgroup \( g^- \).

Let \( Rx \) and \( C \) denote “treatment” and “control” respectively. Let \( \mu^{Rx} \) and \( \mu^C \) denote the arm-specific mean responses over the entire patient population. Let \( \mu^{Rx}_{g^+}, \mu^C_{g^+}, \mu^{Rx}_{g^-}, \mu^C_{g^-} \) denote the arm-specific mean responses in the \( g^+ \) and \( g^- \) subgroups. If a higher mean response is better, and prevalence of the \( g^+ \) subgroup is \( \gamma \) (for both \( Rx \) and \( C \) patients), then

\[
\begin{align*}
\theta_{g^+} & = \mu^{Rx}_{g^+} - \mu^C_{g^+} \\
\theta_{g^-} & = \mu^{Rx}_{g^-} - \mu^C_{g^-} \\
\bar{\theta} & = \gamma \times \theta_{g^+} + (1 - \gamma) \times \theta_{g^-}
\end{align*}
\]

(2.6)

represent effect of the drug in the \( g^+ \) and \( g^- \) subgroups, and over the entire patient population, respectively. (If a lower mean response is better, then \( \theta_{g^+} = \mu^C_{g^+} - \mu^{Rx}_{g^+}, \theta_{g^-} = \mu^C_{g^-} - \mu^{Rx}_{g^-} \))

A treatment is efficacious for a group of patients or a subgroup of patients if its effect is greater than \( \delta \). The quantity, \( \delta \), representing a clinically meaningful difference, can be a non-negative quantity for a superiority study, or a negative quantity for a non-inferiority study. Prevalence, \( \gamma \), can be obtained from the National Center for Biotechnology Information (NCBI).
2.4.1 Defining average for a difference of means

One logical issue in subgroup efficacy testing is the logical relationships among parameters representing efficacy. Specifically, if a drug is efficacious averaged over the entire patient population (as defined in equation (2.6)), then it must be efficacious in at least one of the \( g^+ \) and \( g^- \) subgroups. Similarly, if a drug is efficacious in both the \( g^+ \) and \( g^- \) subgroups, then it must be efficacious averaged over the entire patient population.

The way we defined the average efficacy in equation (2.6)) avoids amalgamation paradox like Simpson’s Paradox. In our research, we defined the drug effect over an entire patient population as \( \bar{\theta} = \gamma \theta_{g^+} + (1 - \gamma) \theta_{g^-} \), which is an average weighted by the prevalence of each genetic subgroup. But, some studies, for example, Wang, Hung, O’Neil (2007) and Spiessens and Debois (2010) instead defined average efficacy weighted by the sample size proportion in each subgroup instead.

The difference between two different average definitions is evident in their formulae. Let \( N^{Rx} \) and \( N^C \) be the patient sample sizes in the treatment and control arms respectively \( (N^{Rx} + N^C = N) \), and \( N^{Rx}_+ \) and \( N^C_+ \) be the number of patients with a \( g^+ \) biomarker in each arm \( (N^{Rx}_+ + N^C_+ = N_+) \). Similar notation is used for the \( g^- \) subgroup. \( \bar{\theta}_n \) represents the average weighted by the sample size and \( \bar{\theta}_\gamma \) is the average.
weighted by prevalence.

\[
\bar{\theta}_n = \Delta = \mu_{Rx} - \mu_{C} \text{( assume } \sigma = 1) \\
= \frac{N^R_x \mu_{g+} + N^R_x \mu_{g-}}{N^{R_x}} - \frac{N^C \mu_{g+} + N^C \mu_{g-}}{N^{C}} \\
= \left[ \frac{N^R_x}{N^{R_x}} \mu_{g+} - \frac{N^C}{N^{C}} \mu_{g+} \right] + \left[ \frac{N^R_x}{N^{R_x}} \mu_{g-} - \frac{N^C}{N^{C}} \mu_{g-} \right] \\
= \frac{N^R_x}{N} \theta_{g+} + (1 - \frac{N^R_x}{N}) \theta_{g-} \text{ (if } \frac{N^R_x}{N^{R_x}} = \frac{N^C}{N^{C}} = \frac{N^R_x}{N}) \\
\bar{\theta}_\gamma = \gamma \theta_{g+} + (1 - \gamma) \theta_{g-}
\]

From equation (2.9), it is clear that \( \bar{\theta}_\gamma > 0 \) if \( \theta_{g+} > 0 \) and \( \theta_{g-} > 0 \). On the contrary, \( \bar{\theta}_n \) is not necessarily greater than zero even if \( \theta_{g+} > 0 \) and \( \theta_{g-} > 0 \). This is called Simpson’s Paradox. From equation (2.8), if the study satisfies the condition \( \frac{N^R_x}{N^{R_x}} = \frac{N^C}{N^{C}} = \frac{N^R_x}{N} = \gamma \) then \( \bar{\theta}_n \) becomes an average weighted by the sample proportions of two genetic subgroups. In addition, if the sample proportion is a good estimator of population prevalence, then \( \bar{\theta}_n \) is approximately equal to \( \bar{\theta}_\gamma \). In the clinical data, this condition might be achieved by randomizing patients into treatment and control arms from a random sample in a patient population.

**A toy example**

Table 2.4 provides a toy example which gives opposite directions of overall efficacy by using different definitions of average. \( \bar{\theta}_n < 0 \) shows that the control arm is better; while \( \bar{\theta}_\gamma > 0 \) exhibits a treatment effect that is larger in the treatment arm. Therefore, we claim that weighting the average by population prevalence is an accurate way to
<table>
<thead>
<tr>
<th>Genetic subgroup</th>
<th>$g^+$</th>
<th>$g^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment ($R_{rx}$)</td>
<td>3.5, 3.5, 3.5 (ave. = 3.5)</td>
<td>7.5, 7.5 (ave. = 7.5)</td>
</tr>
<tr>
<td>Control ($C$)</td>
<td>5, 6 (ave. = 5.5)</td>
<td>4, 5, 6 (ave. = 5)</td>
</tr>
</tbody>
</table>

Table 2.4: A toy example showing difference between $\bar{\theta}_n$ and $\bar{\theta}_\gamma$

adjust the overall average for the genetic subgroup effect.

\[
\hat{\bar{\theta}}_n = \frac{3.5 + 3.5 + 3.5 + 7.5 + 7.5}{5} - \frac{5 + 6 + 4 + 5 + 6}{5} = 5.1 - 5.2 = -0.1 < 0
\]

\[
\hat{\bar{\theta}}_\gamma = \frac{3.5 + 7.5}{2} - \frac{5.5 + 5}{2} = 5.5 - 5.25 = 0.25 > 0
\]

This use of the word *logic* is in the same sense as Shaffer (1986), Holland and Copenhaver (1987), and Westfall and Tobias (2007) in their multiple testing of equality among pairs of three or more treatments. Sections 3.2 and 4.1 will show the importance of recognizing logical relationships among parameters.

Note that the concept of average efficacy, averaged across different subgroups, applies when efficacy is defined as the difference of expectation of treatment and control effects. This is because, when prevalences are taken into account, the average of the expected differences in the subgroups equals the expected broad population difference.

When efficacy is defined in terms of a nonlinear function such as the ratio of hazard rates, or odds ratio, defining "average" efficacy across subgroups becomes treacherous. Assessing consistency of efficacy in subgroups relative to "average" efficacy is difficult in such situations.
2.4.2 Distribution of standardized statistics

Two settings to which our formulation applies are:

1. Continuous response modeled by a General Linear Model (GLM)
2. Binary response modeled by binomial distribution

For a continuous response, predictors in the GLM include indicators for treatment vs. control, subgroup \( g^+ \) vs. \( g^- \), their interaction, and possibly other covariates and blocking factors. Estimates of \( \theta_{g^+}, \theta_{g^-} \), and \( \bar{\theta} \) appropriate for our formulation are computed from Least Squares means (LSmeans in SAS Proc GLM), not marginal means (Means in SAS Proc GLM). The estimate for \( \bar{\theta} \) obtained by fitting a (marginal) model without the “subgroup” and “interaction” predictors is in general inappropriate, because it is affected by the potential artifact of unequal sample size patterns across \( R_x \) and \( C \), between the \( g^+ \) and \( g^- \) subgroups. A model without a subgroup term is an "unconditional" model while a model with a subgroup term is a "conditional" model. We disagree about the "average" interpretation of a model without the subgroup term.

Under the assumption of i.i.d. normally distributed errors, the estimates \( \hat{\theta}_{g^+}, \hat{\theta}_{g^-}, \) and \( \hat{\theta} = \gamma \hat{\theta}_{g^+} + (1 - \gamma) \hat{\theta}_{g^-} \) have a multivariate normal distribution with means \( \theta_{g^+}, \theta_{g^-}, \) and \( \bar{\theta} \). Under this (homoscedastic error) assumption, estimates of their standard errors, \( \hat{\sigma}_{\theta_{g^+}}, \hat{\sigma}_{\theta_{g^-}}, \) and \( \hat{\sigma}_{\bar{\theta}} \), are multiples of \( \sqrt{MSE} \), so the studentized statistics

\[
T_+ = \frac{(\hat{\theta}_{g^+} - \theta_{g^+})}{\hat{\sigma}_{\theta_{g^+}}}
\]

\[
T_- = \frac{(\hat{\theta}_{g^-} - \theta_{g^-})}{\hat{\sigma}_{\theta_{g^-}}}
\]

\[
T = \frac{(\hat{\theta} - \bar{\theta})}{\hat{\sigma}_{\bar{\theta}}}
\]
have a multivariate \( t \) distribution (which can be computed by Monte Carlo simulation).

For binary response, we assume the sample sizes are large enough so that \( \hat{\theta}_{g^+} \) and \( \hat{\theta}_{g^-} \) are approximately (independently) normally distributed with means \( \theta_{g^+} \) and \( \theta_{g^-} \). Let \( \hat{\sigma}_{g^+} \) and \( \hat{\sigma}_{g^-} \) be consistent estimates of their standard errors. Define

\[
\hat{\theta} = \gamma \hat{\theta}_{g^+} + (1 - \gamma) \hat{\theta}_{g^-} \quad \text{and} \quad \hat{\sigma}_\theta = \gamma^2 \hat{\sigma}_{g^+}^2 + (1 - \gamma)^2 \hat{\sigma}_{g^-}^2.
\]

Then

\[
T_+ = \frac{(\hat{\theta}_{g^+} - \theta_{g^+})}{\hat{\sigma}_{g^+}}
\]

\[
T_- = \frac{(\hat{\theta}_{g^-} - \theta_{g^-})}{\hat{\sigma}_{g^-}}
\]

\[
\bar{T} = \frac{(\hat{\theta} - \bar{\theta})}{\hat{\sigma}_\theta}
\]

have approximately a multivariate normal distribution (which can be computed by Monte Carlo simulation).

2.4.3 Measure of efficacy and design of study

The efficacy measure of a drug depends on the disease it is intended to treat. The primary outcome of a treatment for schizophrenia is often decrease in PANSS from a baseline, and the efficacy of a drug is typically measured as the difference in the expected decrease, relative to a control treatment. The primary outcome of a treatment for cardiovascular disease is often time to any major vascular event, and the efficacy of a drug is either measured as the ratio of the risk of such an event, or the hazard rate of such an event, relative to a control treatment. In safety studies, whether a genetic mutation is associated with patients experiencing an adverse event to a drug may be measured by the ratio of the odds of patients experiencing an adverse event having this mutation, relative to patients not experiencing an adverse event having this mutation.
When efficacy is defined as the *difference* of expectation of treatment and control effects, with prevalence defined as the *population* prevalence, our definition of average efficacy is not subject to Simpson’s paradox. This is because the average of the expected difference in subgroups equals the expected difference in the broad population. However, more care needs to be taken in defining "average efficacy" when efficacy is defined in terms of a relative risk, a ratio of hazard rates, or an odds ratio.

We consider two situations that relate to the average definition:

1. **Paradox-free**: We say that a measure of efficacy is *paradox-free* if

   \[
   \max_i \theta_i \neq \bar{\theta} \text{ and } \min_i \theta_i \neq \bar{\theta}
   \]  

   (2.10)

   Here note that paradox is defined for the parameters not for statistics. Paradox-free efficacy in the parameters can be achieved by a clinical design, and we will discuss how to avoid paradox by design in this section.

2. **Mixture-representable**: Even though paradox-free efficacy in the parameters can be achieved by a clinical design, there is still a possibility that the estimates from a realized experiment can suffer from paradox. Efficacy must be *mixture-representable* is required to accurately represent efficacy for the overall patient population if not enough care is taken in the design of experiment.

   We say \( \bar{\theta} \) is *mixture-representable* when broad efficacy is a function of subgroup efficacy, and known population value \( \kappa \):

   \[
   \bar{\theta} = f(\theta_+, \theta_-, \kappa)
   \]

   For example, \( \kappa \) can be the prevalence of a genetic subgroup, which we denote \( \gamma \) in our definition of average in equation (2.6).
In the following subsections, we will discuss our concerns about paradox-free and mixture-representable efficacy for the relative risk, hazard ratio, and odds ratio, respectively.

Paradox-free sample estimators can be achieved by clinical design. We introduce three conditions in the design of an experiment. (Good and Mittal (1987)) . Let \( a_+ \) and \( b_+ \) indicate the number of responders and non-responders in the treatment (\( R_x \)) arm and \( c_+ \) and \( d_+ \) indicate the number of responders and non-responders for the control (\( C \)) arm in the \( g_+ \) subgroup. Analogical notations for the \( g_- \) subgroup and combined group are provided in Table 2.5.

<table>
<thead>
<tr>
<th>treatment (Rx)</th>
<th>( g_+ ) subpopulation</th>
<th>( g_- ) subpopulation</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R )</td>
<td>( a_+ )</td>
<td>( a_- )</td>
<td>( R )</td>
</tr>
<tr>
<td>( NR )</td>
<td>( b_+ )</td>
<td>( b_- )</td>
<td>( NR )</td>
</tr>
<tr>
<td></td>
<td>( N_{Rx}^+ )</td>
<td>( N_{Rx}^- )</td>
<td>( N_{Rx} )</td>
</tr>
<tr>
<td>( c_+ )</td>
<td>( c_- )</td>
<td>( c_+ )</td>
<td>( c_- )</td>
</tr>
<tr>
<td>( d_+ )</td>
<td>( d_- )</td>
<td>( d_+ )</td>
<td>( d_- )</td>
</tr>
<tr>
<td>( R_+ )</td>
<td>( N_{Rx}^+ )</td>
<td>( N_{Rx}^- )</td>
<td>( R_{Rx} )</td>
</tr>
<tr>
<td>( NR_+ )</td>
<td>( N_{Rx}^+ )</td>
<td>( N_{Rx}^- )</td>
<td>( NR_{Rx} )</td>
</tr>
<tr>
<td>( R_- )</td>
<td>( N_{Rx}^+ )</td>
<td>( N_{Rx}^- )</td>
<td>( N_{Rx} )</td>
</tr>
<tr>
<td>( NR_- )</td>
<td>( N_{Rx}^+ )</td>
<td>( N_{Rx}^- )</td>
<td>( NR_{Rx} )</td>
</tr>
<tr>
<td>( N_+ )</td>
<td>( N_{Rx}^+ )</td>
<td>( N_{Rx}^- )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

Table 2.5: responders (R) and Non-responders (NR) for each genetic subgroup in the treatment and control arms and the combined group.

- **row-uniform**: An experimental design is said to be row-uniform for \( g^+ \) and \( g^- \) subgroups if, for some \( \lambda \),

\[
\frac{N_{Rx}^+}{N_{Rx}^-} = \frac{N_{Rx}^+}{N_{Rx}^-} = \lambda
\]  

(2.11)

A *row-uniform* design is a design in which the ratio of the number of patients in the treatment to the number of patients in the control is constant in each genetic subgroup and in the combined group.
• **column-uniform**: An experimental design is said to be column-uniform for \( g^+ \) and \( g^- \) subgroups if, for some \( \eta \),

\[
\frac{R_+}{N_{R+}} = \frac{R_-}{N_{R-}} = \eta
\]  

(2.12)

A *column-uniform* design is a design in which the ratio of the number of responders to the number of non-responders is constant in each genetic subgroup and in the combined group.

• **row-column-uniform**: An experimental design is said to be row-column-uniform for \( g^+ \) and \( g^- \) subgroups if both conditions, (2.11) and (2.12), are satisfied. The famous row-column-uniform design is Fisher’s lady tasting tea experiment.

**Definition of average relative risk (RR)**

A Relative Risk is a ratio of two probabilities and it is susceptible to Simpson’s paradox, unless the design is *row-uniform*.

Let \( RR_+ \), \( RR_- \) and \( \overline{RR} \) indicate RR for the \( g_+ \), \( g_- \) subgroups and for the broad population, respectively.

1. RR is paradox-free if the design is row-uniform.

2. \( \overline{RR} \) is mixture-representable, if the design is a row-uniform design. The average of RR can be defined as \( \overline{RR} = q * RR_+ + (1 - q) * RR_- \). (Theorem 4.2 in Good and Mittal (1987) .)
< Proof >

If the design is row-uniform

\[
\frac{N^{Rx}_+}{N^C_+} = \frac{N^{Rx}_-}{N^C_-} = \lambda
\]

\[
\hat{RR}_+ = \frac{a_+}{N^{Rx}_+} \frac{c_+}{N^C_+} = \frac{1}{\lambda} \frac{a_+}{c_+}
\]

\[
\hat{RR}_- = \frac{a_-}{N^{Rx}_-} \frac{c_-}{N^C_-} = \frac{1}{\lambda} \frac{a_-}{c_-}
\]

\[
\hat{RR} = \frac{1}{\lambda} \frac{a_+ + a_-}{c_+ + c_-}
\]

\[
= \frac{1}{\lambda} \left\{ \frac{c_+}{c_+ + c_-} \lambda \hat{RR}_+ + \frac{c_-}{c_+ + c_-} \lambda \hat{RR}_- \right\}
\]

\[
= \frac{c_+}{c_+ + c_-} \hat{RR}_+ + \frac{c_-}{c_+ + c_-} \hat{RR}_-
\]

\[
= \hat{q} \hat{RR}_+ + (1 - \hat{q}) \hat{RR}_-
\]

where \( \hat{q} = \frac{c_+}{c_+ + c_-} \), the proportion of \( g^+ \) patients among all responders in control group. In the population, the average of RR can be defined as \( \bar{RR} = q \cdot RR_+ + (1 - q) \cdot RR_- \) where \( q = P(g^+ | C, R) \).

Three clinical designs are suggested to achieve row-uniformity.

- **R-UI Design I**: A randomized design (Figure 2.2). A randomized design ensures the treatment and control groups have similar characteristics by making \( \frac{N^{Rx}_+}{N^{Rx}_-} = \frac{N^C_+}{N^C_-} \). This actually implies row-uniformity, \( \frac{N^{Rx}_+}{N^C_+} = \frac{N^{Rx}_-}{N^C_-} = \lambda \). In practice, there is still a possibility that the estimates from a realized experiment can suffer from paradox, if the randomization does not work well. Therefore, the R-UI Design II and R-UI Design III are suggested to ensure row-uniformity.

- **R-UI Design II**: R-UI Design II (Figure 2.3) stratifies the study on genotype and fixes the numbers of patients in treatment and control to achieve (2.11),

53
randomizing the assignment of patients to treatment and control within the $g^+$ patients, and within the $g^-$ patients.

- **R-UI Design III**: If patients are randomly assigned to treatment and control without stratification on genotype, Figure 2.4 provides strict row-uniformity by taking random samples of treated and control $g^+$ patients, and random samples of treated and control $g^-$ patients, with sample sizes that satisfy (2.11).

No matter if we have a R-UI Design I,II or III, we need to make sure the proportion of $g^+$ patients in the responders of the control arm ($\hat{q} = \frac{c_+}{c_+ + c_-}$) can represent $q = P(g^+|C,R)$ in the population by design. Otherwise, the estimates of the average still cannot accurately represent population average. However, RR is mixture-representable if we define average as: $\overline{RR} = q \times RR_+ + (1 - q) \times RR_-$. $q$ can be
estimated by the proportion of patients with the $g^+$ biomarker out of responders in the sample if the usual cure is assumed in the control arm. For example, the control arm with a cancer drug is usually treated by usual care so that $q$ can be estimated.
by \( \hat{p}(g^+ | R) \) in the sample. However, \( \hat{q} = \hat{p}(g^+ | R) \) is not always true for other drugs if their control arm does not stand for usual care.

**Definition of average hazard ratio (HR)**

The hazard rate is a commonly used endpoint for time to event data in survival analysis. It means the instantaneous rate of death at time \( x_j \) given that an individual is alive at time \( x_j \). In the discrete data case, the hazard function can be defined as:

\[
h(x_j) = P(X = x_j | X \geq x_j) = \frac{p(x_j)}{S(X_{j-1})}, j = 1, 2, 3, ...
\]

By this definition, the hazard rate means the number of patients dead among all patients at risk at a fixed time point \( x_j \). The hazard ratio (HR) of treatment versus control behaves as the **Relative Risk (RR)**. However, the difference is that time to event data involves "event" and "time". To express the data by contingency tables, there are not only different contingency tables for different genetic subpopulations but also for different time points. In Table 2.6, \( N_{j+} \) means the number of patients at risk at time \( x_j \) in the \( g^+ \) subpopulation and \( a_{j+} \) indicates the number of patients dead at at time \( x_j \) in the \( g^+ \) subpopulation. Analogical notations for the \( g_- \) subpopulation and in the overall population are shown in Table 2.6.

1. HR is paradox-free only if there are row-uniform designs "over time".

2. \( \overline{HR} \) is mixture-representable only if (1) there are row-uniform designs "over time" and (2) \( \overline{HR}_j \) is constant "over time".

< Proof >

If row-uniformity "over time" is achieved, \( \frac{N_{j+}^{Rx}}{N_{j+}^C} = \frac{N_{j-}^{Rx}}{N_{j-}^C} = \lambda_j \) for different time points \( x_j \), then \( HR_j \) is paradox-free over time by the result shown in RR. In
Table 2.6: Contingency tables for time to event data. D means patients died and ND means patients survived.

\[
\begin{array}{c|cc|c|cc}
& g_+ & & & g_- & \\
\hline
\text{treatment (Rx)} & a_{1+} & b_{1+} & N_{1+}^{Rx} & a_{1-} & b_{1-} & N_{1-}^{Rx} \\
\text{control (C)} & c_{1+} & d_{1+} & N_{1+}^{C} & c_{1-} & d_{1-} & N_{1-}^{C} \\
R_{1+} & NR_{1+} & N_{1+} & R_{1-} & NR_{1-} & N_{1-} \\
\hline
\text{treatment (Rx)} & a_{2+} & b_{2+} & N_{1+}^{Rx} - a_{1+} = b_{1+} & a_{2-} & b_{2-} & N_{1-}^{Rx} - a_{1-} = b_{1-} \\
\text{control (C)} & c_{2+} & d_{2+} & N_{1+}^{C} - c_{1+} = d_{1+} & c_{2-} & d_{2-} & N_{1-}^{C} - c_{1-} = d_{1-} \\
R_{2+} & NR_{2+} & N_{1+} - R_{1+} & R_{2-} & NR_{2-} & N_{1-} - R_{1-} \\
\hline
\text{treatment (Rx)} & a_{j+} & b_{j+} & N_{1+}^{Rx} - \sum_{i=1}^{j-1} a_{i+} & a_{j-} & b_{j-} & N_{1-}^{Rx} - \sum_{i=1}^{j-1} a_{i-} \\
\text{control (C)} & c_{j+} & d_{j+} & N_{1+}^{C} - \sum_{i=1}^{j-1} c_{i+} & c_{j-} & d_{j-} & N_{1-}^{C} - \sum_{i=1}^{j-1} c_{i-} \\
R_{j+} & NR_{j+} & N_{1+} - \sum_{i=1}^{j-1} R_{i+} & R_{j-} & NR_{j-} & N_{1-} - \sum_{i=1}^{j-1} R_{i-} \\
\end{array}
\]

addition, the average \( \overline{HR}_j \) is a linear combination of \( HR_{j+} \) and \( HR_{j-} \).

\[
\overline{HR}_j = q_j \ast HR_{j+} + (1 - q_j) \ast HR_{j-}, \text{ for } j = 1, 2, 3...
\]

Furthermore, if condition (2) is assumed, then

\[
\overline{HR} = \overline{HR}_1 = \overline{HR}_2 = \ldots = \overline{HR}_j
\]

However, the condition (1), row-uniform design over time, is NOT a reasonable design for genetic subgroup analysis. We can take time point \( x_1 \), for example, \( \frac{N_{1+}^{Rx}}{N_{1+}^{C}} = \lambda_1 \) can be designed by randomly assigning patients into treatment and control arms. However, to ensure row-uniformity is continued to time point \( x_2 \), we need to have \( \frac{b_{1+}^{Rx}}{d_{1+}^{C}} = \frac{b_{1-}^{Rx}}{d_{1-}^{C}} = \nu_1 \) since patients at risk at time \( x_2 \) are the patients who survived.
at time $x_1$ ($N_{2+}^{Rx} = b_{1+}^{Rx}$). That is, "row-uniformity over time" actually means,

$$\frac{N_{j+}^{Rx}}{N_{j+}^{C}} = \frac{N_{j-}^{Rx}}{N_{j-}^{C}} = \lambda_j$$

and

$$\frac{b_{j+}^{Rx}}{d_{j+}^{C}} = \frac{b_{j-}^{Rx}}{d_{j-}^{C}} = \nu_j, j = 1, 2, 3..... \quad (2.13)$$

Equation (2.13) means a ratio of patients who survived in treatment vs. control should be constant in each genetic subgroup over time. This is an unreasonable design for genetic subgroup analysis since Equation (2.13) actually implies that the treatment effect is not different between $g^+$ and $g^-$ patients, which is not what we expect. In addition, since $b_j$ or $d_j$ are observed survived patients, equation (2.13) is hard to achieved by a design.

In the Cox proportional model, condition (2) is assumed. If the assumption of "constant hazard ratio over time" is true, then condition (2) combined with condition (1) shows,

$$\frac{a_{i+}/N_{i+}^{Rx}}{c_{i+}/N_{i+}^{C}} = \frac{a_{j+}/N_{j+}^{Rx}}{c_{j+}/N_{j+}^{C}}$$

$$\Rightarrow \frac{1}{\lambda_i} \frac{a_{i+}}{c_{i+}} = \frac{1}{\lambda_j} \frac{a_{j+}}{c_{j+}}$$

$$\Rightarrow \frac{a_{i+}/c_{i+}}{a_{j+}/c_{j+}} = \frac{\lambda_i}{\lambda_j} = \frac{a_{i-}/c_{i-}}{a_{j-}/c_{j-}} \quad (2.14)$$

Equation (2.14) means the ratio of the number of patients dead in the treatment arm divided by the number of patients dead in the control arm at two different time points is constant. Moreover, this ratio is the same for all subgroups. Again, designing a clinical trial to satisfy condition (1) (or (1) combined with (2)) is unreasonable for genetic subgroup studies.
Definition of average odds ratio (OR)

The odds Ratio (OR) is the common parameter of interest in a logistic regression model for binary data. Let $OR_+$, $OR_-$ and $\overline{OR}$ indicate OR for the $g_+$, $g_-$ subgroup and for broad population respectively.

1. $OR$ is only paradox-free if a design has both row and column-uniformity (Theorem 4.3 of Good and Mittal (1987)). This would only happen with very specialized designs. We will discuss a row-column-uniform design later.

2. There is NO mixture-representable $\overline{OR}$, since it cannot be formulated by a function of $OR_+$, $OR_-$ and a population value ($\kappa$).

• **RC-UI Design I:** RC-UI Design I (Figure 2.5) is to stratify the study on genotype and to fix up the 4 margins for treatment and control and also responders and non-responders in each subgroup. Then sampling with replacement until row-column-uniformity conditions (2.11) and (2.12) achieved. RC-UI Design I might be designed to avoid paradox. However, column uniformity in clinical design means biomarkers are independent of responders, $P(R|g^+) = P(R|g^-) = P(R)$. This is usually not true for genetic subgroup studies. Therefore, $\overline{OR}$ is not representable even though we can avoid paradox for OR by this design. So far, it is not clear for us how to define $\overline{OR}$ in a representable way.

To sum up this section, if efficacy is defined in terms of relative risk, then to be paradox-free, a row uniform design is required. This means the proportion of patients receiving treatment must be the same for the $g+$ subgroup and the $g-$ subgroup.
For a binary response, if efficacy is defined in terms of the odds ratio, then to be paradox-free, row-column-uniformity is required. While row-uniformity can be achieved as described above, column-uniformity is not sensible especially in a genetic subgroup study. This will lead to the fact that the average $\hat{OR}$ cannot represent the average odds ratio in a population even if paradox does not occur.

For time-to-event response, if efficacy is defined in terms of a hazard ratio, and if a Cox proportional hazard rate model is assumed, then to be paradox-free, row-uniformity at all time points is required. This means the proportion of treated patients among patients still at risk must be the same for the $g^+$ subgroup and the $g^-$ subgroup, at all time points. In some situations, this would not happen. One example where this would not happen is when the response from a control is the same for the $g^+$ and $g^-$ subgroups, but the $g^+$ subgroup derives more efficacy from the drug than the $g^-$ subgroup.
Chapter 3: Assessing broad efficacy and efficacy in two genetic subgroups

3.1 Simultaneous testing for broad efficacy and efficacy in one pre-determined subgroup

In some studies, there is an expectation that patients with a particular allele of a gene may derive more benefits from a drug, due to the gene’s involvement in the pathway that the drug targets. For example, patients with a BRCA mutation might benefit more from a PARP breast cancer drug. If $g^+$ denotes the patient subgroup with that allele, then in addition to testing for efficacy in the entire patient population, a sponsor can choose to test simultaneously for efficacy in the $g^+$ subgroup.

3.1.1 Formulating null hypotheses

Assume a treatment is efficacious for a group of patients or a subgroup of patients if its effect is greater than $\delta$. The quantity $\delta$, representing a clinically meaningful difference, can be a non-negative quantity for a superiority study, or a negative quantity for a non-inferiority study. The null hypotheses being tested are then

$$H_0 : \bar{\theta} \leq \delta \quad vs. \quad H_a : \bar{\theta} > \delta \quad (3.1)$$

$$H_{0+} : \theta_{g^+} \leq \delta \quad vs. \quad H_{a+} : \theta_{g^+} > \delta \quad (3.2)$$

61
Testing (3.1) and (3.2) simultaneously allows the possibility of getting the drug approved for \( g^+ \) patients, should the study fail to establish broad efficacy. This is the same hypothesis formulation as in Wang, O’Neill, and Hung(2007), Song and Chi (2007), and Alish and Huque (2009).

3.1.2 Decision rule: Partition testing vs. Closed testing

To test (3.1) and (3.2), Closed testing forms the following intersection null hypotheses:

\[
\bar{H}_0^C : \{ \bar{\theta} \leq \delta \cap \theta_{g+} \leq \delta \} \quad \text{vs.} \quad \bar{H}_a^C : \{ \bar{\theta} > \delta \cup \theta_{g+} > \delta \}
\]

\[
\bar{H}_0^C : \{ \bar{\theta} \leq \delta \} \quad \text{vs.} \quad \bar{H}_a^C : \{ \bar{\theta} > \delta \}
\]

\[
H_0^C : \{ \theta_{g+} \leq \delta \} \quad \text{vs.} \quad H_a^C : \{ \theta_{g+} > \delta \}
\]

while Partition testing forms the following partitioning null hypotheses

\[
\bar{H}_0^* : \{ \bar{\theta} \leq \delta \cap \theta_{g+} \leq \delta \} \quad \text{vs.} \quad \bar{H}_a^* : \{ \bar{\theta} > \delta \cup \theta_{g+} > \delta \}
\]

\[
\bar{H}_0^* : \{ \bar{\theta} \leq \delta \cap \theta_{g+} > \delta \} \quad \text{vs.} \quad \bar{H}_a^* : \{ \bar{\theta} > \delta \}
\]

\[
H_0^* : \{ \bar{\theta} > \delta \cap \theta_{g+} \leq \delta \} \quad \text{vs.} \quad H_a^* : \{ \theta_{g+} > \delta \}
\]

An original null hypothesis is rejected if all Closed testing or Partition testing hypotheses implying it are rejected. Table 3.1 displays decision rules for rejecting \( \bar{H}_0 \) and \( H_0+ \) by Closed testing and Partition testing. Provided level-\( \alpha \) tests for Closed and Partition hypotheses are chosen to be the same, for testing (3.1) and (3.2), Closed testing and Partition testing would reach the same inference. This does not hold, however, for testing the null hypotheses in section 3.2, as will be seen. Further, in practice, users tend to (implicitly) choose different tests for Partition testing and Closed testing hypotheses. Currently, one of the most popular Closed test is Hochberg’s step-up test,
<table>
<thead>
<tr>
<th></th>
<th>Reject $H_0$</th>
<th>Reject $H_{0+}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{0+}$</td>
<td>$H_{*+}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\bar{H}$</td>
<td>$\bar{H}_{*}$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$H_{C}$</td>
<td>$H_{*}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Closed testing (C) and Partition testing (P) decision rules for rejecting $\bar{H}_0$ and $H_{0+}$. A $\checkmark$ denotes a C or P null hypothesis that needs to be rejected.

As a result of (implicitly) choosing what Huang and Hsu (2007) call the Simes-Hochberg test to test each Closed testing null hypothesis. In contrast, Partition testers generally recommend choosing the maxT test or, equivalently, a minP test (both described in Appendix B), to test each Partition null hypothesis, resulting in a step-down shortcut. The advantage of a multiple test having a step-wise shortcut is computational, for situations where the number of original hypotheses is large. If the number of original hypotheses is $k$, then step-wise testing reduces the number of Closed testing or Partition testing hypotheses from (up to) $2^k$ to at most $k$. This is obviously important when $k$ is in the hundreds (DMET panel), thousands (drug discovery studies), or millions (bioinformatics). However, for the genetic subgroup testing problem of this article, the number of original hypotheses is small. So whether the multiple test has a step-wise shortcut is not very important. Of consequence, however, is that in contrast to Hochberg’s method, (a) Partition testing recognizes the logical relationships between parameters in the null hypotheses, and (b) Partition testing in the maxT (or minP) form allows critical value computation that takes the joint distribution of the test statistics into account. The difference these make is demonstrated in the next
sections. In the remainder of this article, “Partition testing” is understood to refer to Partition testing in the maxT form.

3.1.3 Inference: consistency issue and off-label use

Only testing for overall patients and for one pre-specified $g^+$ subgroup might cause a problem in inference. For example, even though the test has shown efficacy for overall patient group and for the $g^+$ subgroup, it actually does not guarantee efficacy for the complementary $g^-$ subgroup. Whether patients in the $g^-$ subgroup should take the drug or not might be questionable. There are two important issues that arise from this testing that have been discussed.

The first issue is, if broad efficacy is demonstrated and the drug is approved for use by the entire patient population, then there should be some assurance that such efficacy is not derived from extreme efficacy in one subgroup, i.e., that the drug should be sufficiently efficacious for patients in the complementary subgroup as well. (This is the “consistency” requirement in Alish and Huque 2009.)

The second issue is, if efficacy is only demonstrated in a subgroup and the drug is approved for use in that subgroup, it may be difficult to prevent its use in the complementary subgroup, due to the practice of off-label use. There should be some assurance that the drug is not potentially alarmingly lacking in efficacy for patients in the complementary subgroup. (Song and Chi 2007, p. 3536)

3.2 Simultaneous testing for broad efficacy and efficacy in complementary subgroups

Whereas Wang, O’Neill, and Hung (2007), Song and Chi (2007), and Alish and Huque (2009) test (3.1) and (3.2) only, we believe $g^-$ patients should be enrolled and
(3.3) should also be tested, and we present a third reason in 3.1.3. The third reason is that $g^+$ patients may turn out to not be the patients receiving maximum efficacy. For example, Xalkori’s original target was only MET and the original protocol for its Phase II trials was to enroll only $g^+$ patients, those with amplification in MET. As reported in Ou (2011), fortuitously, the Phase II trials had two patients with ALK translocation that experienced dramatic improvement early on. The protocol was then amended to target ALK. (In fact, none of the patients with ALK translocation had MET amplification.) As another example, in the absence of dosing according to metabolic profile, it is not necessarily easy to predict which allele of a metabolic gene corresponds to maximum efficacy from a drug. In an analysis of how polymorphism in CYP2C19 affected the efficacy of clopidogrel (Plavix) on patients in the CURE study, the point estimate of hazard ratio in the primary endpoint shown in Table 1 of Paré et al. (2010) was actually lower for poor and ultra metabolizers than for patients with wild type CYP2C19.

3.2.1 Formulating null hypotheses

Let $g^+$ denote the patient subgroup with the “normal” allele of the major metabolic gene of a drug and let $g^-$ denote its complementary subgroup. Then, testing (3.3) below in addition to (3.1) and (3.2) can verify that broad efficacy is not only derived from efficacy in the $g^+$ subgroup, but that the drug is efficacious for $g^-$ patients as well. The null hypotheses being tested simultaneously are then
$$H_0 : \overline{\theta} \leq \delta \quad \text{vs.} \quad H_a : \overline{\theta} > \delta$$

$$H_{0+} : \theta_{g+} \leq \delta \quad \text{vs.} \quad H_{a+} : \theta_{g+} > \delta$$

$$H_{0-} : \theta_{g-} \leq \delta \quad \text{vs.} \quad H_{a-} : \theta_{g-} > \delta$$

(3.3)

For testing the hypotheses above, Closed testing forms the following null hypotheses:

$$H_{0+}^C : \{ \overline{\theta} \leq \delta \cap \theta_{g+} \leq \delta \cap \theta_{g-} \leq \delta \} \quad \text{vs.} \quad H_{a+}^C : \{ \overline{\theta} > \delta \cup \theta_{g+} > \delta \cup \theta_{g-} > \delta \}$$

$$H_{0+}^C : \{ \overline{\theta} \leq \delta \cap \theta_{g+} \leq \delta \} \quad \text{vs.} \quad H_{a+}^C : \{ \overline{\theta} > \delta \cup \theta_{g+} > \delta \}$$

$$H_{0-}^C : \{ \overline{\theta} \leq \delta \cap \theta_{g-} \leq \delta \} \quad \text{vs.} \quad H_{a-}^C : \{ \overline{\theta} > \delta \cup \theta_{g-} > \delta \}$$

$$H_{0+}^C : \{ \theta_{g+} \leq \delta \cap \theta_{g-} \leq \delta \} \quad \text{vs.} \quad H_{a+}^C : \{ \theta_{g+} > \delta \cup \theta_{g-} > \delta \}$$

(3.4)

$$H_0^C : \{ \overline{\theta} \leq \delta \} \quad \text{vs.} \quad H_a^C : \{ \overline{\theta} > \delta \}$$

(3.5)

$$H_{0+}^C : \{ \theta_{g+} \leq \delta \} \quad \text{vs.} \quad H_{a+}^C : \{ \theta_{g+} > \delta \}$$

$$H_{0-}^C : \{ \theta_{g-} \leq \delta \} \quad \text{vs.} \quad H_{a-}^C : \{ \theta_{g-} > \delta \}$$

while Partition testing would initially form the null hypotheses
\[ H^*_0 \cap \delta \leq \theta_{g+} \leq \delta \cap \theta_{g-} \leq \delta \] vs. \[ H^*_a \cap \delta > \theta_{g+} > \delta \cup \theta_{g-} > \delta \]

\[ H^*_0 + \cap \delta \leq \theta_{g+} \leq \delta \cap \theta_{g-} > \delta \] vs. \[ H^*_a + \cap \delta > \theta_{g+} > \delta \]

\[ H^*_0 - \cap \delta \leq \theta_{g+} \leq \delta \cap \theta_{g-} \leq \delta \] vs. \[ H^*_a - \cap \delta > \theta_{g-} > \delta \]

\[ H^*_{0+} \cap \delta > \theta_{g+} \leq \delta \cap \theta_{g-} \leq \delta \] = \emptyset \text{ illogical} \tag{3.6}

\[ H^*_{0+} \cap \delta \leq 0 \cap \theta_{g+} > \delta \cap \theta_{g-} > \delta \] = \emptyset \text{ illogical} \tag{3.7}

\[ H^*_0 + \cap \delta > \theta_{g+} \leq \delta \cap \theta_{g-} > \delta \] vs. \[ H^*_a + \cap \theta_{g+} > \delta \]

\[ H^*_0 - \cap \delta > \theta_{g+} \leq \delta \cap \theta_{g-} \leq \delta \] vs. \[ H^*_a - \cap \theta_{g-} > \delta \]

Then Partition testing flags the null hypothesis (3.6), \( H^*_{0+} \), as illogical, testing the null hypothesis that the parameter is in the empty set, because if a drug is efficacious averaged over the entire patient population, then it must be efficacious in at least one of the \( g^+ \) and \( g^- \) subgroups. Similarly, the null hypothesis (3.7), \( H^*_{0} \), is flagged as illogical because if a drug is efficacious in both the \( g^+ \) and \( g^- \) subgroups, then it must be efficacious averaged over the entire patient population.

Routine application of the closed testing principle would test the null hypothesis (3.4) that a drug lacks efficacy on average, as well as the null hypothesis (3.5) that the drug is efficacious in neither subgroup. See Table 1 of Spiessens and Debois (2010) for an example. Hochberg’s method, a special case of Closed testing, implicitly tests (3.4) and (3.5). However, a more careful application of Closed testing, analogous to those in Shaffer (1986), Holland and Copenhaver (1987), and Westfall and Tobias (2007) in their multiple testing of equality among pairs of means, would not test either (3.4) or (3.5), because they are already effectively tested by the first three intersection hypotheses.
3.2.2 Decision rule: Partition testing vs. Closed testing

An original null hypothesis is rejected if all Closed testing or Partition testing hypotheses implying it are rejected. Table 3.2 displays decision rules for rejecting $\bar{H}_0$, $H_{0^+}$, and $H_{0^-}$ by Closed testing and Partition testing. Partition testing tests five hypotheses while (default) Closed testing tests seven hypotheses. This difference has real consequences in practice, as will be illustrated in sections 3.2.3 and 3.2.4.

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Reject $H_0$</th>
<th>Reject $H_{0^+}$</th>
<th>Reject $H_{0^-}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{0^+}^C$</td>
<td>$H_{0^+}^*$</td>
<td>∨</td>
<td>∨</td>
<td>∨</td>
</tr>
<tr>
<td>$H_{0^+}^C$</td>
<td>$H_{0^+}^*$</td>
<td>∨</td>
<td>∨</td>
<td>∨</td>
</tr>
<tr>
<td>$H_{0^-}^C$</td>
<td>$\bar{H}_{0^-}^*$</td>
<td>∨</td>
<td>∨</td>
<td>∨</td>
</tr>
<tr>
<td>$H_{0^+}^C$</td>
<td>$H_{0^-}^*$</td>
<td>n/a</td>
<td>n/a</td>
<td>∨</td>
</tr>
<tr>
<td>$H_{0^+}^C$</td>
<td>$\bar{H}_{0^-}^*$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 3.2: Closed testing (C) and Partition testing (P) decision rules for rejecting $\bar{H}_0$, $H_{0^+}$, and $H_{0^-}$. A ∨ denotes a C or P null hypothesis that needs to be rejected. An n/a denotes a P null hypothesis that need not be tested.

3.2.3 A logic and a distribution example

As shown in Huang and Hsu (2007), Hochberg step-up testing is Closed testing that applies what they call the Simes-Hochberg test to each intersection null hypothesis. For details, see Appendix C. Critical values for Hochberg’s method are computed under the assumption that the test statistics are independent. It is difficult to modify
these critical values to take dependence among the test statistics into account, because the rejection region of each intersection null hypothesis is defined by a vector of order statistics. It is difficult to deal with joint distributions of order statistics of dependent random variables.

In contrast, if the rejection region of each Partition testing or Closed testing null hypothesis depends only on a single order statistic of the test statistics involved, a maxT test or minP test described in Appendix B for example, then it is typically possible to take dependence among the test statistics into account, computing critical values so that the size of the test for each Partition testing or Closed testing null hypothesis is $\alpha$.

For insight, it is sufficient to consider the simple case where

$$
\text{Prevalence of } g^+ \text{ patients} = \text{Prevalence of } g^- \text{ patients} \\
\hat{\theta}_{g^+} \sim \text{Normal}(\theta_{g^+}, 1) \\
\hat{\theta}_{g^-} \sim \text{Normal}(\theta_{g^-}, 1)
$$

The joint distribution of the standardized test statistics $T_+ = \hat{\theta}_{g^+}$, $T_- = \hat{\theta}_{g^-}$, and $\bar{T} = \sqrt{2} \times \hat{\theta} = (\hat{\theta}_{g^+} + \hat{\theta}_{g^-})/\sqrt{2}$, is then multivariate Normal, with variance-covariance matrix in the order of $T_+, T_-, \bar{T}$

$$
\sum = \begin{bmatrix}
1 & 0 & 1/\sqrt{2} \\
0 & 1 & 1/\sqrt{2} \\
1/\sqrt{2} & 1/\sqrt{2} & 1
\end{bmatrix}.
$$

Quantiles of $max\{T_+, \bar{T}\}$, $max\{T_-, \bar{T}\}$, and $max\{T_+, T_-, \bar{T}\}$ can be computed using the algorithm of Genz and Bretz (1999).

In what we call the Logic example, Partition testing has an advantage over Hochberg’s step-up test by recognizing inferential logic among parameters in the null hypotheses.
<table>
<thead>
<tr>
<th>Rejection rule</th>
<th>Logic</th>
<th>Distribution</th>
</tr>
</thead>
</table>
| $\bar{T}$ | $T_+ = 1.65$ | $T_+ = 1.65$
| $\bar{T}$ | $T_- = 1.45$ | $T_- = 1.35$
| $\bar{T}$ | $\bar{T} = 2.19$ | $\bar{T} = 2.12$

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{H}<em>{0+}/\bar{H}</em>{0+}$</td>
<td>$\max{T_+, T_- T} &gt; 2.128$</td>
<td>$T_+ &gt; 2.028$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 2.128$</td>
<td>$\bar{T} &gt; 2.028$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 2.128$</td>
<td>$\bar{T} &gt; 2.028$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+}/\bar{H}</em>{0+}$</td>
<td>$\max{T_+, T} &gt; 1.96$</td>
<td>$T_+ &gt; 1.875$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T} &gt; 1.96$</td>
<td>$\bar{T} &gt; 1.875$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T} &gt; 1.96$</td>
<td>$\bar{T} &gt; 1.875$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+}/\bar{H}</em>{0+}$</td>
<td>$\max{T_-, T} &gt; 1.96$</td>
<td>$T_- &gt; 1.875$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_-, T} &gt; 1.96$</td>
<td>$\bar{T} &gt; 1.875$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_-, T} &gt; 1.96$</td>
<td>$\bar{T} &gt; 1.875$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+}/\bar{H}</em>{0+}$</td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$T_+ &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+}/\bar{H}</em>{0+}$</td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$T_+ &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+}/\bar{H}</em>{0+}$</td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$T_+ &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\max{T_+, T_- T} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

Table 3.3: Analysis of Logic and Distribution examples, FWER=0.05. A Partition null hypothesis is rejected by H or P if there is at least one ×.

Specifically, Partition testing recognizes that $H_{0+/-}$ need not be tested because it states the treatment is efficacious on average but ineffective in either the $g^+$ or $g^-$ subgroup. Partition testing also recognizes $H_{0+/-}$ need not be tested, because it states the treatment is efficacious for both subgroups but lacks efficacy on average. In contrast, Hochberg’s method tests both hypotheses.

The (artificial) data we choose for the Logic example, $T_+ = 1.65, T_- = 1.45, \bar{T} = 2.19$, are such that at the 1-sided 5% level, testing and failing to reject the unnecessary
Table 3.4: Analysis of Logic and Distribution examples (FWER=0.05), presented in terms of p-values. A Partition null hypothesis is rejected by H or P if there is at least one \(\times\).

null hypothesis \(H_{0+}^*\) by Hochberg’s test causes it to fail to reject the original null hypothesis \(H_{0+} : \theta_{g^+} \leq 0\). With details given in Table 3.3, for this Logic example, Partition testing (with maxT testing) infers \(\theta_{g^+} > 0\) and \(\bar{\theta} > 0\), but Hochberg’s test infers only \(\bar{\theta} > 0\).

In what we call the Distribution example, Partition testing has an advantage over Hochberg’s test by being able to take into account the joint distribution of the test statistics which Hochberg’s step-up test cannot.
The (artificial) data we choose for the Distribution example, $T_+ = 1.65, T_- = 1.35, \bar{T} = 2.12$, is such that at the 1-sided 5% level, the inability of Hochberg’s test to take the joint distribution of $\max\{T_+, T_-, \bar{T}\}$ into account causes it to fail to reject $H_{0+/-}^*$. Therefore, Hochberg’s test fails to reject any of the original null hypotheses. In contrast, with details given in Table 3.3, by being able to take the joint distribution into account in setting the critical values, for this Distribution example, Partition testing infers $\theta_g^+ > 0$ and $\bar{\theta} > 0$.

Table 3.4, with the same information as table 3.3, is presented for the convenience of readers used to seeing Hochberg’s method in its p-value form.

### 3.2.4 Analysis of data from a motivating example

Subgroup efficacy studies, testing whether drugs (such as clopidogrel) with proven efficacy, when efficacy is measured averaged over the entire patient population, has efficacy in identifiable subgroups, is an area of intense current activities.

Paré et al. (2010) studied the effect of CYP2C19 polymorphism on the efficacy of clopidogrel (Plavix). They obtained 5059 samples from patients in the Clopidogrel in Unstable Angina to Prevent Recurrent Events (CURE) trial, a randomized, double-blind, placebo controlled trial with 12562 patients. Our illustration is based on the incidence rate of the primary outcome. This primary outcome is said to have occurred if the patient experiences any one of the following three events: \{death from cardiovascular causes, nonfatal myocardial infarction, or stroke\}.

Major alleles of CYP2C19 are *1, *2, *3, and *17, with *1 being normal (wild-type), *2 and *3 being loss-of-function, *17 being gain-of-function alleles. Carriers
of at least one loss-of-function allele (i.e., *2 or *3) are classified as “loss-of-function allele carriers” \((g^-)\); otherwise, they are classified as “non-carriers” \((g^+)\).

One conclusion stated in Parè et al. (2010) was that efficacy of clopidogrel is similar between carriers and non-carriers of loss-of-function alleles. They cited a large p-value from testing for interaction between efficacy and carrier status to support this conclusion. However, as is well known (from equivalence testing literature for example), a large p-value for a null hypothesis of no difference does not necessarily imply the difference is close to zero. Besides, the issue is not whether there is differential efficacy between subgroups, but rather for which groups can efficacy be established. Testing whether the data supports inference of broad efficacy, efficacy for non-carriers of loss-of-function alleles \((g^+)\), and carriers of loss-of-function alleles \((g^-)\), as formulated in section 3.2, addresses this issue.

<table>
<thead>
<tr>
<th>Genomic subgroup</th>
<th>Treatment ((Rx))</th>
<th>Control ((C))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g^+) Loss-of-function noncarriers</td>
<td>178 (1880)</td>
<td>233 (1813)</td>
</tr>
<tr>
<td>(g^-) Loss-of-function carriers</td>
<td>52 (650)</td>
<td>78 (673)</td>
</tr>
<tr>
<td>Entire group</td>
<td>230 (2530)</td>
<td>311 (2486)</td>
</tr>
</tbody>
</table>

Table 3.5: Number of incidence in clopidogrel (Plavix) example. (Number in parentheses is total number of patients in that group.)

Table 3.5 lists the number of incidences of the primary outcome for the \(g^+\) and \(g^-\) subgroups, when patients are given clopidogrel and when they are given a placebo. For our illustration, efficacy is defined as the reduction in the incidence rates, when
patients are given clopidogrel instead of the placebo. Suppose prevalence of the $g^+$
genotype in the population is 74%. Point estimates for the reduction in incidence rates are

$$
\hat{\theta}_{g^+} = 233/1813 - 178/1880 = 0.03384 \\
\hat{\theta}_{g^-} = 78/673 - 52/650 = 0.03590 \\
\hat{\theta} = \gamma \times \hat{\theta}_{g^+} + (1 - \gamma) \times \hat{\theta}_{g^-} = 0.03437
$$

with estimated variances 0.000107, 0.000265, 0.0000767424, and estimated correlation matrix in the order of $T_+, T_-, \bar{T}$

$$
R = \begin{bmatrix}
1 & 0 & 0.8753 \\
0 & 1 & 0.4836 \\
0.8753 & 0.4836 & 1
\end{bmatrix}
$$

Our analysis will be based on the asymptotic normality of the test statistics. Note
that, for real data such as these, different subsets of the test statistics have different
variance-covariance matrices. So, in contrast to the situation in the simple (artificial)
examples in section 3.2.3, critical values of maxT (or minP) tests for each Partitioning
hypothesis do not just depend on how many parameters are involved, but also depend
on the particular set of parameters involved. However, this presents no computation
or implementation difficulty.

We first illustrate efficacy testing with $\delta = 0$. The test statistics are simply the
standardized point estimates, and they are $T_+ = 3.27, T_- = 2.20, \bar{T} = 3.92$ As shown
in Table 3.6, with FWER controlled at 2.5%, Partition testing and Hochberg’s method
make the same inferences, that clopidogrel is not only efficacious averaged over the
entire patient population, but in fact efficacious for each of the $g^+$ and $g^-$ subgroups.

Next we illustrate efficacy testing with $\delta = 0.016$. The test statistics are the point
estimates minus $\delta$, standardized by their SE, so $T_+ = 1.72, T_- = 1.22, \bar{T} = 2.10$. As
<table>
<thead>
<tr>
<th>Rejection rule</th>
<th>clopidogrel (Plavix)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_+ = 3.27$</td>
</tr>
<tr>
<td></td>
<td>$T_- = 2.20$</td>
</tr>
<tr>
<td></td>
<td>$\bar{T} = 3.92$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0+/-}/H^*_0$</td>
<td>$\max{T_+, T_- T} &gt; 2.394$</td>
<td>$T_+ &gt; 2.305$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>$T_+ &gt; 2.305$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\med{T_+, T_- \bar{T}} &gt; 2.241$</td>
<td>$T_- &gt; 2.305$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>$\bar{T} &gt; 2.305$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\min{T_+, T_- \bar{T}} &gt; 1.96$</td>
<td>$\bar{T} &gt; 2.305$</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0+}/H^*_0$</td>
<td>$\max{T_+, \bar{T}} &gt; 2.241$</td>
<td>$T_+ &gt; 2.122$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>$\bar{T} &gt; 2.122$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\min{T_+, \bar{T}} &gt; 1.96$</td>
<td>$T_- &gt; 2.214$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>$\bar{T} &gt; 2.214$</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0-}/H^*_0$</td>
<td>$\max{T_-, \bar{T}} &gt; 2.241$</td>
<td>$T_- &gt; 2.214$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>$\bar{T} &gt; 2.214$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\min{T_-, \bar{T}} &gt; 1.96$</td>
<td>$\bar{T} &gt; 2.214$</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0+}/H^*_0$</td>
<td>$\max{T_+, T_-} &gt; 2.241$</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\min{T_+, T_-} &gt; 1.96$</td>
<td>$\times$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0}/H^*_0$</td>
<td>$\bar{T} &gt; 1.96$</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0+}/H^*_0$</td>
<td>$T_+ &gt; 1.96$</td>
<td>$T_+ &gt; 1.96$</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>$\text{or}$</td>
<td>$\times$</td>
<td></td>
</tr>
<tr>
<td>$H^C_{0-}/H^*_0$</td>
<td>$T_- &gt; 1.96$</td>
<td>$T_- &gt; 1.96$</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

Table 3.6: Analysis of clopidogrel (Plavix) data, $\delta = 0$, FWER = 0.025. A Partition null hypothesis is rejected by H or P if there is at least one $\times$.

shown in Table 3.7, with FWER controlled at 5%, Partition testing and Hochberg’s method reach different results. While Partition testing infers clopidogrel is efficacious averaged over the entire patient population, and in the $g^+$ subgroups, Hochberg’s method fails to make any inference.
<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Rejection rule</th>
<th>clopidogrel (Plavix)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_+ = 1.72$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$T_- = 1.22$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{T} = 2.10$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Rejection rule</th>
<th>Hochberg (H)</th>
<th>Partition (P)</th>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{H}_0^{C-}/\bar{H}_0^{*}$</td>
<td>$\max{T_+,T_-,\bar{T}} &gt; 2.128$ or $\operatorname{med}{T_+,T_-,\bar{T}} &gt; 1.96$ or $\min{T_+,T_-,\bar{T}} &gt; 1.645$</td>
<td>$\bar{T} &gt; 2.015$</td>
<td>$\bar{T} &gt; 2.015$</td>
<td>$\times$</td>
<td></td>
</tr>
<tr>
<td>$\bar{H}_0^{C+}/\bar{H}_0^{*}$</td>
<td>$\max{T_+,\bar{T}} &gt; 1.96$ or $\min{T_+,\bar{T}} &gt; 1.645$</td>
<td>$T_+ &gt; 1.812$</td>
<td>$T_+ &gt; 1.812$</td>
<td>$\times$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$\bar{H}_0^{C-}/\bar{H}_0^{*}$</td>
<td>$\max{T_-,\bar{T}} &gt; 1.96$ or $\min{T_-,\bar{T}} &gt; 1.645$</td>
<td>$\bar{T} &gt; 1.919$</td>
<td>$\times$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0^{C-}/H_0^{*}$</td>
<td>$\max{T_+,T_-,\bar{T}} &gt; 1.96$ or $\min{T_+,T_-,\bar{T}} &gt; 1.645$</td>
<td>$\times$</td>
<td>$\times$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0^{C+}/H_0^{*}$</td>
<td>$\bar{T} &gt; 1.645$</td>
<td>$\times$</td>
<td>$\times$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0^{-}/H_0^{*}$</td>
<td>$T_- &gt; 1.645$</td>
<td>$T_- &gt; 1.645$</td>
<td>$\times$</td>
<td>$\times$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7: Analysis of clopidogrel (Plavix) data, $\delta = 0.016$, FWER = 0.05. A Partition null hypothesis is rejected by H or P if there is at least one $\times$.

### 3.2.5 A simulation comparison of performance

In multiple testing, for methods that control FWER (the probability of rejecting true null hypotheses), there are multiple ways of defining their “power” (the probability of rejecting false null hypotheses). The definition advocated in Appendix C of Hsu (1996) is $P\{\text{correct and useful inference}\}$. It is, in each region of the parameter space, the probability of rejecting exactly all the false null hypotheses in that region. For example, if a drug has efficacy in the $g^+$ subgroup and when averaged over the
patient population, but is insufficiently efficacious in the $g^-$ subgroup, then we measure the “power” of a statistical method by its probability of rejecting simultaneously the two null hypotheses of no efficacy on average, and in the $g^+$ subgroup, and not rejecting the null hypothesis of insufficient efficacy in the $g^-$ subgroup. For example, the upper-right quadrant (the shaded region in Figure 3.1) indicates FFF as the false parameter space for null hypotheses in the order $\bar{H}_0$, $H_{0+}, H_{0-}$. Therefore, its corresponding “correct and useful” inference should be to reject $\bar{H}_0$, reject $H_{0+}$ and reject $H_{0-}$, indicated by RRR. The information of marginal power can also be obtained by summing up $P\{\text{correct and useful inference}\}$ in its corresponding testing region. For example, the marginal power for testing efficacy in the $g^+$ subgroup (the circled region in Figure 3.1) is $P\{RRR\} + P\{RRA\} + P\{ARA\}$.

**Partitioning with maxT versus Hochberg’s step-up method**

In Figure 3.2, we show values of $P\{\text{correct and useful inference}\}$ of Partition testing, and its difference with Hochberg’s method, using heatmaps. Positive $\theta_{g^+}$ and $\theta_{g^-}$ values represent efficacy in the $g^+$ and $g^-$ subgroups, respectively, while broad efficacy is represented by $(\theta_{g^+}, \theta_{g^-})$ values to the upper-right of the $-45^\circ$ line.

We generated $\hat{\theta}_{g^+}$ and $\hat{\theta}_{g^-}$ as independent, normally distributed random variables with means $\theta_{g^+}$ and $\theta_{g^-}$, and variances 1. Dividing a range of $(-10, +10)$ into 99 evenly spaced intervals for $\theta_{g^+}$ and $\theta_{g^-}$, 10,000 i.i.d. pairs of $(\hat{\theta}_{g^+}, \hat{\theta}_{g^-})$ were generated (independently) for each of the 10,000 grid points. Assuming a prevalence of 50% for $g^+$ patients, $\hat{\theta}$ is then computed as $0.5\hat{\theta}_{g^+} + 0.5\hat{\theta}_{g^-}$. Partition testing and Hochberg’s step-up method were then applied to test the null hypotheses $\bar{H}_0 : \theta \leq \delta$, $H_{0+} : \theta_{g^+} \leq \delta$, $H_{0-} : \theta_{g^-} \leq \delta$. 

77
Figure 3.1: Correct and useful inference and its corresponding parameter region in the order $H_0$, $H_{0+}$, $H_{0-}$. T or F means true or false null hypothesis for the parameter region. R or A means rejecting or accepting null hypothesis in its corresponding parameter region. The $g^+$ and $g^-$ values represent $\theta_{g^+}$ and $\theta_{g^-}$

The left column shows, in the three regions of the parameter space where the drug is efficacious for the $g^+$ subgroup, $P\{\text{correct and useful inference}\}$ by Partition testing. “Correct and useful” inference in each region of the parameter space is indicated by the letters A, representing the acceptance (failure to reject), and R, representing rejection, of the null hypotheses $H_0$, $H_{0+}$, $H_{0-}$, in that order. For example, in the part of the lower-right quadrant of the parameter space that is to the upper-right of the $-45^\circ$ line, correct and useful inference is to reject $H_0$, reject $H_{0+}$, and accept $H_{0-}$, as indicated by RRA.

The middle column shows the difference in $P\{\text{correct and useful inference}\}$, in the direction of Partition testing minus Hochberg’s method. It turns out, when the drug has efficacy in both $g^+$ and $g^-$ subgroups (and therefore efficacy on average), there
Figure 3.2: $P\{\text{correct and useful inference}\}$ of Partition testing, and its difference with Hochberg’s method. Positive $g^+$ and $g^-$ value represent efficacy in the $g^+$ and $g^-$ subgroup respectively, while broad efficacy is represented by $(g^+, g^-)$ values to the upper-right of the $-45^\circ$ line.
is no significant difference in the probability. Therefore, that region of the parameter space is not plotted.

For completeness, Figure 3.2 also displays $P\{\text{RAA}\}$, the probability that efficacy is inferred in the broad population but not in either the $g^+$ or the $g^-$ subgroups. The top heatmap of the right column is $P\{\text{RAA}\}$ for Partition testing, showing such a probability is small throughout the parameter space. The bottom heatmap of the right column is the difference in $P\{\text{RAA}\}$ between Partition testing and Hochberg’s method, in the direction of Partition testing minus Hochberg’s method. When the drug is slightly efficacious, approximately equally for $g^+$ and $g^-$ patients, both Partition testing and Hochberg’s method lack power; Hochberg’s method has slightly higher probability of inferring efficacy in the broad population while failing to infer efficacy in either subgroup.

A summary of the simulation study results is as follows.

1. When the drug compound is substantially efficacious for both $g^+$ and $g^-$ patients (and thus on average), Partition testing and Hochberg’s method are equally likely to infer simultaneously broad efficacy and efficacy in both the $g^+$ and $g^-$ subgroups.

2. When the drug compound is efficacious for $g^+$ patients and sufficiently efficacious when averaged over the patient population, Partition testing is more likely to infer simultaneously broad efficacy and efficacy in the $g^+$ subgroup than Hochberg’s method. (Hochberg’s method is more likely to infer efficacy in the $g^+$ subgroup only, compared to Partition testing.)
3. When the drug compound is only efficacious for $g^+$ patients but not sufficiently efficacious when averaged over the patient population, Partition testing is more likely to infer efficacy in the $g^+$ subgroup than Hochberg’s method.

4. When the drug is slightly efficacious, approximately equally for $g^+$ and $g^-$ patients, both Partition testing and Hochberg’s method lack power to infer efficacy. Hochberg’s method has slightly higher probability of inferring efficacy in the broad population while failing to infer efficacy in either subgroup.

**Partitioning with maxT versus closed testing with maxT**

People might argue that Hochberg’s method usually is more powerful when the test statistics are independent or have a distribution with multivariate total positivity of order two or a scale mixture thereof for its validity (Sarkar, 1998) In genetic subgroup analysis, a genetic subgroup is always correlated with the entire patient population since the subgroup is a part of the entire group. Therefore, it is not surprising that Partition testing with maxT can perform better than Hochberg’s method since maxT test takes the joint distribution of the test statistics into account.

In this section, we will show the fundamental difference between Partition testing and Closed testing in genetic subgroup analysis by applying the maxT test to both principles. For example, a graphical approach using parametric tests proposed by Bretz et al. (2011) [5] is based on Closed testing procedure but it considers the joint distribution of the test statistics. In multiple doses and multiple endpoints settings, the graphical approach with a parametric test might produce the same performance as Partition with maxT. However, Partitioning respects the logical relationship between
parameters in the genetic subgroup setting. It might be interesting to compare the difference of performance in this setting.

By using the same simulation method as section 3.2.5, Figure 3.3 demonstrates $P\{\text{correct and useful inference}\}$ for Partitioning with maxT, Closed testing with maxT and the difference of probability between these two principles. The left column shows, in the three regions of parameter space where the drug is efficacious for the $g^+$ subgroup, $P\{\text{correct and useful inference}\}$ by Partition testing with maxT. The middle column shows $P\{\text{correct and useful inference}\}$ by Closed testing with maxT. The right column shows the difference in $P\{\text{correct and useful inference}\}$, in the direction of Partition testing minus Closed testing. For completeness, Figure 3.3 also displays $P\{RAA\}$, the probability that efficacy is inferred in the broad population but not in either the $g^+$ or the $g^-$ subgroups.

A summary of the simulation study results is as follows.

1. It is impossible to infer efficacy in both both $g^+$ or the $g^-$ subgroups but not for broad efficacy, $P\{ARR\} = 0$, no matter where Partitioning or Closed testing principle is used.

2. There is no difference between Partitioning and Closed testing in terms of $P\{AAA\}$, $P\{ARA\}$ and $P\{AAR\}$. That is, Partitioning and Closed testing have the same chance to claim efficacy for only one subgroup or claim no efficacy for all of them.

3. When the drug is slightly efficacious, approximately equally for $g^+$ and $g^-$ patients, Partition testing has a higher probability to claim either a big success ($P\{RRR\}$) or small success ($P\{RRA\}$ or $P\{RAR\}$). On the other hand, Closed
testing has a higher chance to claim efficacy on average only \( P\{RAA\} \), which is not a useful inference.

### 3.3 Simultaneous confidence bounds for broad efficacy and efficacy in complementary subgroups

Confidence bounds are more informative than point estimates and p-values (ICH E9). When the drug has no effect on a subgroup, the chance that the point estimate for its effect on that subgroup will be in the right direction is 50%. A large p-value for equality of efficacy in subgroups does not infer equivalent effect in the subgroups, as it may be due to noisy data, or insufficient sample size. When efficacy is demonstrated for the broad population, and/or a subgroup, our confidence bounds allow assessment of whether there is concern over efficacy in the complementary subgroup. Should there be concern, our recommendation is that the compound be approved for use by the subgroup for which efficacy is shown, but a statement be included on the label for patients in the complementary subgroup, perhaps as part of the Contraindication, that efficacy in that subgroup has not been shown.

In section 3.3.1, we show how to pivot partition tests for (3.1), (3.2), and (3.3) to obtain simultaneous confidence bounds for \( \bar{\theta}, \theta_{g+}, \theta_{g-} \) and generalize the pivoting procedure to a Theorem in section 3.3.2. Then the two artificial examples and Plavix example are demonstrated by assessing the confidence set for \( \bar{\theta}, \theta_{g+}, \theta_{g-} \).

#### 3.3.1 Confidence set for assessing consistency in efficacy

Recall that in section 3.2, to test (3.1), (3.2), and (3.3), Partition testing and Closed testing form the following null hypotheses to test broad efficacy and efficacy in the \( g_+ \) and \( g_- \) subgroups simultaneously.
Figure 3.3: $P\{$correct and useful inference$\}$ of Partition testing with maxT test, and its difference with closed testing with maxT test. Positive $g^+$ and $g^-$ value represent efficacy in the $g^+$ and $g^-$ subgroup respectively, while broad efficacy is represented by $(g^+, g^-)$ values to the upper-right of the $-45^\circ$ line.
<table>
<thead>
<tr>
<th>Closed Testing</th>
<th>Partition Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^C_{0+ -} : { \theta \leq \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } $</td>
<td>$H^*_{0+ -} : { \bar{\theta} \leq \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } $</td>
</tr>
<tr>
<td>$H^C_{0+} : { \bar{\theta} \leq \delta \cap \theta_g^+ \leq \delta } $</td>
<td>$H^*_{0+} : { \bar{\theta} \leq \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- &gt; \delta } $</td>
</tr>
<tr>
<td>$H^C_{0-} : { \bar{\theta} \leq \delta \cap \theta_g^- \leq \delta } $</td>
<td>$H^*_{0-} : { \bar{\theta} \leq \delta \cap \theta_g^+ &gt; \delta \cap \theta_g^- \leq \delta } $</td>
</tr>
<tr>
<td>$H^C_{0+ -} : { \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } $</td>
<td>$H^*_{0+ -} : { \bar{\theta} &gt; \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } = \emptyset $</td>
</tr>
<tr>
<td>$H^C_{0+} : { \theta_g^+ \leq \delta } $</td>
<td>$H^*_{0+} : { \bar{\theta} &gt; \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- &gt; \delta } $</td>
</tr>
<tr>
<td>$H^C_{0-} : { \theta_g^- \leq \delta } $</td>
<td>$H^*_{0-} : { \bar{\theta} &gt; \delta \cap \theta_g^+ &gt; \delta \cap \theta_g^- \leq \delta } $</td>
</tr>
<tr>
<td>$H^C_{0-} : { \theta_g^- \leq \delta } $</td>
<td>$H^*_{0-} : { \bar{\theta} &gt; \delta \cap \theta_g^+ &gt; \delta \cap \theta_g^- &gt; \delta } $</td>
</tr>
</tbody>
</table>

Table 3.8: null hypotheses constructed from closed testing and partition testing principle

To test for efficacy, the last partition null hypothesis $H^*_{\emptyset}$, which states there is efficacy on average and in each of the subgroups, need not be tested. However, displaying it completes the partitioning of the parameter space, and it sets up the construction of the confidence set for the parameters $\bar{\theta}, \theta_g^+, \theta_g^-$, which we do in the next section.

The recommended acceptance regions, shown in Table 3.9, are directed toward rejecting the “$\leq$” in the partitioning null hypotheses. It is clear that, for each partitioning null hypothesis excepting $H^*_{\emptyset}$, the LFC is at $\theta_g^+ = \theta_g^- = \bar{\theta} = \delta$. Under our distribution assumptions, the critical values for those five partitioning hypotheses can then be computed, free of unknown parameters. They can in fact be computed exactly, taking the joint distribution of $T^+, T^-, \bar{T}$ into account.

How to test the parameter points in $H^*_{\emptyset}$, needed for the construction of the confidence set for $(\theta_g^+, \theta_g^-, \bar{\theta})$, is discussed in section 3.3.2.
<table>
<thead>
<tr>
<th>Partition null hypothesis</th>
<th>Acceptance region</th>
<th>Confidence region</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{H}_{0+}^* : { \bar{\theta} \leq \delta \wedge \theta_g^+ \leq \delta \wedge \theta_g^- \leq \delta } )</td>
<td>( \hat{\theta}<em>g^+ - \theta_g^+ &lt; c</em>{\theta_g^+} \hat{\sigma}<em>{\theta_g^+} ) and ( \hat{\theta}<em>g^- - \theta_g^- &lt; c</em>{\theta_g^-} \hat{\sigma}</em>{\theta_g^-} ) and ( \hat{\theta} - \bar{\theta} &lt; c_{\bar{\theta}} \hat{\sigma}_{\bar{\theta}} )</td>
<td>( \theta_g^+ - c_{\theta_g^+} \hat{\sigma}<em>{\theta_g^+} &lt; \theta_g^+ \leq \delta ) and ( \theta_g^- - c</em>{\theta_g^-} \hat{\sigma}<em>{\theta_g^-} &lt; \theta_g^- \leq \delta ) and ( \hat{\theta} - c</em>{\bar{\theta}} \hat{\sigma}_{\bar{\theta}} &lt; \hat{\theta} \leq \delta )</td>
</tr>
<tr>
<td>( \bar{H}_{0+}^* : { \bar{\theta} \leq \delta \wedge \theta_g^+ \leq \delta \wedge \theta_g^- &gt; \delta } )</td>
<td>( \theta_g^+ - \theta_g^+ &lt; c_{\theta_g^+} \hat{\sigma}<em>{\theta_g^+} ) and ( \hat{\theta} - \bar{\theta} &lt; c</em>{\bar{\theta}} \hat{\sigma}_{\bar{\theta}} )</td>
<td>( \theta_g^+ - c_{\theta_g^+} \hat{\sigma}<em>{\theta_g^+} &lt; \theta_g^+ \leq \delta ) and ( \delta &lt; \theta_g^- ) and ( \hat{\theta} - c</em>{\bar{\theta}} \hat{\sigma}_{\bar{\theta}} &lt; \hat{\theta} \leq \delta )</td>
</tr>
<tr>
<td>( \bar{H}_{0-}^* : { \bar{\theta} \leq \delta \wedge \theta_g^+ &gt; \delta \wedge \theta_g^- \leq \delta } )</td>
<td>( \hat{\theta}<em>g^+ - \theta_g^+ &lt; c</em>{\theta_g^+} \hat{\sigma}<em>{\theta_g^+} ) and ( \hat{\theta}<em>g^- - \theta_g^- &lt; c</em>{\theta_g^-} \hat{\sigma}</em>{\theta_g^-} ) and ( \hat{\theta} - \bar{\theta} &lt; c_{\bar{\theta}} \hat{\sigma}_{\bar{\theta}} )</td>
<td>( \theta_g^+ - c_{\theta_g^+} \hat{\sigma}_{\theta_g^+} &lt; \theta_g^+ \leq \delta ) and ( \delta &lt; \theta_g^- ) and ( \delta &lt; \hat{\theta} )</td>
</tr>
<tr>
<td>( \bar{H}_{0-}^* : { \bar{\theta} &gt; \delta \wedge \theta_g^+ \leq \delta \wedge \theta_g^- &gt; \delta } )</td>
<td>( \min{\theta_g^+, \theta_g^-} &gt; ) and ( \min{\theta_g^+} &gt; \max{\delta, ) and ( \min{\hat{\theta}<em>g^+ - c</em>{\theta_g^+} \hat{\sigma}<em>{\theta_g^+}, ) and ( \theta_g^- - c</em>{\theta_g^-} \hat{\sigma}_{\theta_g^-} )</td>
<td>( \theta_g^+ - c_{\theta_g^+} \hat{\sigma}_{\theta_g^+} &lt; \theta_g^+ \leq \delta ) and ( \delta &lt; \theta_g^- ) and ( \delta &lt; \hat{\theta} )</td>
</tr>
</tbody>
</table>

Table 3.9: Confidence set is computed by first pivoting each acceptance region within its partitioning null hypothesis space to obtain a corresponding confidence region, and then taking the union of the confidence regions.
3.3.2 Theorem of constructing a confidence set

A second advantage of Partition Tests is they readily pivot to confidence sets. This is because, by partitioning the parameter space, every parameter value is tested exactly once, and a confidence set can be constructed using the following theorem, which is a careful application of the correspondence between family of tests and confidence sets. In contrast, in Closed testing, each parameter point is apparently tested more than once, making construction of a confidence set less obvious.

**Theorem 3.3.1** Let \( \theta = (\theta_1, \ldots, \theta_p) \). Consider a partition \( \Theta_1, \ldots, \Theta_m \) of the parameter space, that is,

\[
\bigcup_{i=1}^{m} \Theta_i = \Theta
\]

and

\[
\Theta_i \cap \Theta_j = \emptyset \text{ for all } i \neq j.
\]

If each acceptance region \( A_i(\theta^0), i = 1, \ldots, m \), defines a level-\( \alpha \) test for

\[
H_0 : \theta = \theta^0, \theta^0 \in \Theta_i,
\]

by rejecting only if \( \hat{\theta} \notin A_i(\theta^0) \), then

\[
C(\hat{\theta}_1, \ldots, \hat{\theta}_p) = \bigcup_{i=1}^{m} C_i(\hat{\theta}_1, \ldots, \hat{\theta}_p) = \bigcup_{i=1}^{m} \left( \{ \theta : \theta \in A_i(\theta) \} \cap \Theta_i \right)
\]

is a 100(1 - \( \alpha \))% confidence set for \( \theta = (\theta_1, \ldots, \theta_p) \).

**Proof.** Suppose \( \theta = \theta^0 \in \Theta_i \). Then

\[
P_{\theta^0} \{ \theta^0 \in C(\hat{\theta}) \} = P_{\theta^0} \{ \theta^0 \in \bigcup_{i=1}^{m} \left( \{ \theta : \theta \in A_i(\theta) \} \cap \Theta_i \right) \}
\]

\[
= P_{\theta^0} \{ \hat{\theta} \in A_i(\theta^0) \}
\]

\[
= 1 - \alpha.
\]
In applying this theorem to our problem,

\[ \Theta_1 = \{ \boldsymbol{\theta} \mid (\hat{\theta}, \theta_{g+}, \theta_{g-}) \in \bar{H}_0^* \} \]

\[ \vdots \]

\[ \Theta_6 = \{ \boldsymbol{\theta} \mid (\hat{\theta}, \theta_{g+}, \theta_{g-}) \in H_0^* \}. \]

With the acceptance regions \( A_i(\boldsymbol{\theta}), i = 1, \ldots, 5 \), shown in Table 3.9, each of \( A_1(\boldsymbol{\theta}), \ldots, A_5(\boldsymbol{\theta}) \) is equivariant in \( \boldsymbol{\theta} \), so

\[ \{ \boldsymbol{\theta} : \hat{\theta} \in A_i(\boldsymbol{\theta}) \} = -A_i(0) + \hat{\theta}. \]

Thus, to calculate the \( \bigcup_{i=1}^{5} (\{ \boldsymbol{\theta} : \hat{\theta} \in A_i(\boldsymbol{\theta}) \} \cap \Theta_i) \) part of the confidence set \( C(\hat{\theta}_1, \ldots, \hat{\theta}_p) \), we first reflect each \( A_i(0) \) through the origin, shift by \( \hat{\theta} \), intersect with \( \Theta_i \), then take the union. (An empty intersection with a \( \Theta_i \) implies the rejection of the corresponding partitioning null hypothesis.)

Figure 3.4 gives some examples of such unions. Confidence bounds on \( \theta_{g+}, \theta_{g-}, \bar{\theta} \) can then be deduced by projection, as indicated by the dotted lines in the figure.

When efficacy averaged over the subgroups is demonstrated, data may be sufficient to infer efficacy in only one of the subgroups (Case IV in Figure 3.4), or it may be insufficient to infer efficacy in either subgroup (Case II in Figure 3.4). When efficacy in a subgroup is demonstrated, data may be insufficient to infer efficacy on average or in the other subgroup (Case III in Figure 3.4).

A consequence of choosing different acceptance regions to test different partitioning hypotheses is that when efficacy in a patient group is inferred, the lower confidence bound on it equals \( \delta \) (not greater than \( \delta \)). When evidence is insufficient to establish efficacy in a patient group, the lower confidence bound for the corresponding parameter is less than \( \delta \), and how far a lower bound on \( \theta_{g-} \) or \( \theta_{g+} \) is from \( \delta \) lets one assess
whether there is concern regarding “consistency” in efficacy. Taking the conservative approach of using the same acceptance region to test all partitioning hypotheses, the one for testing \( \bar{H}_{0+} \), results in a lower confidence bound greater than \( \delta \) when efficacy in a patient group is inferred, but with loss of power to infer such efficacy compared to our recommended acceptance regions. Direct correspondence between partition tests and confidence sets allow assessment of consistency in patient subgroups that do not reach the efficacy threshold, with no loss in power to infer efficacy.

Figure 3.4: Examples of confidence set. In each graph, the red dot is the point estimate of \((\theta_{g+}, \theta_{g-})\), and the dotted lines are the deduced confidence bounds on \(\theta_{g+}, \theta_{g-}, \bar{\theta}\).
If data are sufficient to infer efficacy in both subgroups (and therefore efficacy averaged over the subgroups), then it would be useful to have a positive lower bound on minimum efficacy.

**Theorem 3.3.2 lemma.IUT** For every \( \theta = \theta^0 = (\bar{\theta}^0, \theta_{g^+}^0, \theta_{g^-}^0) \) in \( \Theta_\theta \), let \( \eta^0 \) denote \( \min\{\theta_{g^+}^0, \theta_{g^-}^0\} \). Then the acceptance region

\[
A_\alpha(\theta^0) = \left\{ \min\{\hat{\theta}_{g^+} - c_+\hat{\sigma}_{\theta_{g^+}}, \hat{\theta}_{g^-} - c_-\hat{\sigma}_{\theta_{g^-}} \} \leq \eta^0 \right\}
\]

defines a level-\( \alpha \) test for

\[ H_0 : \theta = \theta^0, \theta^0 \in \Theta_\theta. \]

**Proof.** Without loss of generality, suppose \( \theta_{g^-}^0 = \min\{\theta_{g^+}^0, \theta_{g^-}^0\} \). Then

\[
P_{\theta^0}\{\text{Reject } H_0 : \theta = \theta^0\} \leq P_{\theta^0}\{\hat{\theta}_{g^-} - \theta_{g^-}^0 > c_-\hat{\sigma}_{\theta_{g^-}}\}
\]

\[= \alpha. \]

The test we propose has the same acceptance region as an *intersection-union* test (Berger 1982) for the intersection alternative hypothesis

\[ H_a : \min\{\theta_{g^+}, \theta_{g^-}\} > \eta^0. \]

As the critical values \( c_+ \) and \( c_- \) are computed without multiplicity adjustment, this choice of tests guarantees the lower confidence bound for all three parameters \( \theta_{g^+}, \bar{\theta}, \theta_{g^-} \) obtained by inverting these tests for \( \theta \in \Theta_\theta \) will be strictly positive when data are sufficient to infer broad efficacy and efficacy in both subgroups. This bound is \( \min\{\hat{\theta}_{g^+} - c_+\hat{\sigma}_{\theta_{g^+}}, \hat{\theta}_{g^-} - c_-\hat{\sigma}_{\theta_{g^-}}\} \), as depicted in Case V of Figure 3.4.

**3.3.3 Analysis of two artificial examples**

The (artificial) data we choose for the Logic example, \( T_+ = 1.65, T_- = 1.45, \bar{T} = 2.19 \), is such that at the 1-sided 5% level, testing and failure to reject the unnecessary
null hypothesis $H_{0+}^{C}$ by Hochberg’s test causes it to fail to reject the original null hypothesis $H_{0+} : \theta_{g+} \leq 0$. With details given in Table 3.3, for this Logic example, Partition testing (with maxT testing) infers $\theta_{g+} > 0$ and $\bar{\theta} > 0$, but Hochberg’s test infers only $\bar{\theta} > 0$. Confidence bounds from Partition testing are $\bar{\theta} > 0$, $\theta_{g+} > 0$, and $\theta_{g-} > -0.195$.

In what we call the Distribution example, Partition testing has an advantage over Hochberg’s test by being able to take into account joint distribution of the test statistics which Hochberg’s step-up test cannot.

The (artificial) data we choose for the Distribution example, $T_+ = 1.65, T_- = 1.35, \bar{T} = 2.12$, are such that at the 1-sided 5% level, the inability of Hochberg’s test to take the joint distribution of $\text{max}\{T_+, T_-, \bar{T}\}$ into account causes it to fail to reject $H_{0+}^{C}$. Therefore, Hochberg’s test fails to reject any of the original null hypotheses. In contrast, with details given in Table 3.3, by being able to take the joint distribution into account in setting the critical values for this Distribution example, Partition testing infers $\theta_{g+} > 0$ and $\bar{\theta} > 0$. Confidence bounds from Partition testing are $\bar{\theta} > 0$, $\theta_{g+} > 0$, and $\theta_{g-} > -0.295$.

### 3.3.4 Analysis of data from a motivating example

We first illustrate efficacy testing with $\delta = 0$. The test statistics are simply the standardized point estimates, and they are $T_+ = 3.27, T_- = 2.20$, and $\bar{T} = 3.92$. As shown in Table 3.6, with FWER controlled at 5%, Partition testing and Hochberg’s method make the same inferences, that clopidogrel is not only efficacious averaged over the entire patient population, but in fact efficacious for each of the $g^+$ and $g^-$ subgroups. Confidence bounds from Partition testing are $\bar{\theta} > 0.00912$, $\theta_{g+} >$
0.00912, and $\theta_g^- > 0.00912$. Next we illustrate efficacy testing with $\delta = 0.016$. The test statistics are the point estimates minus $\delta$, standardized by their SE, so $T_+ = 1.72, T_- = 1.22$, and $\bar{T} = 2.10$. As shown in Table 3.7, with FWER controlled at 5%, Partition testing and Hochberg’s method reach different results. While Partition testing infers clopidogrel is efficacious averaged over the entire patient population, and in the $g^+$ subgroups, Hochberg’s method fails to make any inference. Confidence bounds from Partition testing are $\bar{\theta} > 0.016, \theta_{g^+} > 0.016$, and $\theta_{g^-} > 0.00912$. The detailed computation of simultaneous confidence bounds for the Plavix example is given in Table 3.11.
<table>
<thead>
<tr>
<th>Partition null hypotheses</th>
<th>Logic example</th>
<th>Distribution example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{H}<em>{0+} \setminus {\bar{\theta} \leq 0 \wedge \theta</em>{g} \leq 0 \wedge \theta_{g} \leq 0}$</td>
<td>$-0.378 &lt; \theta_{g} \leq 0$ and $-0.578 &lt; \theta_{g} \leq 0$ and $0.116 &lt; \bar{\theta} \leq 0$</td>
<td>$-0.378 &lt; \theta_{g} \leq 0$ and $-0.678 &lt; \theta_{g} \leq 0$ and $0.066 &lt; \bar{\theta} \leq 0$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+} \setminus {\bar{\theta} \leq 0 \wedge \theta</em>{g} \leq 0 \wedge \theta_{g} &gt; 0}$</td>
<td>$-0.225 &lt; \theta_{g} \leq 0$ and $0 &lt; \theta_{g} \leq 0$ and $0.224 &lt; \bar{\theta} \leq 0$</td>
<td>$-0.225 &lt; \theta_{g} \leq 0$ and $0 &lt; \theta_{g} \leq 0$ and $0.174 &lt; \bar{\theta} \leq 0$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0-} \setminus {\bar{\theta} \leq 0 \wedge \theta</em>{g} &gt; 0 \wedge \theta_{g} \leq 0}$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $-0.425 &lt; \theta_{g} \leq 0$ and $0.224 &lt; \bar{\theta} \leq 0$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $-0.525 &lt; \theta_{g} \leq 0$ and $0.174 &lt; \bar{\theta} \leq 0$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0+} \setminus {\bar{\theta} &gt; 0 \wedge \theta</em>{g} \leq 0 \wedge \theta_{g} &gt; 0}$</td>
<td>${0.005 &lt; \theta_{g} \leq 0}$ and $0 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
<td>${0.005 &lt; \theta_{g} \leq 0}$ and $0 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
</tr>
<tr>
<td>$\bar{H}<em>{0-} \setminus {\bar{\theta} &gt; 0 \wedge \theta</em>{g} &gt; 0 \wedge \theta_{g} \leq 0}$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $-0.195 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $-0.295 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
</tr>
<tr>
<td>$\bar{H}<em>{00} \setminus {\bar{\theta} &gt; 0 \wedge \theta</em>{g} &gt; 0 \wedge \theta_{g} &gt; 0}$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $0 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $0 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
</tr>
<tr>
<td>Confidence bounds</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $-0.195 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
<td>$0 &lt; \theta_{g} \leq 0$ and $-0.295 &lt; \theta_{g} \leq 0$ and $0 &lt; \bar{\theta}$</td>
</tr>
</tbody>
</table>

Table 3.10: 95% confidence bounds calculation for distribution and logic examples
<table>
<thead>
<tr>
<th>Partition null hypotheses</th>
<th>Confidence region</th>
<th>( \delta = 0 )</th>
<th>( \delta = 0.016 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{0+}^* : { \bar{\theta} \leq \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } )</td>
<td>0.01300 &lt; ( \theta_g^+ \leq 0 ) and 0.00310 &lt; ( \theta_g^- \leq 0 ) and {0.01672 &lt; \bar{\theta} \leq 0} = \emptyset</td>
<td>0.01300 &lt; ( \theta_g^+ \leq 0.016 ) and 0.00310 &lt; ( \theta_g^- \leq 0.016 ) and {0.01672 &lt; \bar{\theta} \leq 0.016} = \emptyset</td>
<td></td>
</tr>
<tr>
<td>( H_{0+}^* : { \bar{\theta} \leq \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } )</td>
<td>0.01510 &lt; ( \theta_g^+ \leq 0 ) and 0 &lt; ( \theta_g^- ) and {0.01850 &lt; \bar{\theta} \leq 0} = \emptyset</td>
<td>0.01510 &lt; ( \theta_g^+ \leq 0.016 ) and 0.016 &lt; ( \theta_g^- ) and {0.01850 &lt; \bar{\theta} \leq 0.016} = \emptyset</td>
<td></td>
</tr>
<tr>
<td>( H_{0-}^* : { \bar{\theta} \leq \delta \cap \theta_g^+ &gt; \delta \cap \theta_g^- \leq \delta } )</td>
<td>0 &lt; ( \theta_g^+ ) and 00466 &lt; ( \theta_g^- ) and {0.01756 &lt; \bar{\theta} \leq 0} = \emptyset</td>
<td>0.016 &lt; ( \theta_g^+ ) and 0.00466 &lt; ( \theta_g^- ) and {0.01756 &lt; \bar{\theta} \leq 0.016} = \emptyset</td>
<td></td>
</tr>
<tr>
<td>( H_{0+}^* : { \bar{\theta} &gt; \delta \cap \theta_g^+ \leq \delta \cap \theta_g^- \leq \delta } )</td>
<td>{0.01682 &lt; ( \theta_g^+ \leq 0} = \emptyset and 0 &lt; ( \theta_g^- ) and 0 &lt; ( \bar{\theta} ) and 0.016 &lt; ( \bar{\theta} )</td>
<td>{0.01682 &lt; ( \theta_g^+ \leq 0.016} = \emptyset and 0 &lt; ( \theta_g^- ) and 0.016 &lt; ( \bar{\theta} )</td>
<td></td>
</tr>
<tr>
<td>( H_{0-}^* : { \bar{\theta} &gt; \delta \cap \theta_g^+ &gt; \delta \cap \theta_g^- \leq \delta } )</td>
<td>0 &lt; ( \theta_g^+ ) and {0.00912 &lt; ( \theta_g^- \leq 0} = \emptyset and 0 &lt; ( \bar{\theta} ) and 0.016 &lt; ( \bar{\theta} )</td>
<td>0.016 &lt; ( \theta_g^+ ) and {0.00912 &lt; ( \theta_g^- \leq 0} = \emptyset and 0.016 &lt; ( \bar{\theta} )</td>
<td></td>
</tr>
<tr>
<td>( H_{00}^* : { \bar{\theta} &gt; \delta \cap \theta_g^+ &gt; \delta \cap \theta_g^- \leq \delta } )</td>
<td>00912 &lt; ( \theta_g^+ ) and 00912 &lt; ( \theta_g^- ) and 00912 &lt; ( \bar{\theta} ) and 0.016 &lt; ( \theta_g^+ ) and 0.016 &lt; ( \theta_g^- ) and 0.016 &lt; ( \bar{\theta} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.11: 95% confidence bounds calculation for Plavix example, \( \delta = 0 \) and \( \delta = 0.016 \)
Chapter 4: Simultaneous testing for broad efficacy and efficacy in three or more subgroups

4.1 Simultaneous testing for broad efficacy and efficacy in three or more subgroups

Let $g_1$, $g_2$, and $g_3$ denote three subgroups of patients with different variants of one gene. These could be patients with loss-of-function, Wild Type, and gain-of-function alleles of a P450 metabolic gene, for example. Then the null hypotheses of interest involve not only the broad efficacy of the entire population and each of the subgroups, but also efficacy in the three two-group-combined subgroups. There are seven hypotheses of interest, as follows.
\[
\begin{align*}
\bar{H}_0 : \bar{\theta} &= \gamma_1 \times \theta_{g1} + \gamma_2 \times \theta_{g2} + \gamma_3 \times \theta_{g3} \leq \delta & \text{vs.} & \bar{H}_a : \bar{\theta} > \delta \\
\bar{H}_{012} : \bar{\theta}_{12} &= (\gamma_1/(\gamma_1 + \gamma_2)) \times \theta_{g1} + (\gamma_2/(\gamma_1 + \gamma_2)) \times \theta_{g2} \leq \delta & \text{vs.} & \bar{H}_{a12} : \bar{\theta}_{12} > \delta \\
\bar{H}_{013} : \bar{\theta}_{13} &= (\gamma_1/(\gamma_1 + \gamma_3)) \times \theta_{g1} + (\gamma_3/(\gamma_1 + \gamma_3)) \times \theta_{g3} \leq \delta & \text{vs.} & \bar{H}_{a13} : \bar{\theta}_{13} > \delta \\
\bar{H}_{023} : \bar{\theta}_{23} &= (\gamma_2/(\gamma_2 + \gamma_3)) \times \theta_{g2} + (\gamma_3/(\gamma_2 + \gamma_3)) \times \theta_{g3} \leq \delta & \text{vs.} & \bar{H}_{a23} : \bar{\theta}_{23} > \delta \\
\bar{H}_{01} : \theta_{g1} \leq \delta & \text{vs.} & \bar{H}_{a1} : \theta_{g1} > \delta \\
\bar{H}_{02} : \theta_{g2} \leq \delta & \text{vs.} & \bar{H}_{a2} : \theta_{g2} > \delta \\
\bar{H}_{03} : \theta_{g3} \leq \delta & \text{vs.} & \bar{H}_{a3} : \theta_{g3} > \delta
\end{align*}
\]

Here \(\gamma_1, \gamma_2\) and \(\gamma_3\) are prevalence for subgroups \(g_1, g_2\) and \(g_3\), respectively. Since the entire population is divided into three subgroups, the sum of the prevalences is 1 \((\gamma_1 + \gamma_2 + \gamma_3 = 1)\).

To test the seven hypotheses above, Closed testing using Hochberg’s method would test \(2^7 - 1 = 127\) hypotheses, including illogical hypotheses such as the drug is efficacious in both the \(g1\) and \(g2\) subgroups, but not efficacious averaged over the \(g1\) and \(g2\) subgroups. Recognizing logical relationships among the hypotheses, only 31 hypotheses are tested by Partition testing instead. (Symbolic mathematics computation can be used to generate the logically consistent hypotheses.)

Details of how to test three or more subgroups are beyond the scope of this dissertation. However, we caution against the temptation to test as many subgroups as possible, even using Partition testing, for the following reasons.

First, the more subgroups tested, the larger the multiplicity-adjusted critical values must be for the efficacy claim. Consider, for example, the critical value for the
intersection hypothesis of no broad efficacy nor efficacy in any of the subgroups. Suppose the sample sizes are equal for each of the subgroups, the observations are normally distributed with equal and known variances, and prevalence is the same for each of the subgroups. Then, at the one-sided FWER 5% level, as the number of subgroups tested increases from 0 to 3, critical values are 1.645, 1.875, 2.028, and 2.272. The computation of critical value for 3 subgroups is based on the correlation in Appendix D.

Second, increasing the number of subgroups decreases the sample size (and thus power to prove efficacy) for each subgroup. Suppose observations from the three subgroups have true means $\delta + \epsilon, \delta, \delta - \epsilon$, variances one, and $n$ samples are taken from each subgroup. Then, in a power calculation, the non-centrality parameter for the $g1$ group is $\sqrt{n}(\delta + \epsilon)$, and $\sqrt[3]{3n}\delta$ for the three subgroups combined. So, even without the consideration of multiplicity adjustment of critical values, the efficacy claim for the $g1$ subgroup is only more likely than for the three subgroups combined if $\epsilon > .732\delta$. This reduced sample size for subgroup issues can be seen in the examples in section 3.2.3.

In addition to the situations of a highly polymorphic biomarker, other situations that lead to more than two subgroups include when there is more than one biomarker, or when there is more than one diagnostic test for a single biomarker. In these situations, due to the considerations above, instead of testing three or more subgroups simultaneously, we recommend testing subgroups in a nested fashion, as described in the next section.
4.2 Testing efficacy in nested genetic subgroups

Consider two binary biomarkers, a primary biomarker which classifies patients into a \( g_P^+ \) subgroup and its complementary subgroup \( g_P^- \), and a secondary biomarker which classifies patients into a \( g_S^+ \) subgroup and its complementary subgroup \( g_S^- \). A similar situation is when there is a biomarker with two diagnostic tests, a primary diagnostic test which classifies patients into a \( g_P^+ \) subgroup and its complementary subgroup \( g_P^- \), and a secondary diagnostic test which classifies patients into a \( g_S^+ \) subgroup and its complementary subgroup \( g_S^- \). In testing breast cancer patients for HER2/neu status (for indication of Herceptin), for example, it is common to use an IHC (immunohistochemistry) test as a primary diagnostic test, and a FISH (fluorescence in situ hybridization) test as a secondary diagnostic test.

Suppose a study is being conducted to test for broad efficacy, efficacy for patients in the \( g_P^- \) subgroup, and efficacy for patients who are both \( g_P^- \) and \( g_S^- \). Let \( \bar{\theta}, \theta_{g_P^+}, \) and \( \theta_{g_P^+g_S^+} \) denote efficacy of the drug in the entire patient population, in the \( g_P^+ \) subgroup, and for patients who are both \( g_P^+ \) and \( g_S^+ \). The null hypotheses of insufficient efficacy to test are then

\[
H_0 : \bar{\theta} \leq \delta \quad \text{vs.} \quad H_a : \bar{\theta} > \delta \quad (4.1)
\]

\[
H_{0+} : \theta_{g_P^+} \leq \delta \quad \text{vs.} \quad H_{a+} : \theta_{g_P^+} > \delta \quad (4.2)
\]

\[
H_{0++} : \theta_{g_P^+g_S^+} \leq \delta \quad \text{vs.} \quad H_{a++} : \theta_{g_P^+g_S^+} > \delta \quad (4.3)
\]

If one strongly anticipates that patients in the \( g_P^- \) subgroup derive more benefit from the drug than patients in the \( g_P^- \) subgroup and, within patients in the \( g_P^+ \) subgroup, \( g_S^+ \) patients derive more benefit from the drug than patients who are \( g_S^- \), then testing in the order of (4.3) \( \rightarrow \) (4.2) \( \rightarrow \) (4.1) has intuitive appeal.
It turns out that, without assuming $\bar{\theta} \leq \theta_{g_p^+} \leq \theta_{g_p g_S^+}$, testing in the order of $(4.3) \rightarrow (4.2) \rightarrow (4.1)$ can be done without multiplicity adjustment while controlling FWER, as follows.

Step 1 If $H_{0++} : \theta_{g_p^+ g_S^+} \leq \delta$ is rejected at level-$\alpha$, then infer $\theta_{g_p g_S^+} > \delta$ and go to Step 2, else stop;

Step 2 If $H_{0+} : \theta_{g_p^+} \leq \delta$ is rejected at level-$\alpha$, then infer $\theta_{g_p^+} > \delta$ and go to Step 3, else stop;

Step 3 If $H_0 : \bar{\theta} \leq \delta$ is rejected at level-$\alpha$, then infer $\bar{\theta} > \delta$.

To effect testing in the order of $(4.3) \rightarrow (4.2) \rightarrow (4.1)$ and prove that such testing without multiplicity adjustment controls FWER, Partition testing makes each Partition null hypothesis in a subsequent step disjoint from null hypotheses in previous steps, by removing from it the union of the null hypotheses whose rejections lead up to it. This is an example of how Partition testing ensures that statistical inferences respect decision path logic. Details are given in Appendix C.

Not needing multiplicity adjustment, pre-determined steps testing is more powerful when the anticipated ordering bears out. However, with such testing, the decision process stops when a step fails to reject its null hypothesis. If this is a concern, then (4.1), (4.2), and (4.3) can be tested with equal priority, as in Wang, Hung, and O’Neill (2009) and Spiessens and Debois (2010).
Chapter 5: Conclusion

5.1 Concluding remarks

Partition testing was discovered independently by authors on three continents in the ’70s and ’80s, the earliest being Chapter 8 of Takeuchi (1973). See Takeuchi (2010) and references therein. Given its name by Finner and Strassburger (2002), its earlier use included insights to confidence sets for step-wise tests (Stefansson, Kim, and Hsu 1988, Hsu and Berger 1999) and bioequivalence tests (Berger and Hsu 1996). As multiple testing problems become more complex, testing multiple doses in combination with multiple endpoints for example, the advantage Partition testing has to readily ensure logic in decision paths becomes practically important (Liu and Hsu 2009). For the emerging problem of testing for broad efficacy and in genetic subgroups, Partition testing has the additional advantage that it readily recognizes logical relationships among the parameters defining broad efficacy and efficacy in the subgroups. Most important, Partition testing always can pivot the test to a confidence set which provides transparent information to resolve the consistency issue in subgroup analysis.
5.2 Limitation

Drugs traditionally are approved based on "average" effect, comparing treatment versus control. However, the concept of average efficacy only works when efficacy is defined in terms of the DIFFERENCE of expectations (between treatment and control). If efficacy is defined in terms of ratios, for example, relative risk, hazard ratio, odds ratio, the concept of average efficacy becomes limited as we discussed in section 2.4.3.

Let us define efficacy by comparing the odds ratio between treatment and control in survival data. In practice, it is a common parameter for binary data from a logistic regression model. To achieve a paradox-free condition, the clinical design needs to achieve a row-column-uniformity, which is not a reasonable design in subgroup analysis. In addition, there is no way to define an average of odds ratios that can represent the population average.

Even though the endpoint is the difference of expectations, the concept of average is hard to apply when more subgroups are involved in the analysis. Suppose there are $k$ binary biomarkers. Then the number of subgroups is $2^k$. Statistical learners talk about penalty, as in penalized regression such as LASSO. But their concept of penalty is abstract, not connecting directly to intended use.

In a paper on microarrays [15], the authors have been aware that the more markers, the fewer times each is probed on a diagnostic device (decreasing PPV and NPV). We have realized another PENALTY is that the more markers the smaller the patient sample size in each subgroup. Therefore, the more subgroups, the smaller sample size in each subgroup. The ultimate subgroup is the individual. A major motivation for the average efficacy approach is that sample size for a combined group is larger. The
sample size issue is not just cost, it is also difficult to find patients to fill a clinical trial.

5.3 Future research

Average is not "personal". To be personal, we should give each patient $P\{\text{response|biomarkervalues}\}$, which is PPV. For future research, instead of thinking how to improve the performance of multiple testing methods or to reduce testing null hypotheses, it is probably better to review how we can benefit “patients” most. From patients’ points of view, they might want to know whether a drug can be a benefit, especially for their genetic subgroup but not for the entire patient population.

Cases so far have involved only 2 subgroups, the marker $g^+$ group and its compliment the marker $g^-$ group. When more and more subgroups are involved in the analysis, one could formulate the problem as finding the largest union of subgroups for which average efficacy can be shown.

In Tu and Hsu (2011) [30], Multiple Comparisons with the Best (MCB) invented by [13] is applied to show which genetic subgroup derives maximum efficacy from a drug. For example, in a three subgroups’ setting, the three null hypotheses can be formulated by MCB as follows:

$$H_{01} : \theta_{g_1} - \max\{\theta_{g_2}, \theta_{g_3}\} \leq 0(\text{or } -\delta)$$

$$H_{02} : \theta_{g_2} - \max\{\theta_{g_1}, \theta_{g_3}\} \leq 0(\text{or } -\delta)$$

$$H_{03} : \theta_{g_3} - \max\{\theta_{g_1}, \theta_{g_2}\} \leq 0(\text{or } -\delta)$$

MCB only tests three null hypotheses in the three subgroups’ setting. By constructing MCB confidence intervals, one can provide useful inference about which subgroup is
one of the best, which subgroup is close to the best or even how much subgroups identified not to be the best are worse than the true best.

For future work, a decision path should be involved in this problem setting since MCB can only show the contrast between two subgroups. Therefore, a next step might be required to verify whether the treatment effects in the best subgroup or in the subgroups close to the best are efficacious. How to control FWER in the new problem setting will be developed in future work.
Appendix A: \( \text{max}T \) and \( \text{min}P \) tests

Let \( T_i, i = 1, \ldots, k \), be test statistics for testing the original hypotheses \( H_{0i}, i = 1, \ldots, k \). A Partition or Closed test is of \( \text{max}T \) form if the rejection rule for \( H_{0i}^* \) is

\[
\text{Reject } H_{0i}^* \text{ if } \max_{i \in I} T_i > d_I.
\]

Let \( p_i, i = 1, \ldots, k \), be p-values for testing the original hypotheses \( H_{0i}, i = 1, \ldots, k \). A Partition or Closed test is of \( \text{min}P \) form if the rejection rule for \( H_{0i}^* \) is

\[
\text{Reject } H_{0i}^* \text{ if } \min_{i \in I} p_i < c_I.
\]
Appendix B: Hochberg’s method

Let \( p_i, \ i \in I \subseteq \{1, 2, \ldots, k\} \), be the subset of the \( p \)-values with indices in \( I \). Let 
\([1]_I, \ldots, [|I|]_I\), where \(|I|\) is the number of elements in \( I \), denote the random indices such that

\[
p_{[1]} \leq \cdots \leq p_{|[I]|}.
\]

Huang and Hsu (2007) showed that, if Closed testing is executed by testing each \( H_{0I} \) using what they call the Simes-Hochberg test:

\[
\text{reject } H_{0I} \text{ if } p_{[i]} < \frac{\alpha}{|I| - i + 1} \text{ for some } i,
\]

then the Closed test has a step-up shortcut which is Hochberg’s method, with the following familiar description.

Let \( p_i, \ i = 1, \ldots, k \), denote the sample \( p \)-values of tests for \( H_{0i} \), \( i = 1, \ldots, k \), computed without multiplicity adjustment. Let \([1], \ldots, [k]\) denote the random indices such that

\[
p_{[1]} \leq \cdots \leq p_{[k]}.
\]

That is, \([i]\) is the anti-rank of \( p_i \) among \( p_1, \ldots, p_k \). Then Hochberg’s step-up method proceeds as follows.

1. **Step 1.** If \( p_{[k]} < \alpha \), reject \( H_{0[i]} \), \( i = 1, \ldots, k \), and stop; otherwise go to Step 2.
Step 2. If $p_{[k-1]} < \alpha/2$, reject $H_{0[i]}$, $i = 1, \ldots, k - 1$, and stop; otherwise go to Step 3.

\[ \ldots \]

Step k. If $p_{[1]} < \alpha/k$, reject $H_{0[i]}$, $i = 1$, and stop.
Appendix C: Partitioning decision paths

Suppose decision-making concerning the null hypotheses $H_{0i} : \theta \in \Theta_i$, $i = 1, \ldots, k$, follows a decision path in the sense that the null hypothesis $H_{0i'}$ is not tested unless all $H_{0i}, i < i'$, are rejected. This is referred to as fixed sequence testing or step-wise testing with pre-determined steps. (An example would be to test efficacy of a drug from high dose to low dose.) Decision Path (logic) Partition testing in Hsu and Berger (1999) partitions the parameter space into $k + 1$ subspaces $\tilde{\Theta}_j, i = 1, \ldots, k + 1$, where

\[
\tilde{\Theta}_j = \begin{cases} 
\Theta_j, & j = 1 \\
\Theta_j \cap (\bigcup_{s=1}^{j-1} \Theta_s)^c, & j = 2, \ldots, k \\
(\bigcup_{s=1}^{j-1} \Theta_s)^c, & j = k + 1
\end{cases}
\]

Thus, to effect pre-determined steps testing of (4.3), (4.2), (4.1) without any assumption on the ordering of $\overline{\theta}, \theta_{g_p^+}, \theta_{g_p g_s^+}$ and prove control of FWER, partition the null hypotheses (4.3), (4.2), (4.1) as follows.

\[
H_{0++}^* : \theta_{g_p g_s^+} \leq \delta \\
H_{0^+}^* : \theta_{g_p^+} \leq \delta \cap \theta_{g_p g_s^+} > \delta \quad \text{(C.1)}
\]

\[
\bar{H}_0^* : \overline{\theta} \leq \delta \cap \theta_{g_p^+} > \delta \cap \theta_{g_p g_s^+} > \delta
\]
Since these Partition null hypotheses are disjoint, at most one null hypothesis is true. Therefore, FWER of testing (C.1) is controlled even when each null hypothesis in (C.1) is tested at level-\( \alpha \) without multiplicity adjustment.

Note that a level-\( \alpha \) test for \( H_{0+,} \) is a level-\( \alpha \) test for \( H_{0++}^* \), a level-\( \alpha \) test for \( H_0 \) is a level-\( \alpha \) test for \( H_0^* \). Further,

\[
H_{0++}^* = H_{0++}
\]

\[
H_{0++}^* \cup H_{0+}^* = H_{0++} \cup H_{0+}
\]  \hspace{1cm} (C.2)

\[
H_{0++}^* \cup H_{0+}^* \cup \bar{H}_0^* = H_{0++} \cup H_{0+} \cup \bar{H}_0
\]  \hspace{1cm} (C.3)

A rejection at Step 1 obviously implies one can reject \( H_{0++} \). With a rejection at Step 2, (C.2) shows one can further reject \( H_{0+,} \). Finally, with a rejection at Step 3, (C.3) shows one can reject \( \bar{H}_0 \) as well.

<table>
<thead>
<tr>
<th></th>
<th>Reject ( H_{0++} )</th>
<th>Reject ( H_{0+} )</th>
<th>Reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{0++}^* )</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>( H_{0+}^* )</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>( \bar{H}_0^* )</td>
<td></td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
</tbody>
</table>

Table C.1: Partition testing (P) decision rules for rejecting \( \bar{H}_0 \), \( H_{0+} \), or \( H_{0+,} \). A \( \checkmark \) denotes a null hypothesis that needs to be rejected.

The above shows how to partition with one decision path, now suppose the null hypotheses \( H_{0ij} : \theta \in \Theta_{ij}, i = 1, \ldots , k, j = 1, \ldots, m \), are in \( k \) parallel paths, and in the \( i^{th} \) path, the null hypothesis \( H_{0ij'} \) is not tested unless all \( H_{0ij}, j < j' \), are rejected. (An example would be testing \( k \) un-ordered doses, and \( m \) ordered endpoints.) The idea in Liu and Hsu (2009) is to follow the path in partition the null hypotheses within
each path, and then proceed as in (1.2). Thus Decision Path partitioning, within the $i^{th}$ path, partitions the parameter space into $m + 1$ subspaces $\tilde{\Theta}_{ij}$ where

$$
\tilde{\Theta}_{ij} = \begin{cases} 
\Theta_{ij}, & j = 1 \\
\Theta_{ij} \cap \left( \bigcup_{s=1}^{j-1} \Theta_{is} \right)^c, & j = 2, \cdots, m \\
\left( \bigcup_{s=1}^{j-1} \Theta_{is} \right)^c, & j = m + 1 
\end{cases}
$$

(In the dose-endpoints example, within each dose, the second null hypothesis becomes “there is efficacy in the primary endpoint, but no efficacy in the secondary endpoint”, while the third null hypothesis becomes “there is efficacy in the primary and secondary endpoints, but no efficacy in the tertiary endpoint”, etc.) Let $j = (j_1, \ldots, j_k)$ denote a $k$-tuple of integers in $\{1, \ldots, m + 1\}$. Then, in applying the partitioning technique analogous to (1.2), since the null hypotheses are now disjoint within each path, one finds that the $(m + 1)^k$ hypotheses

$$
H_{0j} : \bigcap_{i=1}^{k} \{ \theta_{ij} \in \tilde{\Theta}_{ij}, j \in \{1, \cdots, m + 1\} \}^k
$$

partition the parameter space. (In the dose-endpoints example, for each dose, a Partitioning null hypothesis either assigns “true” to one endpoint, or assigns “false” to all endpoints.)
Appendix D: Correlation for three subgroups

Assume

1. \( g_1, g_2 \) and \( g_3 \) subgroups are disjoint.

2. The prevalence for \( g_1, g_2 \) and \( g_3 \) biomarkers are 1 : 1 : 1.

3. \( \hat{\theta}_{g_1} \sim \text{Normal}(\theta_{g_1}, 1) \)
   \( \hat{\theta}_{g_2} \sim \text{Normal}(\theta_{g_2}, 1) \)
   \( \hat{\theta}_{g_3} \sim \text{Normal}(\theta_{g_3}, 1) \)

Then the joint distribution of the standardized test statistics \( T_1 = \hat{\theta}_{g_1}, T_2 = \hat{\theta}_{g_2}, T_3 = \hat{\theta}_{g_3} \), \( \bar{T}_{12} = \sqrt{2} \times \hat{\theta}_{12} = (\hat{\theta}_{g_1} + \hat{\theta}_{g_2})/\sqrt{2}, \bar{T}_{13} = \sqrt{2} \times \hat{\theta}_{13} = (\hat{\theta}_{g_1} + \hat{\theta}_{g_3})/\sqrt{2}, \bar{T}_{23} = \sqrt{2} \times \hat{\theta}_{23} = (\hat{\theta}_{g_2} + \hat{\theta}_{g_3})/\sqrt{2} \) and \( \bar{T} = \sqrt{3} \times \hat{\theta} = (\hat{\theta}_{g_1} + \hat{\theta}_{g_2} + \hat{\theta}_{g_3})/\sqrt{3} \) is multivariate Normal, with variance-covariance matrix

\[
\sum = \begin{bmatrix}
1 & 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} & 0 & 1/3 \\
0 & 1 & 0 & 1/\sqrt{2} & 0 & 1/\sqrt{3} & 1/3 \\
0 & 0 & 1 & 0 & 1/\sqrt{2} & 1/\sqrt{3} & 1/3 \\
1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} & 0 & 1 & 1/2 & 1/2 \sqrt{2/3} \\
1/\sqrt{2} & 0 & 1/\sqrt{2} & 1/2 & 1 & 1/2 \sqrt{2/3} & 1/3 \sqrt{2/3} \\
0 & 1/\sqrt{2} & 1/\sqrt{2} & 1/2 & 1/2 & 1 \sqrt{2/3} & \sqrt{2/3} \\
1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} & \sqrt{2/3} & \sqrt{2/3} & \sqrt{2/3} & 1
\end{bmatrix}
\]
Bibliography


