TESTING FOR EFFICACY IN PRIMARY AND SECONDARY ENDPOINTS BY PARTITIONING DECISION PATHS

DISSERTATION

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

By

Yi Liu

Graduate Program in Statistics

The Ohio State University
2009

Dissertation Committee:

Professor Jason Hsu, Advisor

Professor Elizabeth Stasny

Professor Eloise Kaizar

© Copyright by

Yi Liu

2009

ABSTRACT

Testing for efficacy in multiple endpoints has emerged as an important statistical problem. The Food and Drug Administration (FDA) will issue a guidance on Multiple Endpoints in the near future.

When there are primary and secondary endpoints, efficacy in the secondary endpoint is only relevant if efficacy in the primary endpoint has been shown. There are thus defined paths to decision-making.

The current approach to this problem is based on closed testing, that is, testing all possible intersection hypotheses, and collating the results. For decision-making to follow pre-defined paths, strategic choices of test statistics and critical values must be made. As the number of doses and endpoints increases, such strategic choices become increasingly difficult.

Partition testing is an alternative to closed testing. It gives insights on confidence sets for step-wise tests, and can be more powerful than closed testing. It can also simplify problem formulation when decision-making follows specific paths. We show, for the primary-secondary endpoints problem, that partition testing has advantages. Using it to implement what we call the Decision Path Principle, we find that partition testing not only drastically reduces the number of hypotheses to be tested, but also guides decision-making along pre-defined paths. With our way of setting critical

values, we achieve higher probabilities of correctly inferring efficacious primary endpoints as efficacious compared to gatekeeping methods, while maintaining the same level of strong FWER control. These advantages are illustrated with a real data example, and by simulation.

Dedicated to my family

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Professor Jason Hsu, for his constant support and encouragement.

I would like to thank Professor Elizabeth Stasny and Professor Eloise Kaizar for their help.

I am thankful for the Department of Statistics for their support.

I am grateful to my family, especially my husband Zhi Qiu, for their constant support and encouragement.

VITA

July, 1981 Born - Kunshan, China

2004 B.S. in Statistics,
University of Science and Technology
of China, China

2004-2005 Graduate Teaching Associate,
University of Iowa

2005-present Graduate Teaching Associate and Re-
search Associate,
The Ohio State University

PUBLICATIONS

Research Publications

J.C. Hsu and Y. Liu “Discussion of ‘some controversial multiple testing problems in regulatory applications’ ”. *Journal of Biopharmaceutical Statistics*, 19:22–24, 2009.

F. Bretz, J.C. Hsu, J. Pinheiro, and Y. Liu “Dose finding – a challenge in statistics”. *Biometrical Journal* , 50:480–504, 2008.

Y. Liu, J.C. Hsu, and S. Ruberg “Partition testing in dose-response studies with multiple endpoints”. *Pharmaceutical Statistics*, 6:181–192, 2007.

FIELDS OF STUDY

Major Field: Statistics

TABLE OF CONTENTS

	Page
Abstract	ii
Dedication	iv
Acknowledgments	v
Vita	vi
List of Tables	x
List of Figures	xii
Chapters:	
1. Using the partition principle to control multiple testing error rates	1
1.1 FWER, gFWER, E(V), FDR, and Fdr	1
1.2 Partition principle controlling FWER	4
1.3 Generalized partition principle controlling gFWER	7
1.4 E(V) control using the generalized partition principle	9
1.4.1 One-step procedure	10
1.4.2 A step-down procedure	13
1.4.3 A step-up procedure	15
1.5 FDR control using the generalized partition principle	18
1.5.1 Three hypotheses example	20
1.5.2 Step-down shortcut	22
1.5.3 Comparison with other FDR control methods	29

2.	Partition testing in dose-response studies with multiple endpoints	31
2.1	Historical aspect of dose-response trials in pharmaceuticals	31
2.2	A motivating example for dose-response studies	33
2.3	Multiple tests construction using the partitioning principle	36
2.3.1	Closed testing to step-down doses according to sample re- sponses	36
2.3.2	Partition to step-down doses	37
2.4	Partition testing with two doses and two endpoints	40
2.4.1	Partition to step-down endpoints	40
2.4.2	Partitioning to step-down both doses and endpoints	42
2.5	Analysis of the anti-psychotic drug data	49
2.6	Conclusion	51
2.7	Appendix: Model for dose-response studies with two endpoints . .	52
3.	Decision Path Approach	54
3.1	Primary-secondary endpoints with unordered doses	54
3.1.1	Closed testing	54
3.1.2	Iterative graphical method	55
3.1.3	Partition testing	56
3.2	A motivating example	57
3.3	Decision path principle	57
3.4	Paths with ordered endpoints	59
3.4.1	Partition following decision paths	63
3.4.2	Advantages of partition testing	65
3.4.3	Execution of partition testing	66
3.5	Computer algorithm	67
3.5.1	Computation within and between endpoints	69
3.5.2	Recursive computation between endpoints	70
3.6	Performance comparisons	73
3.7	Analysis of motivating example	74
3.8	Discussion	78
3.9	Appendix	80
3.9.1	Appendix A: Proof of Theorem 3.4.1	80
3.9.2	Appendix B: Steps in computing critical values	81
4.	Work in progress and future research	83
4.1	Stability of V	83
4.1.1	$\text{Var}(V)$ increases with correlation for the one-step procedure	84

4.1.2	Comparing the one-step and BH step-up approaches	86
4.2	Non-inferiority and superiority tests	89
4.2.1	Paths between non-inferiority and superiority	90
4.2.2	Additional paths between doses for the non-inferiority test	92
4.2.3	Non-inferiority and superiority tests with two doses and two endpoints	96
4.3	Co-primary endpoints problem	98
4.3.1	Problem setting	101
4.3.2	Predetermined steps	102
4.3.3	Motivation	103
4.3.4	Pre-specifying d for the primary endpoint	104
4.3.5	Calculating critical values	105
4.4	Finding Minimum Effective Dose using confidence bounds in dose-response studies	108
4.4.1	Preliminaries	109
4.4.2	Finding MED using confidence bounds without multiplicity adjustment	112
4.4.3	Appendix	113

Appendices:

A.	R function “DoseEndpoints” implementing Decision Path Approach	115
A.1	Documentation	115
A.2	Sample Code	127
A.3	R codes for “DoseEndpoints”	135
	Bibliography	157

LIST OF TABLES

Table		Page
1.1	Results for testing k null hypotheses	2
1.2	Critical values for the step-down procedure controlling $E(V)$	15
1.3	Critical values for the step-up procedure controlling $E(V)$ in two dimensional case	18
1.4	Critical values in each partitioned subspace	21
1.5	Setting critical values for the test statistics in each partitioned subspace for testing three hypotheses	25
1.6	Critical values for the step-down procedure and two other methods controlling FDR at 0.05 for testing three hypotheses under independence	30
2.1	Summary statistics of two doses and two endpoints from Arvanitis et al. (1997)	34
2.2	Decision rules for closed testing with two doses and single primary endpoint	37
2.3	Decision rules for partition testing to step-down doses	39
2.4	Direction of the pre-determined DE-steps	43
2.5	Critical values for the pre-determined DE-steps	44
2.6	Comparing conservatism in terms of critical value c_2	50
3.1	Summary statistics for the hypertension example	57

3.2	Partition hypotheses following decision paths in Figure 3.3	66
3.3	Scenarios of mean matrix for simulation results in Table 3.4	74
3.4	Probability of correctly rejecting the individual hypotheses from three procedures: DPjd proposed in this chapter, GKjj in Dmitrienko et al. (2006) and GKjd in Xu et al. (2008) in four scenarios described in Table 3.3 with correlation $\rho = 0.3, 0.5, 0.7$ and 0.9	75
3.5	Partition hypotheses for hypertension example with rejection rules . .	76
3.6	Critical values in Table 3.5 using independence or disjointness approximation with strong FWER control at 0.05	77
4.1	Distribution of the number of false rejections for the one-step and BH step-up procedure	88
4.2	Partition hypotheses from a stage II disjointness partition for decision paths in Figure 4.4	95

LIST OF FIGURES

Figure	Page
1.1 Critical values c_2 and c_1 plotted as functions of d for the step-down procedure controlling FDR at $\alpha = 0.05$ level.	29
2.1 Decision process in Dmitrienko et al. (2006)	41
2.2 Decision tree for pre-determined DE-steps	45
3.1 Decision paths for k doses and m endpoints, with one path for each dose going from endpoint 1 (primary endpoint) to endpoint m	60
3.2 Examples of nine (out of 63) partition hypotheses with three doses (low, medium, high) and three endpoints (P, S1, S2). Each partition hypothesis is represented by a non-empty intersection with at most one box (light colored with “ $\leq$ ” sign) from each column. For example, the partition hypotheses in the left column from top to bottom are: $\{\theta_{11} \leq \delta_1 \text{ and } \theta_{21} \leq \delta_1 \text{ and } \theta_{31} \leq \delta_1\}$, $\{\theta_{11} > \delta_1 \text{ and } \theta_{12} \leq \delta_2 \text{ and } \theta_{21} \leq \delta_1 \text{ and } \theta_{31} > \delta_1 \text{ and } \theta_{32} \leq \delta_2\}$, $\{\theta_{11} > \delta_1 \text{ and } \theta_{12} > \delta_2 \text{ and } \theta_{13} > \delta_3 \text{ and } \theta_{21} \leq \delta_1 \text{ and } \theta_{31} \leq \delta_1\}$	62
3.3 Decision paths for low and high doses with one primary and one secondary endpoint.	63
3.4 Graphical representation of two stages of partitioning in the setting of Figure 3.3.	65
3.5 Critical values for the partition hypotheses in Figure 3.2 using two sample t statistics with 57 degrees of freedom, controlling FWER strongly at 0.05.	72
4.1 Rejection region of the one-step procedure (left) and the BH step-up procedure (right).	88

4.2	Probability distribution of the number of false rejections for the one-step procedure (left) and the BH step-up procedure (right).	89
4.3	Decision paths for k doses, with paths from non-inferiority to superiority for each dose.	91
4.4	Decision paths for the non-inferiority (NI) and superiority (SUP) test with three doses. Paths are from high to low for the NI test, and from NI to SUP for each dose.	92
4.5	Decision tree from stage I: path partition for the decision paths problem in Figure 4.4.	94
4.6	Graphical representation of the decision paths (O1, O2, and O3) for non-inferiority and superiority tests with two doses (low, and high) and two endpoints (primary, and secondary).	98
4.7	Decision tree from stage I: path partition, for the decision paths problem in Figure 4.6	99
4.8	Rejection region of pre-determined steps (left) and of the new proposed method (right) for testing $H_0^{S*} : \theta^S \leq \delta^S$ and $\theta^P > \delta^P$	105
4.9	Plot of c_2 as function of c_1 for $d = 2.1$ in blue and $d = 2.5$ in red. . .	108
A.1	Mean change from baseline for placebo (P), five dose groups (Qtp75, Qtp150, Qtp300, Qtp600, Qtp750) and active control group (A) in BPRS total score, CGI Severity of Illness score, and SANS summary score. The two dotted horizontal lines in each graph are mean change of placebo group plus a negative delta (for superiority) and mean change of active control plus a positive delta (for non-inferiority).	120
A.2	Mean change from baseline in systolic and diastolic blood pressure. The two dotted horizontal lines are mean change for placebo group plus delta (-0.4 for systolic blood pressure and -0.2 for diastolic blood pressure).	124

CHAPTER 1

USING THE PARTITION PRINCIPLE TO CONTROL MULTIPLE TESTING ERROR RATES

In the multiple testing setting, different measures of errors have been proposed in various scenarios. In this chapter, we show that the fundamental tool of the partition principle or its generalizations can be used to control different types of multiple testing error rates.

1.1 FWER, gFWER, $E(V)$, FDR, and Fdr

In this section, we introduce the definitions of these five different error rates in multiple testing, and differentiate them in terms of exceedance (tail) probability or expectation error rate, conditional or unconditional error rate. Then we discuss their potential uses in the drug efficacy study setting.

Suppose we test k null hypotheses, k_0 of which are actually true. Let V be the number of falsely rejected true nulls, and R be the total number of rejected hypotheses as described in Table 1.1.

In words, the Familywise Error Rate (FWER) is the probability of making any false rejection. The generalized Familywise Error Rate (gFWER) represents the probability of making more than $m(> 0)$ false rejections. $E(V)$ is the expected number of

Table 1.1: Results for testing k null hypotheses

	Reject	Fail to reject	Total
True null hypotheses	V	$k_0 - V$	k_0
False null hypotheses	S	$k_1 - S$	k_1
All null hypotheses	R	$k - R$	k

false rejections. The False Discovery Rate (FDR) is the expected proportion of false rejections. While Fdr is a conditional version of FDR.

The mathematical definitions of these error rates are:

- $\text{FWER} = P(V > 0)$ - It is an unconditional tail probability of the distribution of the random variable V .
- $\text{gFWER} = P(V > m)$ - It is an unconditional tail probability of the distribution of V , only with a different cut-off value compared to FWER.
- $E(V)$ - It is an unconditional expectation of the random variable V .
- $\text{FDR} = E \left[\frac{V}{R} I(R \neq 0) \right] (= E \left[\frac{V}{R} I(V > 0) \right])$ - It is an unconditional expectation of the random variable V/R when it is defined.
- $\text{Fdr} = E(V)/r$ - Since one can observe the total number of rejections $R = r$, some people choose to estimate the quantity $E(V)/r$ for a specific experiment. In principle, $E(V)/R$ is a random variable, and hence cannot be controlled at a pre-specified level. Researchers usually try several critical values such that the estimated Fdr for that particular experiment is slightly under α . Fdr has a conditional flavor but is different from the conditional expectation of V/R , since $E(V/R|R = r) = E(V|R = r)/r$, for $r > 0$.

In dose response studies with multiple endpoints, FWER is commonly used to guard against making any false claims on the efficacy of any endpoint at any dose level. As secondary endpoints only provide additional information about the target treatment, it might be possible to make one or two mistakes in the secondary endpoints while still preserving the integrity of efficacy demonstration of the treatment. Thus we think when testing for efficacy in both primary and secondary endpoints, gFWER may be used to control the probability of making more than m false efficacy claims in the secondary endpoints.

Controlling FDR is not appropriate in the drug efficacy setting with multiple doses because one can “cheat”. Suppose k doses are tested in a trial. If $k - 1$ of the doses are highly efficacious such that each has the probability of rejection close to 1, then the other dose can be tested with a Type I error rate higher than α while FDR is still controlled at α . This can be seen in the parameter configuration $\{H_{0k}$ is true; $H_{01}, \dots, H_{0(k-1)}$ are false, so false that they will almost surely be rejected $\}$. A testing procedure controlling FDR at α has

$$\alpha \approx \frac{0}{k-1} \Pr(V=0) + \frac{1}{k} \Pr(V=1) = \frac{1}{k} \Pr(\text{Reject } H_{0k}).$$

H_{0k} is tested at $k\alpha$. See Finner and Roter (2001) for detailed examples.

The same situation occurs for Fdr. If a total rejection of $R = k - 1$ hypotheses is observed, then

$$\text{Fdr} = \frac{E(V)}{k-1} = \frac{0 * P(V=0) + 1 * P(V=1)}{k-1} = \frac{1}{k-1} \Pr(\text{Reject } H_{0k}).$$

If $R = k$ then

$$\text{Fdr} = \frac{E(V)}{k} = \frac{0 * P(V=0) + 1 * P(V=1)}{k} = \frac{1}{k} \Pr(\text{Reject } H_{0k}).$$

Either way, the probability of rejecting H_{ok} is inflated by $k - 1$ or k fold. Therefore, using Fdr is not appropriate in the drug efficacy setting either.

When simultaneously testing a large number of hypotheses, for example, analyzing microarray data, FWER is generally regarded as too “stringent” because rejecting one or two true null hypotheses does not seem to be too crucial. FDR, Fdr, and E(V) control are most commonly used in such settings.

1.2 Partition principle controlling FWER

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 (1979) 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 Bonferroni’s inequality and a modification of Simes’ (1986) inequality to component partitioning tests (see Huang and Hsu, 2007). 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).

The idea of partition testing is if, in a subspace of the parameter space only a subset of the null hypotheses are true, then tests for parameter points in that subspace need to be adjusted for the multiplicity of only those true null hypotheses. Partition testing thus applies a *different* test to each subspace of the parameter space depending on how many null hypotheses are true in that subspace, and then collates the results across the subspaces, as follows.

Let $K = \{1, \dots, k\}$ and consider testing $H_{0i} : \theta_i \in \Theta_i, i = 1, \dots, k$. The parameter of interest in this case is $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ and the parameter space is $\bigcup_{i \in K} \Theta_i$. To control FWER (gFWER with $m = 0$), the Partitioning Principle states:

1. Partition $\bigcup_{i \in K} \Theta_i$ into disjoint Θ_I^* , $I \subseteq K$, as follows. For each $I \subseteq K$, let $\Theta_I^* = \bigcap_{i \in I} \Theta_i \cap (\bigcap_{j \notin I} \Theta_j^c)$. Then $\{\Theta_I^*, I \subseteq K\}$, including $\Theta_\emptyset^* = \bigcap_{j \in K} \Theta_j^c$, partition the parameter space. Note that Θ_I^* can be interpreted as the part of the parameter space in which exactly H_{0i} , $i \in I$, are true, and H_{0j} , $j \notin I$ are false.
2. Test each $H_{0I}^P : \boldsymbol{\theta} \in \Theta_I^*$ at level α . Since the null hypotheses are disjoint, at most one null hypothesis is true. Therefore, even though no multiplicity adjustments to the levels of the tests are made, the probability of rejecting at least one true null hypothesis is at most α .
3. For all $I \subseteq K$, infer $\boldsymbol{\theta} \notin \Theta_I$ if all H_{0J}^P such that $I \subseteq J$ are rejected. That is, infer the intersection null hypothesis H_{0I} is false if all null hypotheses H_{0J}^P implying it are rejected. In terms of useful scientific inference of rejecting the original null hypotheses H_{0i} , $i \in K$, since $\Theta_i = \bigcup_{I \ni i} \Theta_I^*$, the partition test rejects H_{0i} if all H_{0I}^P such that $I \ni i$ are rejected.

For example, suppose the desired inferences are $\theta_1 > 0$, or $\theta_2 > 0$, or both. With the alternative hypotheses being $H_{a1} : \theta_1 > 0$ and $H_{a2} : \theta_2 > 0$, the proper null hypotheses are $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$.

Partitioning forms the hypotheses

$$H_{012}^P : \theta_1 \leq 0 \quad \text{and} \quad \theta_2 \leq 0$$

$$H_{01}^P : \theta_1 \leq 0 \quad \text{and} \quad \theta_2 > 0$$

$$H_{02}^P : \theta_1 > 0 \quad \text{and} \quad \theta_2 \leq 0$$

and tests each at level α . Since the three hypotheses are disjoint, at most one null hypothesis is true. Therefore, even though no multiplicity adjustments to the levels of the tests are made, the probability of rejecting at least one null hypothesis is at most α . Applying the Partitioning Principle, statistical inference proceeds as follows:

If $H_{012}^P : \theta_1 \leq 0$ and $\theta_2 \leq 0$ is not rejected, no inference is given.

If only $H_{012}^P : \theta_1 \leq 0$ and $\theta_2 \leq 0$ is rejected but not $H_{01}^P : \theta_1 \leq 0$ and $\theta_2 > 0$ or $H_{02}^P : \theta_1 > 0$ and $\theta_2 \leq 0$, the inference is that at least one of $H_{a1} : \theta_1 > 0$ or $H_{a2} : \theta_2 > 0$ is true, but which one cannot be specified.

If $H_{012}^P : \theta_1 \leq 0$ and $\theta_2 \leq 0$ and $H_{01}^P : \theta_1 \leq 0$ and $\theta_2 > 0$ are rejected but not $H_{02}^P : \theta_1 > 0$ and $\theta_2 \leq 0$, the inference is $\theta_1 > 0$.

If $H_{012}^P : \theta_1 \leq 0$ and $\theta_2 \leq 0$ and $H_{02}^P : \theta_1 > 0$ and $\theta_2 \leq 0$ are rejected but not $H_{01}^P : \theta_1 \leq 0$ and $\theta_2 > 0$, the inference is $\theta_2 > 0$.

If $H_{012}^P : \theta_1 \leq 0$ and $\theta_2 \leq 0$ and $H_{01}^P : \theta_1 \leq 0$ and $\theta_2 > 0$ and $H_{02}^P : \theta_1 > 0$ and $\theta_2 \leq 0$ are all rejected, the inference is both $\theta_1 > 0$ and $\theta_2 > 0$.

Partitioning inference is conditional in the sense that it uses the data to decide in which subspaces the true parameter may lie, and bases its inferences on tests appropriate for those subspaces. The Partitioning Principle has been successfully applied to derive practically useful methods for bioequivalence (Berger and Hsu, 1996), dose-response studies (Hsu and Berger, 1999), and genetic linkage analysis (Lin, Rogers, and Hsu, 2001).

The next section describes how the partitioning principle can be generalized to systematically construct methods controlling gFWER. A key difference between classical partitioning and generalized partitioning is that generalized partition testing must test *each individual* $H_{0i}, i \in K$ in each Θ_I^* .

1.3 Generalized partition principle controlling gFWER

The generalized partitioning principle states:

- P1. Partition the parameter space $\bigcup_{i \in K} \Theta_i$ into disjoint sets: for each $I \subseteq K = \{1, \dots, k\}$, let $\Theta_I^* = \bigcap_{i \in I} \Theta_i \cap (\bigcap_{j \notin I} \Theta_j^c)$.
- P2. In each Θ_I^* , test $\{H_{0i} : \theta_i \in \Theta_i, i \in I\}$ at gFWER level α controlling $\sup_{\theta \in \Theta_I^*} P_{\theta}(V > m) \leq \alpha$. Define H_{0i} to be *I-rejected* if it is rejected in Θ_I^* . *I-reject* all H_{0i} , $i \notin I$.
- P3. Reject H_{0i} if H_{0i} is *I-rejected* for all I , $I \subseteq K$.

The generalized partitioning principle first partitions the parameter space into disjoint sets, so that exactly one Θ_I^* contains the true parameter θ^* . It then controls gFWER by controlling gFWER within each Θ_I^* .

Consider, for example, $k = 3$, $m = 1$, $\Theta_i = \{\theta_i \leq 0\}$. Then there are $2^3 = 8$ Θ_I^* 's. Each $H_{0i} : \theta_i \leq 0$ is tested eight times, once in each Θ_I^* , and is ultimately rejected only if it is rejected in each Θ_I^* . Within each Θ_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 gFWER is controlled at level- α within Θ_I^* . For $I = \{2, 3\}$, say, H_{01} is automatically rejected while H_{02} and H_{03} are tested so that the probability of rejecting *both* is no more than α . If the test statistics are independent, then each can be tested at level- $\sqrt{\alpha}$ within $\Theta_{\{2,3\}}^*$. (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 α .)

To see that the generalized partitioning principle controls gFWER, suppose, for example, $\theta^* = (1, -1, -1)$. Then, with $m = 1$, gFWER is controlled at level α if the

probability of rejecting *both* H_{02} and H_{03} is no more than α . Since H_{02} and H_{03} will be rejected if they are I -rejected in Θ_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 α by P2 of the generalized partitioning principle.

A distinguishing feature of generalized partition testing controlling gFWER with $m > 0$ versus classical partition testing controlling gFWER with $m = 0$ is that generalized partition testing must test *each individual* $H_{0i}, i \in K$ in each Θ_I^* . While such tests are not unique, a reasonable way of testing is as follows.

Suppose T_i^I is a suitable test statistic for testing H_{0i} , $i \in I$. Let $|I|$ be the cardinality of I and let $[1], \dots, [|I|]$ be the indices such that $T_{[1]}^I \leq \dots \leq T_{[|I|]}^I$ for T_i^I , $i \in I$. Suppose the critical value c^I satisfies

$$\sup_{\theta \in \Theta_I^*} P_{\theta}(T_{[|I|-m]}^I > c^I) \leq \alpha. \quad (1.3.1)$$

Then the I -reject generalized partitioning test proceeds as follows:

1. In all Θ_I^* with $|I| \leq m$, I -reject all $H_{0i}, i \in I$.
2. In Θ_I^* with $|I| > m$,
 - I -reject $H_{[1]}^I, H_{[2]}^I, \dots, H_{[|I|-m+1]}^I$ for sure;
 - I -reject H_{0i}^I for all $i \in I$ if $T_{[|I|-m]}^I > c^I$;
 - Do not I -reject any $H_{0i}^I, i \in I \setminus \{[1], [2], \dots, [|I|-m+1]\}$, if $T_{[|I|-m]}^I \leq c^I$.
3. Reject H_{0i} if H_{0i} is I -rejected in all Θ_I^* s.t. $i \in I$.

Theorem 1.3.1. *The I -reject partitioning test controls gFWER strongly at level α .*

Proof (Theorem 1.3.1) Without loss of generality, let $\{H_{0i}, i \in K^*\}$ be the non-empty collection of true null hypotheses, that is, $\theta \in \Theta_{K^*}^*$. If $|K^*| \leq m$, then it is certain that gFWER is zero. For $|K^*| > m$, we show the I -reject partitioning test controls gFWER strongly at level α as follows.

$$\begin{aligned}
gFWER &= \sup_{\theta \in \Theta_{K^*}^*} P_{\theta}\{V > m\} \\
&= \sup_{\theta \in \Theta_{K^*}^*} P_{\theta}\{\text{reject strictly more than } m \text{ } H_{0i}\text{'s, } i \in K^*\} \\
&= \sup_{\theta \in \Theta_{K^*}^*} P_{\theta}\{I\text{-reject } H_{0i} \text{ under } \Theta_I^* \forall I \ni i, \\
&\quad \text{for strictly more than } m \text{ } H_{0i}\text{'s, } i \in K^*\} \\
&\leq \sup_{\theta \in \Theta_{K^*}^*} P_{\theta}\{I\text{-reject strictly more than } m \text{ } H_{0i}\text{'s under } \Theta_{K^*}^*\} \\
&= \sup_{\theta \in \Theta_{K^*}^*} P_{\theta}\{T_{[|K^*|-m]}^{K^*} > c^{K^*}\} \\
&\leq \alpha
\end{aligned}$$

To illustrate the proof, let $k = 3, m = 1$, and suppose $\theta^* \in \Theta_{12}^*$, then

$$\begin{aligned}
gFWER &= \sup_{\theta^* \in \Theta_{12}^*} P_{\theta^*}\{\text{reject } H_{01} \text{ and reject } H_{02}\} \\
&= P\{I\text{-reject } H_{01} \text{ in } \Theta_{12}^*, \Theta_{13}^*, \Theta_{123}^* \text{ and} \\
&\quad I\text{-reject } H_{02} \text{ in } \Theta_{12}^*, \Theta_{23}^*, \Theta_{123}^*\} \\
&\leq P\{I\text{-reject } H_{01} \text{ and } H_{02} \text{ in } \Theta_{12}^*\} \\
&\leq \alpha.
\end{aligned}$$

1.4 E(V) control using the generalized partitioning principle

This section shows that one can use the generalized partitioning principle to control the expected number of false rejections, $E(V)$. The generalized partitioning principle states:

P1. Partition the parameter space $\bigcup_{i \in K} \Theta_i$ into disjoint sets: for each $I \subseteq K = \{1, \dots, k\}$, let $\Theta_I^* = \bigcap_{i \in I} \Theta_i \cap (\bigcap_{j \notin I} \Theta_j^c)$.

P2. In each Θ_I^* , test $\{H_{0i} : \theta_i \in \Theta_i, i \in I\}$ at level α controlling $\sup_{\theta \in \Theta_I^*} E_{\theta}(V_I)$ where V_I is the number of false rejections in Θ_I^* . Define H_{0i} to be *I-rejected* if it is rejected in Θ_I^* . *I-reject* all H_{0i} , $i \notin I$.

P3. Reject H_{0i} if H_{0i} is *I-rejected* for all I , $I \subseteq K$.

Theorem 1.4.1. *Let V_I be the number of true null hypotheses *I-rejected* in Θ_I^* . Controlling $E(V_I)$ for testing each partition hypothesis $\Theta_I^*, I \subseteq K$, following the generalized partition principle P1-P3 guarantees the overall control of $E(V)$.*

Proof: Suppose the true $\theta^* \in \Theta_{I^*}^*$. Then $V \leq V_{I^*}$, 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^* .

$$\begin{aligned} \sup_{\theta \in \Theta_I^*} E_{\theta}(V) &= \sup_{\theta \in \Theta_I^*} E_{\theta} \left(\sum_{i \in I^*} I\{\text{reject } H_{0i}\} \right) \\ &\leq \sup_{\theta \in \Theta_I^*} E_{\theta} \left(\sum_{i \in I^*} I\{I^*\text{-reject } H_{0i}\} \right) \\ &= \sup_{\theta \in \Theta_I^*} E_{\theta}(V_{I^*}) \leq \alpha. \end{aligned}$$

Suppose in each partitioned subspace Θ_I^* , we test H_{0i} using test statistic T_i^I . We describe three types of rejection rules. In addition, we provide some sufficient conditions that lead to different step-wise procedures.

1.4.1 One-step procedure

We first propose the following rejection rule for testing $\{H_{0i}, i \in I\}$ in each Θ_I^* .

$$I\text{-reject } H_{0i} \text{ if } T_i^I > c_I.$$

Collating decisions for each H_{0i} following P3 leads to the following:

$$\text{Reject } H_{0i} \text{ if } T_i^I > c_I, \text{ for all } I \ni i.$$

Although $E(V)$ is controlled using the generalized partitioning principle, each hypothesis needs to be tested in $2^k - 1$ partitioned subspaces. The testing procedure becomes increasingly complicated even when we have moderate number of hypotheses k . However, if the test statistics and the critical values satisfy certain conditions, some of the partition subspace could be skipped, leading to a simplified shortcut procedure with only k or fewer steps.

Assume in each partition of the parameter space Θ_I^* , the test statistics and the critical values satisfy the following conditions:

- S1: The test statistic for each H_{0i} , T_i^I , should not depend on which Θ_I is tested, that is $T_i = T_i^I$.
- S2: The critical value c_I , should only depend on the cardinality of I and satisfy monotonicity.
 - If $|I| = |J|$, then $c_I = c_J$ denoted as $c_{|I|}$.
 - If $I \subseteq J$, then $c_{|I|} \leq c_{|J|}$.

Given condition S1, we reject an individual hypothesis H_{0i} if $T_i > c_I$, for all $I \ni i$. From S2, the critical values are monotone, thus we reject an individual hypothesis H_{0i} if $T_i > c_k$. This is exactly a one step procedure, comparing each test statistic T_i with the same critical value c_k .

Next, we discuss how to calculate critical values. In calculations related to critical values, we assume the test statistics $T_i, i = 1, \dots, k$ follow a multivariate t distribution with ν degrees of freedom.

Note there is only one critical value c_k that needs to be calculated in the partitioned subspace Θ_K^* denoting all hypotheses are true.

$$\begin{aligned}
\alpha = \sup_{\boldsymbol{\theta} \in \Theta_K^*} E_{\boldsymbol{\theta}}(V_K) &= \sup_{\boldsymbol{\theta} \in \Theta_K^*} E_{\boldsymbol{\theta}} \left(\sum_{i \in K} I\{K\text{-reject } H_{0i}\} \right) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_K^*} E_{\boldsymbol{\theta}} \left(\sum_{i \in K} I\{T_i > c_k\} \right) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_K^*} \sum_{i=1}^k P_{\boldsymbol{\theta}}(T_i > c_k) \\
&= k \sup_{\theta_1 \in \Theta_1} P_{\theta_1}(T_1 > c_k).
\end{aligned}$$

Note the last equality comes from the fact that each test statistic T_i is involved with one θ_i only. As T_1 follows a central t distribution with ν d.f. under the supremum, we get the critical value $c_k = t_{\alpha/k, \nu}$. Note that this critical value does not depend on correlations between the test statistics.

Gordon et al. (2007) mentioned that comparing p-values to α/k controls the expected value of false rejections. They called this procedure a Bonferroni procedure because the same critical value is obtained using the Bonferroni inequality for FWER control at level α . We do not have the same view because in developing the critical values, a Bonferroni inequality was never used. The critical values share the same form simply because that the expected values can be written as the sum of probabilities for the one-step procedure.

1.4.2 A step-down procedure

This section uses a different rejection rule for testing $\{H_{0i}, i \in I\}$ in each Θ_I^* which leads to a step-down procedure.

$$I\text{-reject all } H_{0i}, i \in I \text{ if } \max_{i \in I} T_i^I > c_I.$$

Collating decisions following P3, we have the following:

$$\text{Reject } H_{0i} \text{ if } \max_{j \in I} T_j^I > c_I, \text{ for all } I \ni i.$$

Suppose the test statistics and critical values satisfy conditions S1 and S2, let $[1], \dots, [k]$ be the random indices such that $T_{[1]} \leq \dots \leq T_{[k]}$. Then executing all $2^k - 1$ tests for each H_{0i} is equivalent to the following step-down procedure.

- Step 1: If $T_{[k]} > c_k$ then reject $H_{0[k]}$ and go to step 2; otherwise stop.
- Step 2: If $T_{[k-1]} > c_{k-1}$ then reject $H_{0[k-1]}$ and go to step 3; otherwise stop.
- ⋮
- Step k: If $T_{[1]} > c_1$ then reject $H_{0[1]}$ and stop; otherwise stop.

Note that the procedure operates exactly the same as the sample determined procedures controlling FWER. The only difference is that the critical values c_i 's are different.

The proof of equivalence of the partitioning procedure and the step-down procedure is the exactly same as proving the validity of the shortcut of sample determined steps.

In total, k critical values need to be calculated in the step-down procedure. Let $T_{1:|I|} \leq T_{2:|I|} \leq \dots \leq T_{|I|:|I|}$ denote the order statistics for $|I|$ test statistics, $|I| =$

$1, \dots, k$. The largest critical value c_k is first calculated in Θ_K^* as follows:

$$\begin{aligned}
\alpha &= \sup_{\boldsymbol{\theta} \in \Theta_K^*} E_{\boldsymbol{\theta}}(V_K) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_K^*} k P_{\boldsymbol{\theta}}(\max_{i \in K} T_i > c_k) \\
&= k \sup_{\boldsymbol{\theta} \in \Theta_K^*} P_{\boldsymbol{\theta}}(T_{k:k} > c_k).
\end{aligned} \tag{1.4.1}$$

Similarly, the second largest critical value c_{k-1} can be calculated in any subspace Θ_I with $k-1$ hypotheses being true (i.e. $|I| = k-1$). We illustrate the calculation using $I = \{1, \dots, k-1\}$.

$$\begin{aligned}
\alpha &= \sup_{\boldsymbol{\theta} \in \Theta_I^*} E_{\boldsymbol{\theta}}(V_I) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_I^*} (k-1) P_{\boldsymbol{\theta}}(\max_{i \in I} T_i > c_{k-1}) \\
&= (k-1) \sup_{\boldsymbol{\theta} \in \Theta_I^*} P_{\boldsymbol{\theta}}(T_{k-1:k-1} > c_{k-1}).
\end{aligned} \tag{1.4.2}$$

Carrying on the same process, we obtain the critical values by solving the following set of equations.

$$P(T_{k:k}^* > c_k) = \frac{\alpha}{k} \tag{1.4.3}$$

$$P(T_{k-1:k-1}^* > c_{k-1}) = \frac{\alpha}{k-1} \tag{1.4.4}$$

$$\vdots$$

$$P(T_1^* > c_1) = \alpha \tag{1.4.5}$$

where $T_i^*, i = 1, \dots, k$ are assumed to have central multivariate t distribution representing the case after taking supreme in (1.4.1) or (1.4.2). Given the exact correlation structure between the test statistics $T_1, \dots, T_k$, the solution for the critical values might involve multi-dimensional integrals. Table 1.2 gives critical values using either the Bonferroni or independence inequality. It can be shown that the use of

the Bonferroni inequality leads to conservative results (larger critical values) for any correlation structure of the test statistics, while using the independence inequality is conservative if the correlation between any pair of test statistics is nonnegative.

Table 1.2: Critical values for the step-down procedure controlling E(V)

	Bonferroni	Independence
c_k	$t_{\frac{\alpha}{k^2}, \nu}$	$t_{1-(1-\frac{\alpha}{k})^{1/k}, \nu}$
c_{k-1}	$t_{\frac{\alpha}{(k-1)^2}, \nu}$	$t_{1-(1-\frac{\alpha}{k-1})^{1/(k-1)}, \nu}$
$\vdots$	$\vdots$	$\vdots$
c_1	$t_{\alpha, \nu}$	$t_{\alpha, \nu}$

Comparing the critical values in Table 1.2 with the one-step critical value $t_{\frac{\alpha}{k}, \nu}$, we can see that, unlike the FWER control situation, the step-down procedure developed in this section is not uniformly more powerful than the one-step procedure for E(V) control.

1.4.3 A step-up procedure

This section shows that a step-up procedure can be developed using the following rejection rule in each Θ_I^* .

Test $\{H_{0i}, i \in I\}$, the null hypotheses that are true in Θ_I^* , using all the ordered statistics, $T_{[1]}^I \leq T_{[2]}^I \leq \dots T_{[|I|]}^I$ in a step-up fashion:

Step 1: if $T_{[1]}^I > c_1^I$ then I -reject all $H_{0i}, i \in I$ and stop;

otherwise go to step 2.

Step 2: if $T_{[2]}^I > c_2^I$ then I -reject $H_{0i}, i \in I$ corresponding to $T_{[2]}^I, \dots, T_{[|I|]}^I$ and stop; otherwise go to step 3.

$\vdots$

Step $|I|$: if $T_{[|I|]}^I > c_{|I|}^I$ then I -reject $H_{0i}, i \in I$ corresponding to $T_{[|I|]}^I$ and stop.

Assume in each partition of the parameter space Θ_I^* , in addition to S1 from Section 1.4.1, the critical values satisfy the following condition:

- S3: c_i^I does not depend on the set we test, that is the $c_i^I = c_i$ and satisfy the monotonicity $c_1 \leq c_2 \leq \dots \leq c_k$.

Suppose the test statistics and critical values satisfy conditions S1 and S3. Let $[1], \dots, [k]$ be the random indices such that $T_{[1]} \leq \dots \leq T_{[k]}$. Then the procedure after collating results from this step-up test in each Θ_I^* , is equivalent to the following step-up procedure.

- Step 1: If $T_{[1]} > c_1$ then reject all $H_{0i}, i = 1, \dots, k$ and stop; otherwise go to step 2.

- Step 2: If $T_{[2]} > c_2$ then reject $H_{0[i]}, i = 2, \dots, k$ and stop; otherwise go to step 3.

$\vdots$

- Step k : If $T_{[k]} > c_k$ then reject $H_{0[k]}$ and stop; otherwise stop.

Next, we show that S1 and S3 are sufficient conditions for the step-up shortcut.

In Step 1, if $T_{[1]} > c_1$, then for all $\Theta_I^*, I \ni [1]$, all $H_{0i}, i \in I$ are I -rejected because $T_{[1]}^I = T_{[1]} > c_1^I = c_1$. In testing the rest of the partition subspaces, namely $\Theta_I^*, [1] \notin I$, we have $T_{[1]}^I \geq T_{[1]} > c_1$. Thus all $H_{0i}, i \in I$ are I -rejected as well. Therefore, all hypotheses $H_{0i}, i \in K$ are rejected.

In Step 2, if $T_{[1]} \leq c_1$ and $T_{[2]} > c_2$, there are two cases. For those $\Theta_I^*, [1] \in I$, all hypotheses $H_{0i}, i \in I$ except $H_{0[1]}$ are I -rejected because $T_{[1]}^I = T_{[1]} \leq c_1$ and $T_{[2]}^I \geq T_{[2]} > c_2$. In testing the rest partition subspaces $\Theta_I^*, [1] \notin I$, all hypotheses $H_{0i}, i \in I$ are rejected because $T_{[1]}^I \geq T_{[2]} > c_2 > c_1$. Combining these two cases, all hypotheses $H_{0i}, i \in K$ are rejected except for $H_{0[1]}$.

A similar technique can be carried out in proving the other shortcut steps.

The critical values for the step-up procedure can be calculated sequentially from the following set of equations.

$$P(T_1^* > c_1) = \alpha \quad (1.4.6)$$

$$2P(T_{1:2}^* > c_1) + P(T_{1:2}^* \leq c_1 \text{ and } T_{2:2}^* > c_2) = \alpha \quad (1.4.7)$$

$$\vdots$$

$$\begin{aligned} kP(T_{1:k}^* > c_1) + (k-1)P(T_{1:k}^* \leq c_1, T_{2:k}^* > c_2) + \dots \\ + P(\cap_{i=1}^{k-1} \{T_{i:k}^* \leq c_i\}, T_{k:k}^* > c_k) = \alpha \end{aligned} \quad (1.4.8)$$

where $T_i^*, i = 1, \dots, k$ have a central multivariate t distribution. The solution to the set of equations above depends on the joint distribution of the ordered test statistics. Even with known correlations between the test statistics, it is extremely complicated to solve for the critical values. For illustration purposes, Table 1.3 presents the critical values for two special correlation structures between the test statistics in the

two dimensional case. Note the critical values under theses two correlations are not necessarily larger than other cases.

Table 1.3: Critical values for the step-up procedure controlling $E(V)$ in two dimensional case

	$Cor(T_1, T_2) = -1$	$Cor(T_1, T_2) = 0$
c_1	$t_{\alpha, \nu}$	$t_{\alpha, \nu}$
c_2	$t_{\frac{\alpha}{2}, \nu}$	$t_{\alpha - \frac{\alpha}{2(1-\alpha)}, \nu}$

Note: $t_{\alpha - \frac{\alpha}{2(1-\alpha)}, \nu} > t_{\frac{\alpha}{2}, \nu}$

Unlike for the step-down procedure, critical values calculated under negative 1 correlation are actually smaller than the ones calculated under independence. For the two-dimensional case, the critical values for the step-up procedure are all smaller than those for the step-down procedure under either correlation scenario ($t_{\alpha - \frac{\alpha}{2(1-\alpha)}, \nu} < t_{1-\sqrt{1-\alpha}/2, \nu}$ for $\alpha < 0.3234$). This means that the step-up procedure is uniformly more powerful than the step-down procedure under theses two correlations in the two dimensional case. Comparing with the one-step critical value $t_{\alpha/2, \nu}$, if correlation is -1, then the step-up procedure is uniformly more powerful. Otherwise, if correlation is 0, then the step-up procedure is not uniformly more powerful.

1.5 FDR control using the generalized partition principle

In this section, we show that one can use a modified version of generalized partitioning principle to control the false discovery rate, FDR .

The modified generalized Partitioning Principle states:

P1. Partition the parameter space $\bigcup_{i \in K} \Theta_i$ into disjoint sets: for each $I \subseteq K = \{1, \dots, k\}$, let $\Theta_I^* = \bigcap_{i \in I} \Theta_i \cap (\bigcap_{j \notin I} \Theta_j^c)$.

P'2. In each Θ_I^* , test $\{H_{0i} : \theta_i \in \Theta_i, i \in K\}$ at level α controlling

$$\sup_{\theta \in \Theta_I^*} E_{\theta} \left(\frac{V_I}{V_I + S_I} I(V_I > 0) \right) \quad (1.5.1)$$

where V_I is the number of false rejections and S_I is the number of correct rejections in Θ_I^* . Define H_{0i} to be *I-rejected* if it is rejected in Θ_I^* .

P'3. Reject H_{0i} if H_{0i} is *I-rejected* for all $I, i \in I$.

Note, in each Θ_I^* , we test not only hypotheses $H_{0i}, i \in I$ that are true, but also hypotheses that are false $H_{0i}, i \in K \setminus I$ to ensure that controlling FDR in each Θ_I^* , guarantees the overall control of FDR (see proof of Theorem 1.5.1).

Under assumption A1 given below, we prove in Theorem 1.5.1 that the overall FDR is controlled at the pre-specified level α following the modified generalized partition principle P1, P'2, P'3.

Assumption

A1. For any set $J \subset K = \{1, \dots, k\}$, if H_{0i} is *J-rejected* for any $i \notin J$, then we assume it is *I-rejected* for every $I \in i$ meaning it is finally rejected.

Theorem 1.5.1. *Controlling (1.5.1) for testing each partition hypothesis $\Theta_I^*, I \subseteq K$ following the modified generalized partition principle P1, P'2, P'3 guarantees the overall control of FDR under assumption A1.*

Proof: Suppose the true $\theta^* \in \Theta_{I^*}^*$, then $H_{0i}, i \in I^*$ 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 A1 ($i \notin I^*$), 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

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_{I^*}^*} FDR_{\boldsymbol{\theta}} &= \sup_{\boldsymbol{\theta} \in \Theta_{I^*}^*} E_{\boldsymbol{\theta}} \left(\frac{V^*}{V^* + S^*} I(V^* > 0) \right) \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta_{I^*}^*} E_{\boldsymbol{\theta}} \left(\frac{V_{I^*}}{V_{I^*} + S_{I^*}} I(V_{I^*} > 0) \right) \leq \alpha. \end{aligned}$$

We propose the following tests for $\{H_{0i}, i \in K\}$ using test statistic T_i^I in each Θ_I^* .

- Reject all $H_{0i}, i \in I$, if $\max \{T_i^I, i \in I\} > c_I$.
- Reject H_{0j} if $T_j^I > d_I, j \in K \setminus I$.

The following two conditions on test statistics and critical values are then sufficient for assumption A1.

- S1: The test statistic for each H_{0i}, T_i^I should not depend on which Θ_I^* is tested, that is $T_i = T_i^I$.
- S4: The critical values corresponding to the true null hypotheses are smaller than the ones corresponding to the false null hypotheses. That is $c_I \leq d_J, \forall I, J \subseteq K$.

1.5.1 Three hypotheses example

We illustrate FDR control using the modified generalized partitioning principle in the case of testing three hypotheses. For $k = 3$, there are $2^3 - 1 = 7$ partitions of the

parameter space that must be tested. The critical values of the test statistics T_i , for each hypothesis H_{0i} in each partitioned subspace Θ_I^* are presented in Table 1.4.

Table 1.4: Critical values in each partitioned subspace

Partitioned space	T_1	T_2	T_3
$\Theta_{\{123\}}^* : H_{01}, H_{02}, H_{03} \text{ are all true}$	$c_{\{1,2,3\}}$	$c_{\{1,2,3\}}$	$c_{\{1,2,3\}}$
$\Theta_{\{12\}}^* : H_{01}, H_{02} \text{ are true, } H_{03} \text{ is false}$	$c_{\{1,2\}}$	$c_{\{1,2\}}$	$d_{\{1,2\}}$
$\Theta_{\{13\}}^* : H_{01}, H_{03} \text{ are true, } H_{02} \text{ is false}$	$c_{\{1,3\}}$	$d_{\{1,3\}}$	$c_{\{1,3\}}$
$\Theta_{\{23\}}^* : H_{02}, H_{03} \text{ are true, } H_{01} \text{ is false}$	$d_{\{2,3\}}$	$c_{\{2,3\}}$	$c_{\{2,3\}}$
$\Theta_{\{1\}}^* : H_{01} \text{ is true, } H_{02}, H_{03} \text{ are false}$	$c_{\{1\}}$	$d_{\{1\}}$	$d_{\{1\}}$
$\Theta_{\{2\}}^* : H_{02} \text{ is true, } H_{01}, H_{03} \text{ are false}$	$d_{\{2\}}$	$c_{\{2\}}$	$d_{\{2\}}$
$\Theta_{\{3\}}^* : H_{03} \text{ is true, } H_{01}, H_{02} \text{ are false}$	$d_{\{3\}}$	$d_{\{3\}}$	$c_{\{3\}}$

Note: Assumption S4 implies $d_J \geq c_I, \forall I, J \subseteq \{1, 2, 3\}$.

Based on Table 1.4, some of the rejection rules are illustrated below:

- In $\Theta_{\{1,2,3\}}^*$ where all three hypotheses are true,
 - if $\max\{T_1, T_2, T_3\} > c_{\{1,2,3\}}$, then $\{1, 2, 3\}$ -reject all H_{01}, H_{02}, H_{03} ;
 - otherwise, do not $\{1, 2, 3\}$ -reject any.
- In $\Theta_{\{1,3\}}^*$ where H_{01}, H_{03} are true, H_{02} is false:
 - if $\max\{T_1, T_3\} > c_{\{1,3\}}$, then $\{1, 3\}$ -reject both H_{01}, H_{03} ;
 - otherwise, do not $\{1, 3\}$ -reject H_{01}, H_{03} .
 - if $T_2 > d_{\{1,3\}}$, $\{1, 3\}$ -reject H_{02} ; otherwise do not $\{1, 3\}$ -reject H_{02} .
- In $\Theta_{\{1\}}^*$ where H_{01} is true, H_{02}, H_{03} are false:
 - if $T_1 > c_{\{1\}}$, then $\{1\}$ -reject H_{01} ; otherwise do not $\{1\}$ -reject H_{01} .

- if $T_2 > d_{\{1\}}$, $\{1\}$ -reject H_{02} ; otherwise do not $\{1\}$ -reject H_{02} .
- if $T_3 > d_{\{1\}}$, $\{1\}$ -reject H_{03} ; otherwise do not $\{1\}$ -reject H_{03} .

Once we have all 7 I -reject or not results for each hypothesis H_{0i} , $i = 1, 2, 3$, we collate results for each hypothesis. For example, H_{01} is rejected if it is I -rejected for every $I \ni 1$, namely, $\{1, \}$ -rejected, $\{1, 2\}$ -rejected, $\{1, 3\}$ -rejected, and $\{1, 2, 3\}$ -rejected. (These I -rejections are made through the c critical values.)

Suppose the true $\theta^* \in \Theta_{12}^*$. That means H_{01}, H_{02} are true, and H_{03} 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. $V^* \leq V_{\{1,2\}}$ since the rejection of either H_{01} or H_{02} implies they need to be $\{1, 2\}$ -rejected.
- b. $S_{\{1,2\}} \leq S^*$. If H_{03} is $\{1, 2\}$ -rejected ($S_{\{1,2\}} = 1$), by condition S4, $T_3 > d_{\{1,2\}}$ implies $T_3 > c_I$ for every $I \ni 3$. This leads to the rejection of H_{03} , i.e. $S^* = 1$.

We have $\frac{V^*}{S^*} \leq \frac{V_{\{1,2\}}}{S_{\{1,2\}}}$, and thus FDR is controlled.

1.5.2 Step-down shortcut

A step-down procedure can be developed provided that additional conditions are met.

- S5: The critical value c_I, d_I for Θ_I^* should only depend on the cardinality of I and satisfy monotonicity.
 - If $|I| = |J|$, then $c_I = c_J, d_I = d_J$ denoted as $c_{|I|}, d_{|I|}$.
 - Equivalent to S4, $c_{|I|} \leq d_{|J|}, \forall I, J \subseteq K$.

- If $I \subseteq J$, then $c_{|I|} \leq c_{|J|}$.

As the final rejection of a particular hypothesis does not depend on the I -rejection in subspaces where it is false, the step-down shortcut only involves critical values of the c 's, leaving the d 's flexible only to control the FDR_I in each subspace.

Under conditions S1 and S5, let $[1], \dots, [k]$ be the random indices such that $T_{[1]} \leq \dots \leq T_{[k]}$, then the step-down procedure is as follows:

- Step 1: If $T_{[k]} > c_k$ then reject $H_{0[k]}$ and go to step 2; otherwise stop.
- Step 2: If $T_{[k-1]} > c_{k-1}$ then reject $H_{0[k-1]}$ and go to step 3; otherwise stop.
- $\vdots$
- Step k : If $T_{[1]} > c_1$ then reject $H_{0[1]}$ and stop; otherwise stop.

Note that the procedure operates exactly the same as the sample determined procedures controlling FWER or the step-down procedure controlling $E(V)$. The only difference is that the critical values, c_i 's, are different and they are determined by the choices of d_i 's.

FWER control as a special case

To gain an insight on the flexibility of choices of the d_i 's and no obvious involvement of them in the final decision process, we illustrate the FWER control as a special case among all the choices of the d_i 's.

Let d_i 's be some large positive numbers such that all the test statistics, T_i 's, have almost zero probability of exceeding those. Then in each partitioned subspace Θ_I^* , $T_j \leq d_{|I|}, j \notin I$, i.e. no false null hypotheses could be I -rejected ($S_I = 0$). In this

case, the control of $FDR_I = E(\frac{V_I}{V_I+S_I}I(V_I > 0)) = E(\frac{V_I}{V_I+0}I(V_I > 0)) = E(I(V_I > 0)) = P(V_I > 0)$ is equivalent to the control of $FWER_I$.

The proof of equivalence of the partitioning procedure and the step-down procedure is the same as proving the validity of the shortcut of sample-determined steps in the $FWER$ control case.

Three hypotheses example continued

We again use the three dose groups example to illustrate the step-down procedure. Let T_i be the test statistic of null hypothesis H_{0i} and $c_i, d_j, i = 1, 2, 3, j = 1, 2$ be the critical values. Note from the monotonicity assumption S5, $c_1 \leq c_2 \leq c_3 \leq d_j, j = 1, 2$.

The step-down shortcut proceeds as follows:

- Step 1: If $T_{[3]} > c_3$ then reject $H_{0[3]}$ and go to step 2; otherwise stop.
- Step 2: If $T_{[2]} > c_2$ then reject $H_{0[2]}$ and go to step 3; otherwise stop.
- Step 3: If $T_{[1]} > c_1$ then reject $H_{0[1]}$ and stop; otherwise stop.

Calculation of critical values

In the FDR control case, one has to calculate both sets of critical values $c_{|I|}$'s and $d_{|I|}$'s. However, given only one equation by setting FDR_I at α , one has to choose some appropriate values for the $d_{|I|}$'s and then solve for the $c_{|I|}$'s. One issue that differentiates the FDR control from all other error rate control cases is the fact that the calculation involves parameters from null hypotheses that are false.

We illustrate the steps of calculating critical values, and point out the difficulties in the process in the simple setting of testing $k = 3$ hypotheses. For illustration purposes, we assume the null hypotheses are of the form $H_{0i} : \theta_i \leq 0, i = 1, 2, 3$ and

the test statistics T_i are i.i.d from $N(\theta_i, 1)$. Table 1.5 presents the way of setting the critical values with $d = d_1 = d_2$.

Table 1.5: Setting critical values for the test statistics in each partitioned subspace for testing three hypotheses

Partitioned space	T_1	T_2	T_3
$\Theta_{\{123\}}^* : H_{01}, H_{02}, H_{03}$ are all true	c_3	c_3	c_3
$\Theta_{\{12\}}^* : H_{01}, H_{02}$ are true, H_{03} is false	c_2	c_2	d
$\Theta_{\{13\}}^* : H_{01}, H_{03}$ are true, H_{02} is false	c_2	d	c_2
$\Theta_{\{23\}}^* : H_{02}, H_{03}$ are true, H_{01} is false	d	c_2	c_2
$\Theta_{\{1\}}^* : H_{01}$ is true, H_{02}, H_{03} are false	c_1	d	d
$\Theta_{\{2\}}^* : H_{02}$ is true, H_{01}, H_{03} are false	d	c_1	d
$\Theta_{\{3\}}^* : H_{03}$ is true, H_{01}, H_{02} are false	d	d	c_1

Note: The critical values satisfy $d \geq c_3 \geq c_2 \geq c_1$.

In $\Theta_{\{123\}}^* = \{\theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \text{ and } \theta_3 \leq 0\}$ where all null hypotheses are true, the FDR control is equivalent to FWER control, thus we have c_3 as the $1 - (1 - \alpha)^{1/3}$

quantile from standard normal distribution shown as follows.

$$\begin{aligned}
\alpha &= FDR_{\{123\}} \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{123\}}^*} E_{\boldsymbol{\theta}} \left[\frac{V_{\{123\}}}{V_{\{123\}} + S_{\{123\}}} I\{V_{\{123\}} > 0\} \right] \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{123\}}^*} P_{\boldsymbol{\theta}}(V_{\{123\}} > 0) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{123\}}^*} P_{\boldsymbol{\theta}}(T_1 > c_3 \text{ or } T_2 > c_3 \text{ or } T_3 > c_3) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{123\}}^*} [1 - P_{\boldsymbol{\theta}}(T_1 \leq c_3 \text{ and } T_2 \leq c_3 \text{ and } T_3 \leq c_3)] \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{123\}}^*} [1 - P_{\theta_1}(T_1 \leq c_3)P_{\theta_2}(T_2 \leq c_3)P_{\theta_3}(T_3 \leq c_3)] \\
&= 1 - P_{\theta_1=0}(T_1 \leq c_3)P_{\theta_2=0}(T_2 \leq c_3)P_{\theta_3=0}(T_3 \leq c_3) \\
&= 1 - \Phi(c_3)^3
\end{aligned}$$

where Φ is the cdf of standard normal distribution

$$\begin{aligned}
&\Rightarrow \Phi(c_3) = (1 - \alpha)^{(1/3)} \\
&\Rightarrow c_3 = \Phi^{-1}((1 - \alpha)^{(1/3)})
\end{aligned}$$

In $\Theta_{\{12\}}^* = \{\theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \text{ and } \theta_3 > 0\}$, the number of false rejections $V_{\{1,2\}}$ is either 0 or 2 depending on whether $\max(T_1, T_2) > c_2$. The critical value c_2 is

calculated in this subspace shown as follows.

$$\begin{aligned}
\alpha &= FDR_{\{12\}} \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{12\}}^*} E_{\boldsymbol{\theta}} \left[\frac{V_{\{12\}}}{V_{\{12\}} + S_{\{12\}}} I\{V_{\{12\}} > 0\} \right] \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{12\}}^*} \left\{ E_{\boldsymbol{\theta}} \left[\frac{V_{\{12\}}}{V_{\{12\}} + 1} I\{V_{\{12\}} > 0\} \middle| S_{\{12\}} = 1 \right] P_{\boldsymbol{\theta}}(S_{\{12\}} = 1) \right. \\
&\quad \left. + E_{\boldsymbol{\theta}} \left[\frac{V_{\{12\}}}{V_{\{12\}} + 0} I\{V_{\{12\}} > 0\} \middle| S_{\{12\}} = 0 \right] P_{\boldsymbol{\theta}}(S_{\{12\}} = 0) \right\} \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{12\}}^*} \left\{ \frac{2}{3} P_{\boldsymbol{\theta}}(V_{\{12\}} = 2) P_{\boldsymbol{\theta}}(T_3 > d) + P_{\boldsymbol{\theta}}(V_{\{12\}} > 0) P_{\boldsymbol{\theta}}(T_3 \leq d) \right\} \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{12\}}^*} P_{\boldsymbol{\theta}}(T_1 > c_2 \text{ or } T_2 > c_2) \left\{ \frac{2}{3} P_{\boldsymbol{\theta}}(T_3 > d) + P_{\boldsymbol{\theta}}(T_3 \leq d) \right\} \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{12\}}^*} [1 - P_{\theta_1}(T_1 \leq c_2) P_{\theta_2}(T_2 \leq c_2)] \left\{ \frac{2}{3} + \frac{1}{3} P_{\theta_3}(T_3 \leq d) \right\} \\
&= \frac{1}{3} [1 - P_{\theta_1=0}(T_1 \leq c_2) P_{\theta_2=0}(T_2 \leq c_2)] [2 + P_{\theta_3=0}(T_3 \leq d)] \\
&= \frac{1}{3} [1 - \Phi(c_2)^2] [2 + \Phi(d)] \\
&\Rightarrow c_2 = \Phi^{-1} \left(\sqrt{1 - \frac{3\alpha}{2 + \Phi(d)}} \right) \tag{1.5.2}
\end{aligned}$$

Thus, given appropriate choice of d ($> c_3$), c_2 can be calculated through (1.5.2). For large values of d where $\Phi(d)$ is almost 1, the critical value c_2 becomes the $1 - \sqrt{1 - \alpha}$ quantile of the standard normal distribution, coinciding with the critical value in the step-down FWER control procedure.

In $\Theta_{\{1\}}^* = \{\theta_1 \leq 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 > 0\}$, the critical value c_1 is derived as follows.

$$\begin{aligned}
\alpha &= FDR_{\{1\}} \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{1\}}^*} E_{\boldsymbol{\theta}} \left[\frac{V_{\{1\}}}{V_{\{1\}} + S_{\{1\}}} I\{V_{\{1\}} > 0\} \right] \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{1\}}^*} E_{\boldsymbol{\theta}} \left[\frac{V_{\{1\}}}{V_{\{1\}} + 2} I\{V_{\{1\}} > 0\} \middle| S_{\{1\}} = 2 \right] P_{\boldsymbol{\theta}}(S_{\{1\}} = 2) \\
&+ E_{\boldsymbol{\theta}} \left[\frac{V_{\{1\}}}{V_{\{1\}} + 1} I\{V_{\{1\}} > 0\} \middle| S_{\{1\}} = 1 \right] P_{\boldsymbol{\theta}}(S_{\{1\}} = 1) \\
&+ E_{\boldsymbol{\theta}} \left[\frac{V_{\{1\}}}{V_{\{1\}} + 0} I\{V_{\{1\}} > 0\} \middle| S_{\{1\}} = 0 \right] P_{\boldsymbol{\theta}}(S_{\{1\}} = 0) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{1\}}^*} \left[\frac{1}{3} P_{\boldsymbol{\theta}}(S_{\{1\}} = 2) + \frac{1}{2} P_{\boldsymbol{\theta}}(S_{\{1\}} = 1) + P_{\boldsymbol{\theta}}(S_{\{1\}} = 0) \right] P_{\boldsymbol{\theta}}(V_{\{1\}} = 1) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{1\}}^*} \left[\frac{1}{3} P_{\theta_2}(T_2 > d) P_{\theta_3}(T_3 > d) + \frac{1}{2} P_{\theta_2}(T_2 > d) P_{\theta_3}(T_3 \leq d) \right. \\
&+ \left. \frac{1}{2} P_{\theta_2}(T_2 \leq d) P_{\theta_3}(T_3 > d) + P_{\theta_2}(T_2 \leq d) P_{\theta_3}(T_3 \leq d) \right] P_{\theta_1}(T_1 > c_1) \\
&= \sup_{\boldsymbol{\theta} \in \Theta_{\{1\}}^*} \left[\frac{1}{3} \bar{\Phi}(d - \theta_2) \bar{\Phi}(d - \theta_3) + \frac{1}{2} \bar{\Phi}(d - \theta_2) \Phi(d - \theta_3) \right. \\
&+ \left. \frac{1}{2} \Phi(d - \theta_2) \bar{\Phi}(d - \theta_3) + \Phi(d - \theta_2) \Phi(d - \theta_3) \right] \bar{\Phi}(c_1 - \theta_1) \tag{1.5.3}
\end{aligned}$$

Let $f(\theta_2, \theta_3) =$

$$\frac{1}{3} \bar{\Phi}(d - \theta_2) \bar{\Phi}(d - \theta_3) + \frac{1}{2} [\bar{\Phi}(d - \theta_2) \Phi(d - \theta_3) + \Phi(d - \theta_2) \bar{\Phi}(d - \theta_3)] + \Phi(d - \theta_2) \Phi(d - \theta_3)$$

Then $\frac{\partial f}{\partial \theta_2} = -\frac{1}{6} \phi(d - \theta_2) (1 + 2\Phi(d - \theta_3)) < 0$ and $\frac{\partial f}{\partial \theta_3} = -\frac{1}{6} \phi(d - \theta_3) (1 + 2\Phi(d - \theta_2)) < 0$,

where ϕ is the pdf of the standard normal distribution. Therefore the supremum over

$\Theta_{\{1\}}^* = \{\theta_1 \leq 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 > 0\}$ in (1.5.3) should take place at $\theta_1 = \theta_2 = \theta_3 =$

0. This leads to the solution of c_1 as a function of d in (1.5.4),

$$c_1 = \Phi^{-1} \left(1 - \frac{3\alpha}{1 + \Phi(d) + \Phi(d)^2} \right). \tag{1.5.4}$$

Figure 1.1: Critical values c_2 and c_1 plotted as functions of d for the step-down procedure controlling FDR at $\alpha = 0.05$ level.

Figure 1.1 plots both critical values c_2 and c_1 as functions of d ranging from c_3 to 5. One can easily check that the monotonicity assumption between the critical values, $c_3 > c_2 > c_1$, is satisfied. FWER is the special case when d gets large in the sense that c_2 and c_1 are the same as the critical values for a step-down procedure controlling FWER under independence.

1.5.3 Comparison with other FDR control methods

The two commonly used FDR control procedures are the step-up procedure in Benjamini and Hochberg (1995) and the step-down procedure in Benjamini and Liu (1999). Both procedures control FDR if the test statistics are independent. In the

proof of overall FDR control, both papers avoided the calculation involving test statistics from the false null hypotheses by conditioning on those test statistics and controlling the conditional expectation of V/R . However, this approach only works under independence.

The generalized partition principle in this section provides a general framework for developing FDR controlling procedures and is flexible in terms of how each test can be chosen.

The step-down procedure developed through the generalized partition principle in Section 1.5.2 can be applied to test statistics with any correlation structure. The downside, however, is that the critical values ARE only slightly smaller than the ones corresponding to the FWER control case, as shown in Figure 1.1. In fact, from Table 1.6, we see that they are larger than the ones in Benjamini and Liu (1999). Compared to Benjamini and Hochberg (1995), the step-down procedure produces larger c_2 but smaller c_1 and c_3 .

Table 1.6: Critical values for the step-down procedure and two other methods controlling FDR at 0.05 for testing three hypotheses under independence

	SD	BH	BL
c_3	2.1212	2.1280	2.1212
c_2	1.9520	1.8339	1.7716
c_1	1.6366	1.6449	1.0364

Note: SD denotes the step-down procedure in Section 1.5.2, BH denotes the step-up procedure in Benjamini and Hochberg (1995), and BL denotes the step-down procedure in Benjamini and Liu (1999). The critical value for SD is calculated when $d = 2.1212$.

CHAPTER 2

PARTITION TESTING IN DOSE-RESPONSE STUDIES WITH MULTIPLE ENDPOINTS

Dose-response studies with multiple endpoints can be formulated as closed testing or partition testing problems. When the endpoints are primary and secondary, whether the order in which the doses are to be tested is pre-determined or sample-determined leads to different partitioning of the parameter space corresponding to the null hypotheses to be tested. We use the case of two doses and two endpoints to illustrate how to apply the partitioning principle to construct multiple tests that control the appropriate error rate when we step through both doses and endpoints. Graphical representation can be useful in visualizing the decision process.

2.1 Historical aspect of dose-response trials in pharmaceuticals

Pharmaceutical drug development has long been divided into four Phases with Phase II being the part of the development cycle where dose-response studies are conducted. The goal is usually to identify the best dose to use in confirmatory Phase III trials. While this has been the framework for several decades, there are still many drugs that fail to confirm efficacy and safety in Phase III (estimates range from 30-50%) with a substantial number of failures attributed to improper dose selection.

Thus, the importance of the design and analysis of dose response studies is as relevant today as it has ever been.

In the past, efficacy of a new drug was typically demonstrated by showing its superiority to a placebo (called a *negative* control). However, in recent years, when treatments known to be effective exist, and the disease does not cause mortality or irreversible morbidity, efficacy of a new drug might be defined as superiority or non-inferiority to a known effective treatment (called an *active* control). In the case of non-inferiority trials, the determination of non-inferiority might be based on what is a clinically meaningful difference. (See ICH E10.) Even in the case of superiority trials against a negative control, the definition of superiority might consider risk versus benefit, if there are toxicity concerns for the drug. (See Temple 2006.) Some early discussion of these concepts can be seen in Ruberg (1989, 1995a, 1995b), and Ruberg and Cairns (1998).

Dose-response studies may have multiple endpoints. A *primary endpoint* is one such that efficacy of a new drug relative to the control in this single endpoint constitutes evidence of efficacy. *Secondary endpoints* are ones such that efficacy of a new treatment in any secondary endpoints supports evidence of efficacy, but by themselves (i.e. in the absence of efficacy in a primary endpoint) do not constitute evidence of efficacy.

When there are primary and secondary endpoints, inference on the secondary endpoint is given only if the compound is efficacious for the primary endpoint at that dose. This ordering guides the selection of test statistics for each intersection hypothesis in closed testing (as in Dmitrienko et al. 2006). We give a different perspective in this dissertation, which is that the ordering guides the partitioning of

the parameter space in using the partitioning principle to construct multiple tests that control the appropriate FWER.

In some (but not all) dose-response studies, it may be appropriate to pre-determine the order in which inferences on the doses are given. For example, one might start with the high dose and proceed to inference on the low dose only if the high dose shows efficacy. We show how this “ordering” further guides the partitioning of the parameter space. The resulting partitioning test is in the form of a decision tree which can be represented graphically.

Interestingly, the joint distribution of t test statistics for multiple endpoints is *not* what is usually called the multivariate t distribution. This dissertation discusses the computation of this distribution in the bivariate case which, to avoid confusion, we call the *dual* t distribution. We show, for example, that using algorithms for multivariate t distributions results in slightly liberal critical values, while computing as if the t statistics were independent results in somewhat conservative critical values.

2.2 A motivating example for dose-response studies

Consider, for example, Arvanitis et al. (1997), a 26-center double-blind trial comparing the effect of five doses of the anti-psychotic drug “Seroquel” (Quetiapine) and the placebo with parallel design on a total of 361 patients. For the purpose of illustration, we will focus on two of the doses (75 mg/day and 600 mg/day) with the primary endpoint being the Clinical Global Impression (CGI) Global Improvement score and the secondary endpoint being CGI Severity of Illness score. Summary statistics for these two doses and endpoints are presented in Table 2.1.

Table 2.1: Summary statistics of two doses and two endpoints from Arvanitis et al. (1997)

Endpoint	Sample size	Dose		
		0 mg/day	75 mg/day	600 mg/day
CGI Global Improvement score (Primary)	Mean	4.78	4.22	3.58
	SE ¹	0.23	0.22	0.23
CGI Severity of Illness score (Secondary)	Mean	5.2	4.8	4.4
	SE	1.2	1.3	1.5

¹Standard error

Throughout this chapter, the placebo (0 mg/day), low dose (75 mg/day) and high dose (600 mg/day) groups will be indexed as $i = 0, 1, 2$. Primary and secondary endpoints will be indexed by superscripts $L = P, S$. For discussion involving only the primary endpoint, the superscript P will be dropped for convenience.

Let μ_i^L denote the mean response of dose group i for endpoint L , $i = 0, 1, 2$, $L = P, S$. Define $\theta_i^L = \mu_i^L - \mu_0^L$ as the true mean difference between dose group i and the placebo for endpoint L , $i = 0, 1, 2$, $L = P, S$. Let δ^L denote the clinically meaningful difference for endpoint L , $L = P, S$. The family of null hypotheses of interest consists of four null hypotheses with the first two in (2.2.1) concerning the primary endpoint and the second two in (2.2.2) concerning the secondary endpoint.

$$H_{01}^P : \theta_1^P \leq \delta^P \quad \text{vs.} \quad H_{a1}^P : \theta_1^P > \delta^P \quad (2.2.1)$$

$$H_{02}^P : \theta_2^P \leq \delta^P \quad \text{vs.} \quad H_{a2}^P : \theta_2^P > \delta^P$$

$$H_{01}^S : \theta_1^S \leq \delta^S \quad \text{vs.} \quad H_{a1}^S : \theta_1^S > \delta^S \quad (2.2.2)$$

$$H_{02}^S : \theta_2^S \leq \delta^S \quad \text{vs.} \quad H_{a2}^S : \theta_2^S > \delta^S$$

Assume the samples from dose group i for endpoint L , $Y_{i1}^L, \dots, Y_{in_i}^L$, $i = 0, 1, 2$, $L = P, S$, come from a model:

$$Y_{ir}^L = \mu_i^L + \epsilon_{ir}^L, \quad i = 0, 1, 2, r = 1, \dots, n_i, L = P, S, \quad (2.2.3)$$

where $\begin{pmatrix} \epsilon_{ir}^P \\ \epsilon_{ir}^S \end{pmatrix} \quad i.i.d. \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\sigma^P)^2 & \rho\sigma^P\sigma^S \\ \rho\sigma^P\sigma^S & (\sigma^S)^2 \end{pmatrix} \right) \quad i = 0, 1, 2, r = 1, \dots, n_i.$

Define T_i^L , $i = 1, 2$, $L = P, S$, in (2.2.4) to be the t-statistic for testing the null hypothesis H_{0i}^L , $i = 1, 2$, $L = P, S$,

$$T_i^L = \frac{\bar{Y}_{i\cdot}^L - \bar{Y}_{0\cdot}^L - \delta^L}{\hat{\sigma}^L \sqrt{1/n_i + 1/n_0}} \quad i = 1, 2, L = P, S, \quad (2.2.4)$$

where $\bar{Y}_{i\cdot}^L = \frac{1}{n_i} \sum_{r=1}^{n_i} Y_{ir}^L$, $\hat{\sigma}^L = \sqrt{\frac{\sum_{i=0}^2 \sum_{r=1}^{n_i} (Y_{ir}^L - \bar{Y}_{i\cdot}^L)^2}{\sum_{i=0}^2 (n_i - 1)}}$, $i = 0, 1, 2$, $L = P, S$.

The purpose of dose-response studies is to find which doses are effective. Control of the multiple testing error rate should control the probability of incorrectly inferring a dose is efficacious for some endpoint when in fact it is not. In this situation, controlling the False Discovery Rate (FDR) will not control this probability (see Huang and Hsu 2007), while controlling the Familywise Error Rate (FWER) strongly for appropriately formulated null hypotheses will. We thus consider methods that strongly control FWER in this chapter.

With four null hypotheses in (2.2.1) and (2.2.2), a straight-forward application of the closed testing principle would test $2^4 - 1 = 15$ intersection null hypotheses to control FWER. However, we will show that if inferences are ordered by both endpoints and doses, then testing only four (disjoint) null hypotheses according to the partitioning principle controls FWER. These hypotheses are tested in three steps,

with a multiplicity adjustment involved in only one of the steps. When there are $k > 2$ doses and $m > 2$ endpoints, a similar step-down procedure could, in theory, be developed following the same principle.

2.3 Multiple tests construction using the partitioning principle

The purpose of this chapter is to demonstrate how to use the partitioning principle, a fundamental multiple tests constructing technique, to construct multiple tests when there are multiple doses and multiple endpoints. We illustrate the idea with the two hypotheses in (2.2.1), concerning the primary endpoint first.

2.3.1 Closed testing to step-down doses according to sample responses

The closed testing technique of Marcus, Peritz and Gabriel (1976) tests all possible non-empty intersections of the hypotheses in (2.2.1), leading to the three hypotheses in (2.3.1), each at level- α :

$$\begin{aligned} H_0^\cap : \quad & \theta_1 \leq \delta \text{ and } \theta_2 \leq \delta && \text{(Neither dose is efficacious)} \\ H_{01} : \quad & \theta_1 \leq \delta && \text{(Low dose is not efficacious)} \\ H_{02} : \quad & \theta_2 \leq \delta && \text{(High dose is not efficacious)} \end{aligned} \tag{2.3.1}$$

The logical implications of testing are:

- If $H_0^\cap$ is not rejected, no inference is given even if H_{01} or H_{02} or both are rejected, because $H_0^\cap$ implies H_{01} and H_{02} .
- If only $H_0^\cap$ and H_{01} (H_{02}) are rejected but not H_{02} (H_{01}), then infer the low (high) dose is efficacious.

- If all three hypotheses are rejected, then infer both doses are efficacious.

We use t-statistics T_1 and T_2 defined in (2.2.4) (with superscript P dropped) to test the three intersection hypotheses in (2.3.1) with critical values $d_{\alpha,2,\nu}$ (the upper α quantile of Dunnett's distribution with 2 and $\nu = \sum_{i=0}^2 (n_i - 1)$ degrees of freedom) and $t_{\alpha,\nu}$ (the upper α quantile of the t distribution with ν degrees of freedom). Let (1) and (2) denote the random indices such that $T_{(1)} < T_{(2)}$. Then the rejection rules can be shown as in Table 2.2.

Table 2.2: Decision rules for closed testing with two doses and single primary endpoint

Hypothesis	Rejection rule	Test level
$H_0^\cap : \theta_1 \leq \delta \text{ and } \theta_2 \leq \delta$	$T_{(2)} > d_{\alpha,2,\nu}$	α
$H_{01} : \theta_1 \leq \delta$	$T_1 > t_{\alpha,\nu}$	α
$H_{02} : \theta_2 \leq \delta$	$T_2 > t_{\alpha,\nu}$	α

Since the critical values satisfy $d_{\alpha,2,\nu} > t_{\alpha,\nu}$, a step-down procedure with sample-determined steps exists, as follows:

- Step 1: If $T_{(2)} > d_{\alpha,2,\nu}$, infer dose (2) is efficacious and go to step 2; else stop.
- Step 2: If $T_{(1)} > t_{\alpha,\nu}$, infer dose (1) is efficacious and stop; else stop.

2.3.2 Partition to step-down doses

If the higher dosage is expected to be more efficacious, then one may choose to always test the high dose first. If the high dose is effective, then proceed to test the low dose; otherwise, stop. We can formally state this pre-determined sequence of testing as:

Condition A (*order in doses*): The low dose can not be claimed efficacious unless the high dose has shown evidence of efficacy.

The partitioning principle of Stefansson, Kim, Hsu (1988) and Finner and Strassburger (2002) is a general principle for constructing multiple tests. Under *condition A*, Hsu and Berger (1999) partition the null space, $\{\boldsymbol{\theta} \in \mathbb{R}^2 | \theta_1 \leq \delta \text{ or } \theta_2 \leq \delta\}$ into two disjoint subspaces $\{\boldsymbol{\theta} \in \mathbb{R}^2 | \theta_2 \leq \delta\}$ and $\{\boldsymbol{\theta} \in \mathbb{R}^2 | \theta_1 \leq \delta \text{ and } \theta_2 > \delta\}$ corresponding to the two hypotheses:

$$H_{02}^\downarrow : \quad \theta_2 \leq \delta \quad (\text{High dose is not efficacious})$$

$$H_{01}^\downarrow : \quad \theta_1 \leq \delta \text{ and } \theta_2 > \delta \quad (\text{Low dose is not efficacious but high dose is})$$

The logical implications of testing are:

- If $H_{02}^\downarrow$ is not rejected (regardless of $H_{01}^\downarrow$), then no inference is given, since “neither dose is efficacious” (which is contained in $H_{02}^\downarrow$) is not rejected.
- If $H_{02}^\downarrow$ and $H_{01}^\downarrow$ are rejected, then since the union of $H_{02}^\downarrow$ and $H_{01}^\downarrow$ is “either the low dose or the high dose is not efficacious”, the implication is “both the high dose and the low dose are efficacious”.
- If $H_{02}^\downarrow$ is rejected but $H_{01}^\downarrow$ is not rejected, then one infers the high dose is efficacious.

The interesting thing is, in testing $H_{0i}^\downarrow$, $i = 1, 2$ simultaneously, no multiplicity adjustment is needed to control FWER, the probability of rejecting any true null hypothesis. This is because the null spaces of $H_{0i}^\downarrow$, $i = 1, 2$ are disjoint. In other words, at most one null hypothesis can be true: it cannot be the case that the high dose is ineffective

(H_{02}) and that the high dose is effective but the low dose is ineffective (H_{01}), for example. We thus test each $H_{0i}^\downarrow$, $i = 1, 2$, at level- α .

Level α tests for each $H_{0i}^\downarrow$, $i = 1, 2$, are of course not unique. Note, however, a level- α test for $H_{01} : \theta_1 \leq \delta$ is also a level- α test for $H_{01}^\downarrow : \theta_1 \leq \delta$ and $\theta_2 > \delta$. For example, a test which rejects no more than 5% of the time when the low dose is ineffective, regardless of whether the high dose is effective, will reject no more than 5% of the time in particular when the low dose is ineffective and the high dose is effective.

So the simplest level- α test for $H_{0i}^\downarrow$ is to use a one-sided two-sample size α t-test based on T_i , $i = 1, 2$ defined in (2.2.4) for each $H_{0i}^\downarrow$ as shown in Table 2.3.

Table 2.3: Decision rules for partition testing to step-down doses		
Hypothesis	Rejection rule	Test level
$H_0^\cap : \theta_2 \leq \delta$	$T_2 > t_{\alpha, \nu}$	α
$H_{01} : \theta_1 \leq \delta$ and $\theta_2 > \delta$	$T_1 > t_{\alpha, \nu}$	α

In terms of the rejection rules in Table 2.3, the pre-determined D-steps (D stands for doses) proceed as follows.

- Step 1: If $T_2 > t_{\alpha, \nu}$, infer the high dose is efficacious and go to step 2; else stop.
- Step 2: If $T_1 > t_{\alpha, \nu}$, infer the low dose is efficacious and stop; else stop.

Note that, even though the method above controls FWER regardless of the shape of the true response function, it is recommended only when the response is expected

to be monotonically increasing (for otherwise it might stop too soon and miss an efficacious dose).

2.4 Partition testing with two doses and two endpoints

Suppose the dose-response study has both a primary endpoint and a secondary endpoint, so that the family of null hypotheses of interest includes both those in (2.2.1) and (2.2.2). We will show that, in addition to stepping through doses, the partitioning principle can also be used to derive multiple tests that step through endpoints in a pre-determined sequence.

2.4.1 Partition to step-down endpoints

Multiple testing procedures for dose-response studies with a primary endpoint and hierarchically ordered secondary endpoints were developed in Dmitrienko et al. (2006). In the case of two doses and two endpoints (one secondary endpoint), the procedure has to satisfy the condition that, for the same dose, the secondary endpoint is tested only if the primary endpoint is claimed to be efficacious, which is formally stated as:

Condition B (*order in endpoints*): *For the same dose, a secondary endpoint can not be claimed efficacious unless its primary endpoint has been shown to be efficacious.*

Within each endpoint, Dunnett's method was used by Dmitrienko et al. (2006) to adjust multiplicity for multiple doses. The FWER of the whole procedure was controlled at α using the principle of closed testing. Figure 2.1 is a graphical representation of their procedure with two doses and two endpoints.

Figure 2.1: Decision process in Dmitrienko et al. (2006)

Note: S means "Stop", c is calculated based on sample correlation between endpoints.

The procedure in Dmitrienko et al. (2006) can be reproduced by partitioning the parameter space with constraints on ordered endpoints for each dose, similar to partitioning to step-down through the doses technique as illustrated in Section 2.3.2. For example, for the high dose, we would test the following two hypotheses in a

step-down fashion (pre-determined E-steps).

$$H_{02}^{P\downarrow} : \quad \theta_2^P \leq \delta^P \quad (\text{Primary endpoint is not effective})$$

$$H_{02}^{S\downarrow} : \quad \theta_2^S \leq \delta^S \text{ and } \theta_2^P > \delta^P \quad (\text{Secondary endpoint is not effective, but primary endpoint is})$$

2.4.2 Partitioning to step-down both doses and endpoints

The procedure developed in Dmitrienko et al. (2006) steps through the endpoints in a pre-determined sequence, satisfying *condition B*, but does not necessarily step through the doses in a pre-determined sequence.

To step through both doses and endpoints, we test the following four partition hypotheses in a step-down fashion (pre-determined DE-steps):

- Step 1: Test $H_{02}^{P\downarrow} : \theta_2^P \leq \delta^P$ (High dose is not effective for primary endpoint).
 - If rejected, infer the high dose for the primary endpoint is effective and proceed to step 2; else stop.

- Step 2: Test two hypotheses:

$$H_{01}^{P\downarrow} : \theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \quad (\text{Low dose is not effective, but high dose is effective for primary endpoint})$$

$$H_{02}^{S\downarrow} : \theta_2^S \leq \delta^S \text{ and } \theta_2^P > \delta^P \quad (\text{High dose is not effective for secondary endpoint, but is effective for primary endpoint})$$

- If both hypotheses are rejected, infer both doses are effective for the primary endpoint; the high dose is also effective for the secondary endpoint and proceed to the next step.

- Otherwise, if only $H_{01}^{P\downarrow}$ is rejected, infer both doses are effective for the primary endpoint and stop.
- If only $H_{02}^{S\downarrow}$ is rejected, infer the high dose is effective for both endpoints and stop.
- Step 3: Test $H_{01}^{S\downarrow} : \theta_1^S \leq \delta^S$ and $\theta_2^S > \delta^S$ and $\theta_1^P > \delta^P$ and $\theta_2^P > \delta^P$ (Low dose is not effective for secondary endpoint, but is effective for primary endpoint and high dose is effective for both endpoints.)
 - If rejected, infer both doses are effective for both endpoints and stop. Otherwise stop.

The direction of these steps is presented in Table 2.4:

Table 2.4: Direction of the pre-determined DE-steps			
Endpoint\Dose	Low		High
Primary	Step 2 ²	←	Step 1
	↓		↓
Secondary	Step 3	←	Step 2 ²

²Proceed only if both hypotheses in step 2 are rejected

Decision tree

The rejection rules for these hypotheses in each step are presented in Table 2.5 in the form of the usual t-statistics $T_i^L, i = 1, 2, L = P, S$, as defined in (2.2.4).

Table 2.5: Critical values for the pre-determined DE-steps

Step	Hypothesis	Rejection rule	Test level
1	$H_{02}^{P\downarrow} : \theta_2^P \leq \delta^P$	$T_2^P > c_1$	α
2	$H_{01}^{P\downarrow} : \theta_1^P \leq \delta^P$ and $\theta_2^P > \delta^P$	$T_1^P > c_2$	MA ³
	$H_{02}^{S\downarrow} : \theta_2^S \leq \delta^S$ and $\theta_2^P > \delta^P$	$T_2^S > c_2$	MA
3	$H_{01}^{S\downarrow} : \theta_1^S \leq \delta^S$ and $\theta_2^S > \delta^S$ and $\theta_1^P > \delta^P$ and $\theta_2^P > \delta^P$	$T_1^S > c_1$	α

³Multiplicity adjustment needed

For steps 1 and 3, the simplest choice of critical value c_1 is $t_{\alpha, \nu}$. For step 2, since the null spaces of $H_{01}^{P\downarrow}$ and $H_{02}^{S\downarrow}$ are not disjoint, we need to adjust for multiplicity to make the FWER for the whole procedure controlled at level α .

The decision tree of this step-down procedure with pre-determined DE-steps is presented in Figure 2.2.

Computing the critical value c_2

According to Table 2.5, we need to find the critical value c_2 accounting for multiplicity for the two hypotheses in step 2. That is, one needs to focus on the joint distribution of (T_1^P, T_2^S) . Notice that between doses the observations are independent, but within a dose observations between endpoints are dependent; the numerators as well as the denominators of T_1^P and T_2^S are correlated through endpoints for the same dose. For one-sided tests, with the rejection rule of the form in (2.4.1), the supremum of the probability is obtained at the boundary values (involving θ_1^P and θ_2^S only, since

Figure 2.2: Decision tree for pre-determined DE-steps

neither T_1^P nor T_2^S involves θ_2^P). The critical value c_2 needs to satisfy:

$$\begin{aligned}
 1 - \alpha &\leq 1 - \sup_{\theta_1^P \leq \delta^P \text{ and } \theta_2^S \leq \delta^S} \Pr_\rho \{T_1^P > c_2 \text{ or } T_2^S > c_2\} \\
 &= \Pr_\rho \left\{ \frac{\bar{Y}_{1\cdot}^P - \bar{Y}_{0\cdot}^P - \theta_1^P}{\hat{\sigma}^P \sqrt{1/n_1 + 1/n_0}} \leq c_2 \text{ and } \frac{\bar{Y}_{2\cdot}^S - \bar{Y}_{0\cdot}^S - \theta_2^S}{\hat{\sigma}^S \sqrt{1/n_2 + 1/n_0}} \leq c_2 \right\} \\
 &= \Pr_\rho \left\{ \frac{Z_1}{\hat{\sigma}^P / \sigma^P} \leq c_2 \text{ and } \frac{Z_2}{\hat{\sigma}^S / \sigma^S} \leq c_2 \right\} \tag{2.4.1}
 \end{aligned}$$

$$= \int_0^\infty \int_0^\infty \Pr_\rho \{Z_1 \leq c_2 s_1 \text{ and } Z_2 \leq c_2 s_2\} \gamma_{\rho, \nu}(s_1, s_2) ds_1 ds_2 \tag{2.4.2}$$

where $(Z_1, Z_2) = (\frac{\bar{Y}_1^P - \bar{Y}_0^P - \theta_1^P}{\sigma^P \sqrt{1/n_1 + 1/n_0}}, \frac{\bar{Y}_2^S - \bar{Y}_0^S - \theta_2^S}{\sigma^S \sqrt{1/n_2 + 1/n_0}})$ with mean $\mathbf{0} = (0, 0)$, variances equal to 1 and covariance $\rho/(\sqrt{(n_0/n_1 + 1)(n_0/n_2 + 1)})$ ($= \rho/2$ in the balanced case), and $\gamma_{\rho, \nu}(s_1, s_2)$ is the joint density of $(\frac{\hat{\sigma}^P}{\sigma^P}, \frac{\hat{\sigma}^S}{\sigma^S})$.

Note that the definition of a bivariate t distribution, as originally introduced by Dunnett and Sobel (1954) and computed by the ProbMC function in SAS and the qmvnorm function in the R package mvtnorm, requires the denominators in (2.4.1) to be the same random variable $\hat{\sigma}^P \equiv \hat{\sigma}^S$, which is not the case here. In fact, computing (2.4.1) as if the distribution were bivariate t overestimates the probability, resulting in somewhat liberal critical values, as we will demonstrate. To avoid confusion, we call the distribution of (T_1^P, T_2^S) the *dual t* distribution. (This distribution was studied by Siddiqui 1967.)

The probability (2.4.2) involves the unknown parameter ρ . Below we describe five ways of approximating (2.4.2). The first four methods (Normal, Bivariate t, Dual t and Independent Standard Errors) require ρ be known or approximated. The last method computes by pretending T_1^P and T_2^S are independent, which we prove results in a conservative critical value.

Normal approximation

If the sample sizes n_1, n_2 are large, by Slutsky's Theorem, $\hat{\sigma}^P/\sigma^P$ and $\hat{\sigma}^S/\sigma^S$ converge in probability to one. Then, the joint distribution of (T_1^P, T_2^S) (under the true mean difference) is distributed asymptotically bivariate normal as (Z_1, Z_2) .

Bivariate t distribution

Assume $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ are perfectly correlated, or they are the same. This situation coincides with the bivariate t case. Specifically, the probability in (2.4.1) reduces to

$$\int_0^\infty \Pr_\rho \{Z_1 \leq c_2 s \text{ and } Z_2 \leq c_2 s\} \gamma_\nu(s) ds$$

where $\gamma_\nu(s)$ is the density of $\sqrt{\chi_\nu^2/\nu}$, and $\nu = \sum_{i=0}^2 (n_i - 1)$ is the degrees of freedom.

Dual t distribution

It can be shown that $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ are diagonal elements of a Wishart random matrix (see Section 2.7). We use simulations from a Wishart distribution (which depends on ρ , the correlation between endpoints from a single individual) and Monte Carlo methods to approximate the probability based on exact dual t distribution of the test statistics (T_1^P, T_2^S) .

Independent Standard Errors $\hat{\sigma}^P, \hat{\sigma}^S$

The difficulty in computing (2.4.1) is caused, in part, by pooling data across doses in estimating the standard errors σ^P, σ^S , resulting in the denominators of test statistics (T_1^P, T_2^S) , $(\frac{\hat{\sigma}^P}{\sigma^P}, \frac{\hat{\sigma}^S}{\sigma^S})$ being correlated. One can obtain a conservative upper bound of c_2 (see Theorem 2.4.1) by setting equation (2.4.3) equal to $= 1 - \alpha$ and solving for c_2 by pretending $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ are independent, where (2.4.3) can be viewed as a simplified version of (2.4.2).

$$\int_0^\infty \int_0^\infty \Pr_\rho \{Z_1 \leq c_2 s_1 \text{ and } Z_2 \leq c_2 s_2\} \gamma_\nu(s_1) \gamma_\nu(s_2) ds_1 ds_2 \quad (2.4.3)$$

Theorem 2.4.1. *Probability (2.4.1) is greater when $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ are correlated than when they are independent.*

Proof of Theorem 2.4.1: We first condition on Z_1 and Z_2 , and note that $c_2 > 0$

$$\begin{aligned}
& \Pr_\rho \left\{ \frac{Z_1}{\hat{\sigma}^P/\sigma^P} \leq c_2 \text{ and } \frac{Z_2}{\hat{\sigma}^S/\sigma^S} \leq c_2 \right\} \\
&= E_{(Z_1, Z_2)} \left[\Pr_\rho \left\{ \frac{z_1}{\hat{\sigma}^P/\sigma^P} \leq c_2 \text{ and } \frac{z_2}{\hat{\sigma}^S/\sigma^S} \leq c_2 \middle| (Z_1, Z_2) = (z_1, z_2) \right\} \right] \\
&= E_{(Z_1, Z_2)} \left[\Pr_\rho \left\{ \frac{z_1}{\hat{\sigma}^P/\sigma^P} \leq c_2 \text{ and } \frac{z_2}{\hat{\sigma}^S/\sigma^S} \leq c_2 \right\} \right] \\
&= E_{(Z_1, Z_2)} \left[\Pr_\rho \left\{ \hat{\sigma}^P/\sigma^P \geq \frac{z_1}{c_2} \text{ and } \hat{\sigma}^S/\sigma^S \geq \frac{z_2}{c_2} \right\} \right] \tag{2.4.4}
\end{aligned}$$

Since $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ are *associated* (see Theorem 6.1 of Karlin and Rinott 1981), for any given Z_1 and Z_2

$$\Pr_\rho \left\{ \hat{\sigma}^P/\sigma^P \geq \frac{z_1}{c_2} \text{ and } \hat{\sigma}^S/\sigma^S \geq \frac{z_2}{c_2} \right\} \geq \Pr \left(\hat{\sigma}^P/\sigma^P \geq \frac{z_1}{c_2} \right) \Pr \left(\hat{\sigma}^S/\sigma^S \geq \frac{z_2}{c_2} \right) \tag{2.4.5}$$

The result follows.

Independent t statistics T_1^P, T_2^S

To further break down the four dimensional integral in (2.4.3), we add the assumption of independent numerators (Z_1, Z_2) of (T_1^P, T_2^S) in addition to independent denominators, which is reasonable since the correlation between Z_1 and Z_2 is only $\rho/2$ in the balanced case. This method is more conservative than the previous method as shown in Theorem 2.4.2.

Theorem 2.4.2. *The probability (2.4.3) assuming $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ are independent is greater than the probability assuming test statistics (T_1^P, T_2^S) are independent. (This is equivalent to further assuming the numerators of (T_1^P, T_2^S) , (Z_1, Z_2) are independent.)*

Proof of Theorem 2.4.2:

Assuming independence of $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$, we have

$$\begin{aligned} & \Pr_{\rho} \left\{ \frac{Z_1}{\hat{\sigma}^P/\sigma^P} \leq c_2 \text{ and } \frac{Z_2}{\hat{\sigma}^S/\sigma^S} \leq c_2 \right\} \\ &= \int_0^\infty \int_0^\infty \Pr_{\rho} \{Z_1 \leq c_2 s_1 \text{ and } Z_2 \leq c_2 s_2\} \gamma_{\nu}(s_1) \gamma_{\nu}(s_2) ds_1 ds_2 \\ & \text{since } \text{Cov}(Z_1, Z_2) = \rho / \sqrt{(n_0/n_1 + 1)(n_0/n_2 + 1)} > 0 \end{aligned}$$

By Slepian's inequality (Corollary A.3.1 on p. 229 of Hsu, 1996)

$$\begin{aligned} & \geq \int_0^\infty \int_0^\infty \Pr(Z_1 \leq c_2 s_1) \Pr(Z_2 \leq c_2 s_2) \gamma_{\nu}(s_1) \gamma_{\nu}(s_2) ds_1 ds_2 \\ &= \int_0^\infty \Pr(Z_1 \leq c_2 s_1) \gamma_{\nu}(s_1) ds_1 \int_0^\infty \Pr(Z_2 \leq c_2 s_2) \gamma_{\nu}(s_2) ds_2 \\ &= \Pr \left(\frac{Z_1}{\hat{\sigma}^P/\sigma^P} \leq c_2 \right) \Pr \left(\frac{Z_2}{\hat{\sigma}^S/\sigma^S} \leq c_2 \right) \\ &= \Pr(T_1^P \leq c_2) \Pr(T_2^S \leq c_2) \end{aligned}$$

The degrees of conservatism for the methods introduced above are compared in terms of critical value c_2 for different combinations of correlation between endpoints ρ and error degrees of freedom ν in Table 2.6. Critical values for the dual t and Independent Standard Errors methods are based on 100,000 simulations.

As shown in Table 2.6, for each combination of ρ and ν , the normal approximation is somewhat liberal and the bivariate t approximation is slightly liberal. Assuming independence of $\frac{\hat{\sigma}^P}{\sigma^P}$ and $\frac{\hat{\sigma}^S}{\sigma^S}$ is slightly conservative, while assuming independence of t statistics is somewhat conservative. It should be noted that, in the setting of simultaneous efficacy and safety studies, Tamhane and Logan (2002) considered using the Bonferroni inequality and the bootstrap technique to compute such probabilities.

2.5 Analysis of the anti-psychotic drug data

We illustrate the step-down procedure with pre-determined DE-steps using the anti-psychotic drug example introduced at the beginning of the Chapter with $\delta^P =$

Table 2.6: Comparing conservatism in terms of critical value c_2

ν	Method	ρ^4				
		0	0.2	0.4	0.6	0.8
50	Bivariate t ⁵	2.0015	1.9972	1.9913	1.9833	1.9730
	Dual t ⁶	2.0026	1.9988	1.9934	1.9851	1.9741
	Ind. SE	2.0028	1.9988	1.9940	1.9866	1.9771
	Ind. T	2.0028	2.0028	2.0028	2.0028	2.0028
100	Bivariate t	1.9777	1.9737	1.9682	1.9607	1.9507
	Dual t	1.9783	1.9744	1.9692	1.9617	1.9513
	Ind. SE	1.9783	1.9745	1.9694	1.9623	1.9527
	Ind. T	1.9783	1.9783	1.9783	1.9783	1.9783
200	Bivariate t	1.9660	1.9622	1.9569	1.9495	1.9398
	Dual t	1.9659	1.9627	1.9572	1.9498	1.9401
	Ind. SE	1.9664	1.9628	1.9573	1.9503	1.9408
	Ind. T	1.9664	1.9664	1.9664	1.9664	1.9664
∞	Normal ⁷	1.9545	1.9508	1.9456	1.9385	1.9289

⁴Note ρ refers to the correlation between endpoints for a single individual. The correlation between numerators (Z_1, Z_2) of the test statistics, is assumed to be $\rho/2$.

⁵The critical values of the bivariate t are obtained by the *qmv*t function in the R package *mvtnorm*.

⁶The Wishart is generated first using Bartlett's decomposition and then appropriately transformed (see Johnson 1987).

⁷The bivariate normal quantiles are obtained by *qmvnorm* of the R package *mvtnorm*.

$\delta^S = 0$. Using pooled standard error from these three groups, the t-statistics corresponding to each dose and endpoint and the error degrees of freedom are:

$$T_1^P = -1.7499, T_2^P = -3.7318, T_1^S = -0.2116, T_2^S = -0.4212 \text{ and } \nu = 151.$$

If we want to control the FWER at 5% level, then $c_1 = t_{.05,151} = 1.6550$, while $c_2 = 1.9702$, based on the independent t method (which is appropriate because the correlation is unknown). Note in this particular example, a lower score indicates better drug effect. To make this ordering consistent with our hypotheses set up, we take the negatives of the t-statistics and then apply our procedure.

- Step 1: Is $3.7318 > 1.6550$? Yes, go to step 2.
- Step 2: Is $1.7499 > 1.9702$ or $0.4212 > 1.9702$? No, stop.

So we infer that the high dose (600 mg/day) is efficacious for the primary endpoint only, and no evidence of efficacy can be made for the low dose (75 mg/day).

2.6 Conclusion

In this Chapter, we presented a systematic way of constructing null hypotheses by partitioning the parameter space according to conditions that order inferences by dose (high to low), or endpoint (primary to secondary), or both. The hypotheses so constructed automatically form step-down procedures (pre-determined D-steps, E-steps, or DE-steps) and control the proper error rate at a pre-specified level. This way of test construction can be generalized to situations with $k > 2$ doses and $m > 2$ endpoints. The technique is as follows. Start by writing down the null hypotheses whose rejections correspond to desired inferences on primary and secondary endpoints. Then, following stated conditions for the decision process, make each 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). Adjust for multiplicity within a step (only) if the null hypotheses are not disjoint, but do not adjust for multiplicity between steps.

An important issue in multiple endpoints problems is how to deal with correlations among the test statistics induced by correlations among measurements on the multiple endpoints. We made a systematic study of liberalism and conservatism of five approximation methods.

Graphical representation makes the decision process clear and easy to understand. Multiplicity adjustment may be needed for some steps but not for all. Based on the decision tree, there are two general rules. First, if there are two or more statistics involved in one box, then a multiplicity adjustment may be needed. Second, if there are two or more branches from a box, and the hypotheses corresponding to these boxes are not disjoint, then further multiplicity adjustment may be needed among those boxes.

2.7 Appendix: Model for dose-response studies with two endpoints

Suppose we have samples $Y'_{0r} = (Y_{0r}^P, Y_{0r}^S), r = 1, \dots, n_0$ from a placebo group, and $Y'_{1r} = (Y_{1r}^P, Y_{1r}^S), r = 1, \dots, n_1, Y'_{2r} = (Y_{2r}^P, Y_{2r}^S), r = 1, \dots, n_2$ from two dose groups, measuring two endpoints. Let $\boldsymbol{\mu}_0 = (\mu_0^P, \mu_0^S)', \boldsymbol{\mu}_1 = (\mu_1^P, \mu_1^S)', \boldsymbol{\mu}_2 = (\mu_2^P, \mu_2^S)'$ be the group means, $\Sigma = \begin{pmatrix} (\sigma^P)^2 & \rho\sigma^P\sigma^S \\ \rho\sigma^P\sigma^S & (\sigma^S)^2 \end{pmatrix}$ be the covariance matrix of ϵ_n , the n th row of $\boldsymbol{\epsilon}$, and $\mathbf{1}_n$ be a column vector of 1s of length n . The model can be written in the following form:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} Y_{01}^P & Y_{01}^S \\ \vdots & \vdots \\ Y_{0n_0}^P & Y_{0n_0}^S \\ Y_{11}^P & Y_{11}^S \\ \vdots & \vdots \\ Y_{1n_1}^P & Y_{1n_1}^S \\ Y_{21}^P & Y_{21}^S \\ \vdots & \vdots \\ Y_{2n_2}^P & Y_{2n_2}^S \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_0^P & \mu_0^S \\ \mu_1^P & \mu_1^S \\ \mu_2^P & \mu_2^S \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X} \begin{pmatrix} \mu'_0 \\ \mu'_1 \\ \mu'_2 \end{pmatrix} + \boldsymbol{\epsilon}$$

$$\text{The projection matrix } \mathbf{H} = \begin{pmatrix} \frac{1}{n_0}\mathbf{1}_{n_0}\mathbf{1}_{n_0}' & 0 & 0 \\ 0 & \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}' & 0 \\ 0 & 0 & \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}' \end{pmatrix} = \begin{pmatrix} H_0 & 0 & 0 \\ 0 & H_1 & 0 \\ 0 & 0 & H_2 \end{pmatrix}.$$

$$\mathbf{I} - \mathbf{H} = \begin{pmatrix} I - H_0 & 0 & 0 \\ 0 & I - H_1 & 0 \\ 0 & 0 & I - H_2 \end{pmatrix} \text{ and} \\ N\hat{\Sigma} = N\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} = N \sum_{i=0}^2 \mathbf{Y}'_i(I - H_i)\mathbf{Y}_i$$

By Result 7.10. on p. 390 of Johnson and Wichern (2002), $N\hat{\Sigma}$ follows the Wishart distribution $W_2(N - 3, \Sigma)$, where $N = \sum_{i=0}^2 n_i$, and $N - 3$ is the degrees of freedom.

Notice, the diagonal elements of $N\hat{\Sigma}$ are $(N - 3)(\hat{\sigma}^P)^2$ and $(N - 3)(\hat{\sigma}^S)^2$, where $\hat{\sigma}^P = \sqrt{\frac{\sum_{i=0}^2 \sum_{r=1}^{n_i} (Y_{ir}^P - \bar{Y}_{i\cdot}^P)^2}{N - 3}}$ and $\hat{\sigma}^S = \sqrt{\frac{\sum_{i=0}^2 \sum_{r=1}^{n_i} (Y_{ir}^S - \bar{Y}_{i\cdot}^S)^2}{N - 3}}$. To obtain the joint distribution of $(\frac{\hat{\sigma}^P}{\sigma^P}, \frac{\hat{\sigma}^S}{\sigma^S})$, let $U = \text{diag}\{(\sigma^P)^{-1}, (\sigma^S)^{-1}\}$, then $NU'\hat{\Sigma}U \sim W_2\left(N - 3, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with diagonal elements $(N - 3)(\frac{\hat{\sigma}^P}{\sigma^P})^2$ and $(N - 3)(\frac{\hat{\sigma}^S}{\sigma^S})^2$.

CHAPTER 3

DECISION PATH APPROACH

3.1 Primary-secondary endpoints with unordered doses

In this chapter, we propose a new approach to analyzing dose-response data with primary-secondary endpoints. When multiple doses are coupled with multiple endpoints in a clinical trial, the question of interest is for which dose and endpoint combinations are the compound effective. What distinguishes the primary-secondary endpoints situation is, for a given dose, efficacy in the secondary endpoint is a relevant question only if efficacy in the primary endpoint has been shown.

3.1.1 Closed testing

Closed testing is a general principle that can be used to control Familywise Error Rate (FWER) strongly in testing multiple hypotheses. It tests all intersections of the individual hypotheses, rejecting a hypothesis only if all intersection hypotheses implying it are rejected. So, with multiple doses and endpoints, it tests every intersection hypothesis that states a particular subset of dose and endpoint combinations are not effective, then collates the results for individual hypotheses from all such tests. The pioneering paper Dmitrienko et al. (2006) used closed testing to test for efficacy in primary and secondary endpoints.

Since closed testing places no restriction on how each intersection hypothesis is tested, it may lead to a claim of efficacy in the secondary endpoint even if efficacy in the primary endpoint has not been shown. To avoid this, Dmitrienko et al. (2006) placed conditions on the choice of test statistics and critical values for testing the intersection hypotheses. Such strategic choice becomes increasingly complex to achieve as the number of doses and endpoints increases, however, because the number of intersection hypotheses, the number of possible subsets of all dose-endpoint combinations, increases dramatically as the number of doses increases.

3.1.2 Iterative graphical method

Bretz et al. (2009) took a graphical approach to represent and perform closed tests. In the multiple doses primary-secondary endpoints setting, tests are represented by directed, weighted graphs, with each node corresponding to a hypothesis that a dose-endpoint combination lacks efficacy. The graphical approach, very effective in communicating the structure of the tests, differs from the partition testing approach in this dissertation in the following respects.

- While the graphical approach reduces the number of hypotheses actually tested by guaranteeing shortcuts to closed testing, partition testing formulates fewer hypotheses to start with;
- While the graphical approach focuses on testing each hypothesis using a weighted Bonferroni test, partition testing focuses on testing each hypothesis using the joint distribution of the test statistics.

3.1.3 Partition testing

Partition testing (Berger 1982) is an alternative to closed testing. It gives insights on confidence sets for step-wise tests (Stefansson, Kim, and Hsu 1988), and can be more powerful than closed testing (Finner and Strassburger 2002). It can also simplify problem formulation when decision-making follows specific paths (Hsu and Berger 1999). In the primary-secondary endpoints problem, instead of testing intersections of null hypotheses each stating a particular dose-endpoint combination is not efficacious, partition testing forms (only) null hypotheses appropriate for the *decision paths* of the problem in a way that partitions the parameter space. Collating the results from testing the partition hypotheses then leads to decisions that are guaranteed to be correct, at least with a pre-specified high probability.

Even though in simple settings partition testing often gives the same inference as closed testing (cf. Huang and Hsu 2007), partition testing seems to have a genuine advantage in the primary-secondary endpoints setting. It significantly reduces the number of hypotheses to be tested compared to closed testing. For example, just using paths between endpoints within each dose reduces the number of hypotheses from $(2^m)^k - 1$ to $(m+1)^k - 1$, where k is the number of doses and m is the number of endpoints. The reduction is even more dramatic when there are paths between doses as well. With a smaller set of hypotheses to be tested, development and execution of multiple tests become much easier.

3.2 A motivating example

Dmitrienko, Offen and Westfall (2003) gave an example of a dose-response study conducted to evaluate the effect of a hypertension drug, comparing three doses (low, medium and high) to the placebo. Summary statistics are given in Table 3.1.

Table 3.1: Summary statistics for the hypertension example

Endpoint		Dose			
		Placebo	Low	Medium	High
	Sample size	29	23	24	24
Change in SBP (mmHg) (Primary)	Mean	4.02	−2.16	−5.03	−2.60
	SD ⁸	7.85	8.71	11.01	8.96
Change in DBP (mmHg) (Secondary)	Mean	2.31	−0.67	−3.18	−1.44
	SD	5.88	5.51	6.89	6.15

⁸standard deviation

For illustration, let the primary endpoint be reduction in systolic blood pressure (SBP), and let the secondary endpoint be reduction in diastolic blood pressure (DBP). That is, at a given dose, the question of whether the compound is efficacious in terms of DBP is only relevant if it is efficacious in terms of SBP.

3.3 Decision path principle

A sequence of potential inferences is termed a *path*. If a hypotheses testing problem involves some paths then one should follow the Decision Path Principle.

Decision Path Principle: *Null hypotheses should be formulated so that decision-making naturally follows logical paths.*

To apply the Decision Path Principle, we use partition testing. In fact, partition testing has already been used to formulate hypotheses in the case of a path between doses with single endpoint as shown in Hsu and Berger (1999).

In this chapter, we investigate hypotheses testing problems with pre-specified paths in the setting of a single primary endpoint and multiple secondary endpoints coupled with multiple doses. Consider a dose-response study comparing k doses of a drug to a placebo. Efficacy of the drug is assessed using a single primary endpoint. In addition, $m - 1$ secondary endpoints are measured for supportive evidence of the drug.

Let μ_{ij} denote the mean response of group i for endpoint j , $i = 0, 1, \dots, k$, $j = 1, \dots, m$, where $i = 0$ denotes the placebo group and $j = 1$ denotes the primary endpoint. Define $\theta_{ij} = \mu_{ij} - \mu_{0j}$ to be the difference in mean efficacy measurement between dose group i and the placebo group for endpoint j . To be general, we allow the clinically meaningful differences for endpoints to differ, denoted as δ_j for endpoint j . Assuming a larger measurement indicates a better treatment, the statistical inference of interest is to test, for each dose-endpoint combination,

$$H_{0ij} : \theta_{ij} \leq \delta_j, \quad i = 1, \dots, k, \quad j = 1, \dots, m. \quad (3.3.1)$$

In this chapter, we assume the vector of measurements from patient r in dose group i , $\mathbf{Y}_{ir} = (Y_{ir1}, Y_{ir2}, \dots, Y_{irm})'$, comes from a model:

$$\mathbf{Y}_{ir} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ir}, \quad i = 0, 1, \dots, k; \quad r = 1, \dots, n_i \quad (3.3.2)$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{im})'$, $\boldsymbol{\epsilon}_{ir} = (\epsilon_{ir1}, \dots, \epsilon_{irm})'$ are i.i.d. m -dimensional random vectors from a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Define T_{ij} to be the two sample t-statistic for testing the null

hypothesis H_{0ij} ,

$$T_{ij} = \frac{\bar{Y}_{i.j} - \bar{Y}_{0.j} - \delta_j}{\hat{\sigma}_j \sqrt{1/n_i + 1/n_0}} \quad i = 1, \dots, k, j = 1, \dots, m \quad (3.3.3)$$

where $\bar{Y}_{i.j} = \frac{1}{n_i} \sum_{r=1}^{n_i} Y_{irj}$, $\hat{\sigma}_j = \sqrt{\frac{\sum_{i=0}^k \sum_{r=1}^{n_i} (Y_{irj} - \bar{Y}_{i.j})^2}{\sum_{i=0}^k (n_i - 1)}}$.

Since the efficacy claim of the new experimental drug is based on the primary endpoint alone, additional claims on secondary endpoints are of interest only if the primary endpoint has shown efficacy. If more than one secondary endpoint is measured, ordering the secondary endpoints within each dose makes for logical statements in the Clinical Study section of a drug label:

Efficacy in [the primary endpoint] has been established for all doses. In addition, efficacy in [secondary endpoint 1] has been established for medium and high doses. In addition, efficacy in [secondary endpoint 2] has been established for the medium dose.

Note this setting is different from Dmitrienko et al. (2006) or Chapter 2 as it requires the ordering of secondary endpoints within each dose and it does not order doses. If prior information suggests an ordering among the doses, the principle proposed in this paper is still applicable and procedures can be easily modified.

3.4 Paths with ordered endpoints

In this section, we show how to apply the Decision Path Principle by using partition testing to situations in which there is an ordering among the endpoints, but no such ordering among the doses. Decision paths are thus within, but not between, doses. Figure 3.1 illustrates such paths with k doses and m endpoints.

The lower ordered secondary endpoints, in this situation, are required to be tested only if higher ordered ones are proven efficacious. Following the Decision Path Principle, null hypotheses should be formulated so that, for a given dose i , the inference

Figure 3.1: Decision paths for k doses and m endpoints, with one path for each dose going from endpoint 1 (primary endpoint) to endpoint m .

$\theta_{ij} > \delta_j$ in (3.3.1) is not given unless the inference $\theta_{ij^*} > \delta_{j^*}$ is established for all $j^* < j$.

In the case of a single dose ($k = 1$), the path between endpoints leads to the predetermined steps method of Hsu and Berger (1999).

In the case of a single endpoint ($m = 1$) and no paths between doses, it is known that partition testing leads to step-down or step-up methods. If the joint distribution of the test statistics is ignored and each partition hypothesis is tested using the Bonferroni adjustment, then partition testing leads to Holm's (1979) step-down

method. If the test statistics are assumed to be independent (or have a joint distribution with the MTP_2 property), and each partition hypothesis is tested using what Huang and Hsu (2007) called the Sime-Hochberg adjustment, then partition testing leads to Hochberg's (1988) step-up method. If the test statistics have a multivariate t distribution for which Dunnett's (1955) method is applicable, then this formulation leads to a step-down version of Dunnett's method, as described in Section 3.1.1.2 of Hsu (1996).

An advantage of partition testing over closed testing in the presence of paths described in Figure 3.1 is that it drastically reduces the number of hypotheses to be tested. The intuition is illustrated in Figure 3.2 for the case of three doses and three endpoints. For each of the nine partition hypotheses illustrated in Figure 3.2, the nine small boxes represent lack of efficacy of nine dose-endpoint combinations, with rows being endpoints and columns being doses. Thus, any non-empty intersection of the nine boxes (sets) represents a closed testing intersection hypothesis. So there are $2^{mk} - 1 = 2^9 - 1 = 511$ such hypotheses. However, as will be shown in (3.9.2) of Section 3.9.1 in the proof of Theorem 3.4.1, each partition hypothesis is represented by a non-empty intersection with *at most one box (light colored) from each column*. Thus, the number of partition hypotheses to be tested reduces to $(m + 1)^k - 1 = 4^3 - 1 = 63$.

Formally, we state

Theorem 3.4.1. *For a clinical trial with m ordered endpoints and k (unordered) doses, partition testing following the Decision Path Principle leads to testing $(m + 1)^k - 1$ hypotheses which partition the parameter space, with results that*

(i) *automatically respect the pre-specified paths;*

	Low	Medium	High		Low	Medium	High		Low	Medium	High
P	$\theta_{11} \leq \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} \leq \delta_1$	P	$\theta_{11} > \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} > \delta_1$	P	$\theta_{11} > \delta_1$	$\theta_{21} > \delta_1$	$\theta_{31} > \delta_1$
S ₁				S ₁	$\theta_{12} \leq \delta_2$		$\theta_{32} > \delta_2$	S ₁	$\theta_{12} \leq \delta_2$	$\theta_{22} \leq \delta_2$	$\theta_{32} \leq \delta_2$
S ₂				S ₂			$\theta_{33} \leq \delta_3$	S ₂			
	Low	Medium	High		Low	Medium	High		Low	Medium	High
P	$\theta_{11} > \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} > \delta_1$	P	$\theta_{11} > \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} > \delta_1$	P	$\theta_{11} \leq \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} > \delta_1$
S ₁	$\theta_{12} \leq \delta_2$		$\theta_{32} \leq \delta_2$	S ₁	$\theta_{12} > \delta_2$		$\theta_{32} > \delta_2$	S ₁			$\theta_{32} \leq \delta_2$
S ₂				S ₂	$\theta_{13} \leq \delta_3$		$\theta_{33} \leq \delta_3$	S ₂			
	Low	Medium	High		Low	Medium	High		Low	Medium	High
P	$\theta_{11} > \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} \leq \delta_1$	P	$\theta_{11} > \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} > \delta_1$	P	$\theta_{11} > \delta_1$	$\theta_{21} \leq \delta_1$	$\theta_{31} > \delta_1$
S ₁	$\theta_{12} > \delta_2$			S ₁	$\theta_{12} > \delta_2$		$\theta_{32} \leq \delta_2$	S ₁	$\theta_{12} > \delta_2$		$\theta_{32} > \delta_2$
S ₂	$\theta_{13} > \delta_3$			S ₂	$\theta_{13} > \delta_3$			S ₂	$\theta_{13} > \delta_3$		$\theta_{33} > \delta_3$

Figure 3.2: Examples of nine (out of 63) partition hypotheses with three doses (low, medium, high) and three endpoints (P, S1, S2). Each partition hypothesis is represented by a non-empty intersection with at most one box (light colored with “ $\leq$ ” sign) from each column. For example, the partition hypotheses in the left column from top to bottom are: $\{\theta_{11} \leq \delta_1 \text{ and } \theta_{21} \leq \delta_1 \text{ and } \theta_{31} \leq \delta_1\}$, $\{\theta_{11} > \delta_1 \text{ and } \theta_{12} \leq \delta_2 \text{ and } \theta_{21} \leq \delta_1 \text{ and } \theta_{31} > \delta_1 \text{ and } \theta_{32} \leq \delta_2\}$, $\{\theta_{11} > \delta_1 \text{ and } \theta_{12} > \delta_2 \text{ and } \theta_{13} > \delta_3 \text{ and } \theta_{21} \leq \delta_1 \text{ and } \theta_{31} \leq \delta_1\}$.

(ii) *control FWER strongly at level α , if each partition hypothesis is tested at level α .*

The proof of this theorem is given in Section 3.9.1. In the next section, we illustrate the idea behind this theorem in the case of two doses and two endpoints. The case of three doses and two endpoints will be illustrated in Section 3.7 by analyzing the motivating real-data example.

3.4.1 Partition following decision paths

If a single secondary endpoint is involved, the decision paths in Figure 3.1 become what is shown in Figure 3.3 below. For notational convenience, instead of using the second subscript to index the endpoint, we use superscripts P and S to denote primary and secondary endpoints.

Figure 3.3: Decision paths for low and high doses with one primary and one secondary endpoint.

Given the decision paths in Figure 3.3, the parameter space is partitioned in two stages:

- Path partition: Partition within each path.

- Disjointness partition: Further partition by taking intersections to make hypotheses between paths disjoint.

The path partition is within each dose i . Starting with the primary endpoint, we test $H_{0i}^P : \theta_i^P \leq \delta^P$. If this hypothesis is rejected, then efficacy in the primary endpoint has been established.

Then follow the path to the secondary endpoint. But, instead of testing $H_{0i}^S : \theta_i^S \leq \delta^S$, we make it disjoint with H_{0i}^P and test $H_{0i}^{*S} : \theta_i^S \leq \delta^S$ and $\theta_i^P > \delta^P$. If both H_{0i}^P and H_{0i}^{*S} are rejected, then we logically conclude efficacy in both the primary and the secondary endpoints.

Whereas the reason for the path partition is that inference in the secondary endpoint is irrelevant unless efficacy in the primary endpoint is established, the disjointness partition is for proper multiplicity adjustment. If two or more hypotheses can be true simultaneously, then a multiplicity adjustment is needed in testing them. If the hypotheses tested are made disjoint, then since (at most) one of them is true, FWER is controlled strongly if each is tested at level α .

Between paths, connect an *edge* between subspaces of the parameter space that are not disjoint, as illustrated in Figure 3.4. Edges represent hypotheses that can be true simultaneously. (There is no edge between $\{\theta_1^S > \delta^S$ and $\theta_1^P > \delta^P\}$ and $\{\theta_2^S > \delta^S$ and $\theta_2^P > \delta^P\}$, since the intersection of these two hypotheses, the ideal situation of efficacy in all doses and endpoints, does not need to be tested.) Take intersections of connected subspaces to form new hypotheses. The resulting set of hypotheses, as presented in Table 3.2, partitions the parameter space. Therefore, so long as each partition hypothesis is tested at level α , the FWER is controlled strongly at level α .

Inferences on which of the $m \times k$ combinations of dose and endpoint are efficacious are then obtained by collating the results from the $(m + 1)^k - 1$ tests.

Specifically, the inference $\theta_{ij} > \delta_j$ is made if all hypotheses implying that $\theta_{ij} \leq \delta_j$ could be true are rejected. This includes partition hypotheses which do not explicitly state an inequality for θ_{ij} . For example, the hypothesis which states $\theta_1^P \leq \delta^P$ and $\theta_2^P \leq \delta^P$ includes the possibility that $\theta_i^S \leq \delta^S$ (as well as the possibility $\theta_i^S > \delta^S$), and must be rejected before inference on any secondary endpoint is given.

Figure 3.4: Graphical representation of two stages of partitioning in the setting of Figure 3.3.

3.4.2 Advantages of partition testing

Advantages of the partitioning approach now become clear.

Table 3.2: Partition hypotheses following decision paths in Figure 3.3

Partition hypothesis	Rejection rule
$\theta_1^P \leq \delta^P, \theta_2^P \leq \delta^P$	$t_1^P > c_1$ or $t_2^P > c_1$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_2^S \leq \delta^S$	$t_1^P > c_1$ or $t_2^S > c_2$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_2^S > \delta^S$	$t_1^P > c_3$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_1^S \leq \delta^S$	$t_2^P > c_1$ or $t_1^S > c_2$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_1^S > \delta^S$	$t_2^P > c_3$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S$	$t_1^S > c_1$ or $t_2^S > c_1$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S$	$t_1^S > c_3$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S \leq \delta^S$	$t_2^S > c_3$

NOTE: The comma “,” in each partition hypothesis represents “and”.

First, the number of hypotheses tested is reduced from the $(2^2)^2 - 1 = 15$ that would be tested by closed testing to $(2 + 1)^2 - 1 = 8$, as shown in Table 3.2, because intersection hypotheses incompatible with decision paths are omitted. For example, the intersection hypothesis $H_0 : \theta_1^P \leq \delta^P$ and $\theta_1^S > \delta^S$ is not tested, because if there is no efficacy in the primary endpoint, efficacy in the secondary endpoint is of no interest.

Second, decision-making automatically respects the principle that inference on the secondary endpoints should only be given when efficacy in the primary endpoint is shown. If both H_{0i}^P and $H_{0i}^{\star S}$ are rejected, then we logically conclude efficacy in both the primary and the secondary endpoints. But if H_{0i}^P fails to be rejected, then even if $H_{0i}^{\star S}$ is rejected, there is no point reporting the non-informative rejection of $H_{0i}^{\star S}$, and partition testing makes no inference on the secondary endpoint.

3.4.3 Execution of partition testing

Level- α tests for the partition hypotheses are not unique. One can use a generalized likelihood ratio test which is asymptotically optimal, but computationally

intensive. One can also use strong partitioning (Finner and Strassburger 2002) to obtain more powerful tests. We recommend a test of the form that is the union of rejection regions of the individual two sample t statistics. This test does not require the hard work of strong partitioning or too much computing time but is reasonably powerful.

For each partition hypothesis, if we let N denote the (i, j) pairs for which the hypothesis states $\theta_{ij} \leq \delta_j$, then it is common partition testing practice to test the null hypothesis which states $\theta_{ij} \leq \delta_j$ for all $(i, j) \in N$ instead for convenience. For example, instead of testing the hypothesis which states

$$\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S \leq \delta^S \quad (3.4.1)$$

one would test the hypothesis which states

$$\theta_1^P \leq \delta^P \text{ and } \theta_2^S \leq \delta^S. \quad (3.4.2)$$

Strong control of FWER is maintained, because a level- α test for (3.4.2) is also a level- α test for (3.4.1).

Suppose a large value of t_{ij} is indicative of $\theta_{ij} > \delta_j$. We recommend tests of the form which reject if $\max_{(i,j) \in N} (t_{ij} - c_{ij}) > 0$. For example, we would reject hypothesis (3.4.2), thus (3.4.1), if $t_1^P > c_1$ or $t_2^S > c_2$. The rejection rules in Table 3.2 are of this form, and the rules in setting the critical values can be found in Section 3.5.

3.5 Computer algorithm

Given the decision paths for individual hypotheses in Figure 3.1, we recommend the following computer implementation of our Decision Path approach.

1. Generate all partition hypotheses.

2. Choose tests for each partition hypothesis.
3. Calculate critical values for testing each hypothesis so that the Type I error rate is controlled at level α .
4. Make decisions on each partition hypothesis.
5. Collate the results.

Generating all partition hypotheses is facilitated by visualizing each hypothesis as a non-empty intersection of at most one box from each column, in a row-column representation with each of the m rows corresponding to an endpoint, and each of the k columns corresponding to a dose.

The level- α test for each partition hypothesis is not unique. However, as mentioned in Section 3.4.3, we recommend a test of the form that is the union of rejection regions of the two sample t statistics.

As each partition hypothesis may involve multiple t statistics from different endpoints or doses, there are infinitely many ways to set the critical values for keeping the Type I error rate controlled. We recommend that if an entire row or column is efficacious, then multiplicity should be adjusted as if this row or column were not included. Specifically,

- **Dose Multiplicity Adjustment:** If a dose is shown to be efficacious in all endpoints, then that dose should no longer be included in the multiplicity adjustment. (This would be expected in testing multiple doses, in the case of a single endpoint.)

- **Endpoint Prioritization:** If efficacy in a higher ordered endpoint is shown for all doses, then testing proceeds as if the next highest ordered endpoint were one order higher. (This would be expected in testing multiple ordered endpoints, in the case of a single dose.)

3.5.1 Computation within and between endpoints

We discuss the calculation of critical values in the two-dimensional (two endpoints) case under the assumption of model (3.3.2) with ρ being the correlation between the primary and secondary endpoints.

Test statistics involving the same endpoint, such as (t_1^P, t_2^P) or (t_1^S, t_2^S) , have a multivariate t distribution. The numerators have known, non-ignorable correlations, due to comparisons to a common control. In the case of a balanced one-way design, this correlation is 0.5, for example. Probabilities associated with such statistics can be computed using the factor-analytic algorithm of Hsu (1992), the default multiplicity adjustment for Dunnett's method under the LSMeans option in SAS, or the pseudo-Monte Carlo algorithm of Genz and Bretz (1999).

Test statistics involving different endpoints, such as (t_1^P, t_2^S) or (t_2^P, t_1^S) , do *not* have a multivariate t distribution however. Besides the numerators not being scaled by a common random variable, both the numerators and the denominators have joint distributions that depend on the unknown correlation ρ . It is, thus, difficult to compute probabilities associated with the maximum of such statistics. Liu et al. (2007) pointed out that critical values computed using algorithms meant for the multivariate t are in fact somewhat liberal.

Correlations between numerators of test statistics involving different endpoints are typically small though. This follows from the important fact (proven in 3.9.2 of Section 3.9.1) that, for each partition hypothesis, doses corresponding to different endpoints are *disjoint*. For a balanced one-way design, correlation between such test statistics in any partition hypothesis equals $\rho/2$. If the correlation ρ between the primary and the secondary endpoint is .3 say, then the correlation between the numerators of the test statistics is only .15. We thus consider computing critical values ignoring the correlation between endpoints.

One can always use the Bonferroni inequality (between endpoints) to obtain conservative critical values. This is the proposal of Xu et al. (2008), in the setting of gate-keeping methods. But if it is known that $\rho \geq 0$, then less conservative critical values can be computed as if $\rho = 0$. (See Liu et al. 2007 for a proof of conservatism.)

We thus recommend using the exact joint multivariate t distribution for the multiplicity adjustment within each endpoint, while using either the independence (for positive correlation) or disjointness of events (for other cases) to adjust for multiplicity between endpoints.

3.5.2 Recursive computation between endpoints

After using the Dose Multiplicity Adjustment for testing a partition hypothesis, multiplicity is adjusted according to the number of non-effective *doses* in that hypothesis. For example, each hypothesis in the top two rows of Figure 3.5 is adjusted for a multiplicity of three doses, while in the bottom row, multiplicity is adjusted only for two doses for hypotheses of the two left squares, and there is no multiplicity adjustment for the hypothesis of the right square.

After applying Endpoint Prioritization, we compute critical values for test statistics in the order of the endpoints, starting with the primary endpoint. For convenience of collation of results, we use the same critical value for testing different doses within each endpoint.

Consider computing critical values for test statistics of partition hypotheses which state that doses $i, i \in I$, are not efficacious. We first compute the critical value for the hypothesis that every dose $i, i \in I$, of the primary endpoint lacks efficacy. For example, for $I = \{1, 2, 3\}$ in Figure 3.5, this critical value is 2.107, corresponding to the partition hypothesis of the upper left square. This critical value (2.017 in Figure 3.5) is then the starting candidate critical value for all test statistics, for hypotheses that state doses $i, i \in I$, lack efficacy, the six squares in the top two rows.

The critical value for test statistics involving each successively ordered endpoint is then computed *recursively*, as follows.

If the test remains level α using the candidate critical value, then it is kept. This is the case for the critical value 2.107 for endpoint S_1 , in the top middle and right squares of the first row.

Otherwise, the candidate critical value is replaced by the smallest value possible while keeping the test at level α . In Figure 3.5, after one iteration, the critical value becomes 2.185 for the left and middle squares and 2.255 for the right square in the second row. After two iterations, the critical value becomes 2.377 for the middle square in the first row.

The last row of Figure 3.5 illustrates the recursive computation for $I = \{2, 3\}$ and $I = \{2\}$. The detailed step by step algorithm is presented in Appendix B of this chapter.

	Low	Medium	High		Low	Medium	High		Low	Medium	High
P	2.107	2.107	2.107	P		2.107		P			
S ₁				S ₁	2.107			S ₁	2.107	2.107	2.107
S ₂				S ₂			2.377	S ₂			
	Low	Medium	High		Low	Medium	High		Low	Medium	High
P		2.107		P		2.107		P	2.107	2.107	
S ₁	2.185		2.185	S ₁				S ₁			2.255
S ₂				S ₂	2.185		2.185	S ₂			
	Low	Medium	High		Low	Medium	High		Low	Medium	High
P		1.954	1.954	P		1.954		P		1.672	
S ₁				S ₁			2.056	S ₁			
S ₂				S ₂				S ₂			

Figure 3.5: Critical values for the partition hypotheses in Figure 3.2 using two sample t statistics with 57 degrees of freedom, controlling FWER strongly at 0.05.

3.6 Performance comparisons

We compare the performance of the following three procedures:

- DPjd: Decision Path approach with critical value calculation based on joint distribution within endpoint and disjointness (Bonferroni inequality) across endpoints.
- GKjj: Gatekeeping procedure in Dmitrienko et al. (2006) with critical value calculation based on the joint multivariate t distribution both within and across endpoints.
- GKjd: Gatekeeping procedure in Xu et al. (2008) with critical value calculation based on the joint distribution within endpoint and disjointness (Bonferroni inequality) across endpoints.

The simulation study is in the setting of a hypothetical clinical trial comparing three dose groups, low, medium, and high to a placebo group. There are 50 patients in each group. Two endpoints, one primary and one secondary are measured with variances equal to 1 and correlation ρ . Four different scenarios for the mean matrix are used, as presented in Table 3.3. In addition, four values of $\rho = 0.3, 0.5, 0.7, 0.9$ are chosen to match the scenarios presented in Dmitrienko et al. (2006) or Xu et al. (2008). The results are based on 10000 simulations with FWER controlled strongly at 0.025.

The probability of rejecting each individual hypothesis is presented in Table 3.4 for all three procedures. DPjd has a clear advantage over GKjd and GKjj, especially in terms of the probability of correctly inferring efficacy in the primary endpoint. The

Table 3.3: Scenarios of mean matrix for simulation results in Table 3.4

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	L	M	H	L	M	H	L	M	H	L	M	H
P	.65	.65	.65	0	.65	.65	0	0	.65	0	.65	.65
S	.65	.65	.65	0	.65	.65	0	0	.65	.65	0	.65

NOTE: P and S stand for primary and secondary endpoint, while L, M, H indicate low, medium, and high doses. The mean of the placebo group is 0 for both endpoints in all scenarios.

more effective doses, the more power DPjd gains by adjusting for multiplicity among noneffective doses only.

3.7 Analysis of motivating example

We analyze the motivating example to illustrate the decision path principle. With two endpoints and three doses, there are $((2 + 1)^3 - 1 =)$ 26 partition hypotheses (compared to $(2^3)^2 - 1 = 63$ closed testing hypotheses) to be tested.

Critical values are set following the rules described in Section 3.5 with FWER controlled strongly at 0.05. Note that, due to unequal sample sizes across the doses in this example, critical values for some combinations of doses (e.g., (Low, High) and (Medium, High)) are different, as illustrated in Table 3.5.

Table 3.6 presents and compares critical values computed by independence or disjointness approximation.

Table 3.4: Probability of correctly rejecting the individual hypotheses from three procedures: DPjd proposed in this chapter, GKjj in Dmitrienko et al. (2006) and GKjd in Xu et al. (2008) in four scenarios described in Table 3.3 with correlation $\rho = 0.3, 0.5, 0.7$ and 0.9

		correlation $\rho = 0.3$											
		Scenario 1			Scenario 2			Scenario 3			Scenario 4		
		L	M	H	L	M	H	L	M	H	L	M	H
DPjd	P	.854	.854	.855	.832	.834		.808			.829	.804	
	S	.751	.747	.752	.701	.704		.658			.011		.667
GKjd	P	.807	.806	.812	.810	.814		.808			.811	.803	
	S	.710	.706	.714	.690	.694		.658			.008		.667
GKjj	P	.807	.806	.812	.810	.814		.808			.811	.803	
	S	.711	.706	.714	.691	.696		.661			.008		.670
		correlation $\rho = 0.5$											
DPjd	P	.862	.859	.862	.827	.824		.812			.828	.814	
	S	.778	.770	.778	.715	.717		.678			.014		.699
GKjd	P	.816	.806	.820	.807	.805		.812			.807	.814	
	S	.740	.728	.744	.705	.707		.678			.011		.699
GKjj	P	.816	.806	.820	.807	.805		.812			.807	.814	
	S	.741	.729	.744	.707	.709		.685			.011		.703
		correlation $\rho = 0.7$											
DPjd	P	.853	.850	.854	.832	.826		.810			.826	.810	
	S	.788	.784	.791	.742	.735		.706			.014		.714
GKjd	P	.803	.804	.809	.810	.803		.810			.806	.810	
	S	.749	.748	.756	.732	.725		.706			.009		.714
GKjj	P	.803	.804	.809	.810	.803		.810			.806	.810	
	S	.750	.750	.757	.735	.729		.714			.009		.720
		correlation $\rho = 0.9$											
DPjd	P	.862	.861	.857	.830	.838		.812			.834	.813	
	S	.821	.822	.819	.777	.783		.749			.014		.757
GKjd	P	.808	.813	.808	.808	.818		.812			.815	.813	
	S	.784	.786	.783	.767	.777		.749			.009		.757
GKjj	P	.808	.813	.808	.808	.818		.812			.815	.813	
	S	.785	.788	.784	.771	.781		.762			.009		.765

NOTE: The results are based on 10,000 simulations with FWER controlled strongly at 0.025. The probabilities for all three procedures in each scenario are evaluated using the same simulated data set to reduce variation on the difference of the probabilities between procedures. The blank entries represent noneffective dose-endpoint combinations.

Table 3.5: Partition hypotheses for hypertension example with rejection rules

Partition hypothesis	Rejection rule
$\theta_1^P \leq \delta^P, \theta_2^P \leq \delta^P, \theta_3^P \leq \delta^P$	$t_1^P > c_1$ or $t_2^P > c_1$ or $t_3^P > c_1$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_1^S > c_1$ or $t_2^S > c_1$ or $t_3^S > c_1$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_1^P > c_1$ or $t_2^S > c_{22}$ or $t_3^S > c_{22}$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_2^P > c_1$ or $t_1^S > c_{21}$ or $t_3^S > c_{21}$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P \leq \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_3^P > c_1$ or $t_1^S > c_{21}$ or $t_2^S > c_{21}$
$\theta_1^P \leq \delta^P, \theta_2^P \leq \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_1^P > c_1$ or $t_2^P > c_1$ or $t_3^S > c_{31}$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_3^P \leq \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_1^P > c_1$ or $t_3^P > c_1$ or $t_2^S > c_{31}$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_3^P \leq \delta^P, \theta_1^S \leq \delta^S, \theta_2^S \leq \delta^S, \theta_3^S \leq \delta^S$	$t_2^P > c_1$ or $t_3^P > c_1$ or $t_1^S > c_{32}$
$\theta_1^P \leq \delta^P, \theta_2^P \leq \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^P > c_{41}$ or $t_2^P > c_{41}$ or $t_3^S > c_{51}$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_3^P \leq \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^P > c_{41}$ or $t_3^P > c_{41}$ or $t_2^S > c_{51}$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_3^P \leq \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_2^P > c_{42}$ or $t_3^P > c_{42}$ or $t_1^S > c_{51}$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^P > c_{41}$ or $t_2^S > c_{51}$ or $t_3^S > c_{51}$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^P > c_{41}$ or $t_2^S > c_{51}$ or $t_3^S > c_{51}$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_2^P > c_{41}$ or $t_1^S > c_{51}$ or $t_3^S > c_{52}$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_2^P > c_{42}$ or $t_1^S > c_{51}$ or $t_3^S > c_{52}$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P \leq \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_3^P > c_{41}$ or $t_1^S > c_{51}$ or $t_2^S > c_{52}$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P \leq \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_3^P > c_{42}$ or $t_1^S > c_{51}$ or $t_2^S > c_{52}$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^S > c_{41}$ or $t_2^S > c_{41}$ or $t_3^S > c_{41}$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^S > c_{41}$ or $t_2^S > c_{41}$ or $t_3^S > c_{41}$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_2^S > c_{42}$ or $t_3^S > c_{42}$ or $t_1^S > c_{41}$
$\theta_1^P \leq \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^P > c_6$ or $t_2^S > c_6$ or $t_3^S > c_6$
$\theta_1^P > \delta^P, \theta_2^P \leq \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_2^P > c_6$ or $t_1^S > c_6$ or $t_3^S > c_6$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P \leq \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_3^P > c_6$ or $t_1^S > c_6$ or $t_2^S > c_6$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S \leq \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_1^S > c_6$ or $t_2^S > c_6$ or $t_3^S > c_6$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_2^S > c_6$ or $t_3^S > c_6$ or $t_1^S > c_6$
$\theta_1^P > \delta^P, \theta_2^P > \delta^P, \theta_3^P > \delta^P, \theta_1^S > \delta^S, \theta_2^S > \delta^S, \theta_3^S > \delta^S$	$t_3^S > c_6$ or $t_1^S > c_6$ or $t_2^S > c_6$

NOTE: Due to unequal sample sizes, we use a second index to denote critical values c_{22} and c_{21} , for example, for (Medium, High) and (Low, High) in the third and fourth partition hypotheses from the top.

Table 3.6: Critical values in Table 3.5 using independence or disjointness approximation with strong FWER control at 0.05

Approximation			c_1	c_{41}	c_{42}	c_6
Independence or Disjointness			2.099	1.946	1.945	1.661
	c_{21}	c_{22}	c_{31}	c_{32}	c_{51}	c_{52}
Independence	2.154	2.154	2.206	2.204	2.015	2.016
Disjointness	2.163	2.162	2.221	2.219	2.027	2.028

NOTE: Critical values involving the multivariate t are calculated using the *qmv*t function in R.

From the summary statistics in Table 3.1 and using $\delta^P = \delta^S = 0$, the two sample t statistics for each dose and endpoint are:

$$t_1^P = -2.4202, t_2^P = -3.5861, t_3^P = -2.6232$$

$$t_1^S = -1.7434, t_2^S = -3.2498, t_3^S = -2.2198$$

As a lower response is better for drugs treating high blood pressure, we take the negatives of the t statistics in testing the hypotheses to be consistent with the sign of the rejection region in Table 3.5.

An individual hypothesis is rejected if all the partition hypotheses for which it could be true are rejected. For example, $H_{01}^S : \theta_1^S \leq \delta^S$ is rejected if partition hypotheses 1-10, 12-14, 16, 18, 19, 21, 24 from top in Table 3.5 are all rejected. After collating results for each individual hypothesis, we claim all doses are efficacious for all endpoints using either the independence or the disjointness approximation.

3.8 Discussion

Since approval of test drug is based about inference on the primary endpoint only, it is reasonable to require that inference on secondary endpoints be given only if efficacy is proven in the primary endpoint.

One can ask whether there should be the additional requirement that inference on the primary endpoint be totally unaffected by the presence of secondary endpoints. The presence of secondary endpoints in addition to the primary endpoint affects decision-making in three ways.

Decision path: At a given dose, inference on secondary endpoints should only be given if efficacy is shown in the primary endpoint. This can be achieved by appropriate formulation of hypotheses, as we have shown.

Multiplicity adjustment: Inclusion of secondary endpoints in the analysis may necessitate multiplicity adjustment in inference on the primary endpoint. For example, in Table 3.2, one rejects the hypothesis H_{01}^P if

$$\{\theta_1^P \leq \delta^P \text{ and } \theta_2^P \leq \delta^P\}, \quad (3.8.1)$$

$$\{\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S \leq \delta^S\}, \quad (3.8.2)$$

$$\text{and } \{\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S > \delta^S\} \quad (3.8.3)$$

are all rejected. Thus, inference on low dose of the primary endpoint (θ_1^P) may need a multiplicity adjustment, to account for the possibility that the high dose of the secondary endpoint lacks efficacy ($\theta_2^S \leq \delta^S$ in (3.8.2)). From our perspective, this is reasonable, because there should be a cost associated with incorrectly inferring efficacy in a secondary endpoint by chance alone.

Note, however, that this cost of using a multiplicity adjustment is removed if efficacy in the secondary endpoint at the high dose is shown (i.e., if partition hypothesis (3.8.2) is rejected). Our approach has the feature that, if efficacy at a dose has been established for all endpoints, then that dose is no longer included in the multiplicity adjustment, a familiar feature of step-wise testing in the case of a single endpoint.

Appearance: The method we propose rejects the individual hypothesis H_{01}^P if $\{t_1^P > c_1\}$ or $\{c_3 < t_1^P \leq c_1 \text{ and } t_2^P > c_1 \text{ and } t_2^S > c_2\}$. That is, inference about the efficacy at the low dose of the primary endpoint can be made using the smaller critical value c_3 when there is evidence of efficacy at the high dose of the secondary endpoint. It thus appears that inference on the primary endpoint may depend on observations on a secondary endpoint (at a different dose and thus on a different group of subjects).

One can give the appearance that such dependence does not occur by artificially increasing the critical value c_3 to c_1 , so that efficacy of the low dose of the primary endpoint or indeed any dose of the primary endpoint, is inferred when $\{t_i^P > c_1\}$, adjusting for the multiplicity of all doses, regardless of observations on secondary endpoints. This is the approach taken in Dmitrienko et al. (2006) and Xu et al. (2008), to ensure what they call “inference made in primary endpoints is not affected by the inference made in secondary endpoints”. Such appearance comes at a loss of power. It is, in fact, the reason that our Decision Path approach is more powerful as shown in Table 3.4. We feel there is no need for such loss of power, for the sake of appearance.

3.9 Appendix

3.9.1 Appendix A: Proof of Theorem 3.4.1

Proof of Theorem 3.4.1: Let $\Theta_{ij} = \{\theta = (\theta_{11}, \dots, \theta_{km}) \in \Omega : \theta_{ij} \leq \delta_{ij}\}$ denote the null space of the individual hypothesis H_{0ij} , $i = 1, \dots, k$, $j = 1, \dots, m$, and the superscript c denote the complement of a set.

Path partitioning, within a fixed dose i , partitions the parameter space into $m + 1$ subspaces $\tilde{\Theta}_{ij}$ where

$$\tilde{\Theta}_{ij} = \begin{cases} \Theta_{ij}, & j = 1 \\ \Theta_{ij} \cap (\bigcup_{s=1}^{j-1} \Theta_{is})^c, & j = 2, \dots, m \\ (\bigcup_{s=1}^{j-1} \Theta_{is})^c, & j = m + 1 \end{cases} \quad (3.9.1)$$

Note that the subspaces $\tilde{\Theta}_{ij}$ are disjoint within each dose i but not across doses.

Consider a mapping which, for each dose i , assigns a unique “endpoint” j_i , $j_i \in \{1, \dots, m + 1\}$. For each such mapping, form a partition hypothesis by intersecting $\tilde{\Theta}_{ij_i}$ across all doses (i.e., intersect those connected by “edges” in Figure 3.4). The $(m + 1)^k$ possible intersections

$$\mathcal{P} = \left\{ \theta \in \bigcap_{i=1}^k \tilde{\Theta}_{ij_i}, j_i \in \{1, \dots, m + 1\} \right\} \quad (3.9.2)$$

partition the parameter space. The number of partition hypotheses to be tested is therefore $(m + 1)^k - 1$, since the hypothesis $\theta \in \bigcap_{i=1}^k \tilde{\Theta}_{i(m+1)}$, indicating that all doses are efficacious for all endpoints, does not need to be tested.

The next step is to show that any procedure derived by testing this set of partition hypotheses has properties (i) and (ii). To show (i), note that the individual hypothesis H_{0ij} is rejected if all partition hypotheses indicating non-efficacious endpoints with no lower order than endpoint j for dose i are rejected. Mathematically all partition

hypotheses in

$$\mathcal{R}_{ij} = \left\{ \theta \in \bigcap_{l=1}^k \tilde{\Theta}_{lj_l}, j_l \in \{1, \dots, m+1\}, j_i \leq j \right\}$$

need to be rejected where j_l is a dummy index for the endpoint assigned to dose $l, l = 1, \dots, k$. Since $\mathcal{R}_{ij} \subset \mathcal{R}_{ij'}$ if $j < j'$, $H_{0ij'}$ can not be rejected if H_{0ij} is not rejected first. Therefore, any procedure developed through testing the partition hypotheses in $\mathcal{P}$ satisfies the ordering between endpoints. To show (ii), note that we already know the partition hypotheses are disjoint; only one of them could be true. Testing each at α level suffices the strong FWER control at the same level.

3.9.2 Appendix B: Steps in computing critical values

For each partition hypothesis testing $\theta = (\theta_{11}, \dots, \theta_{km}) \in \Theta_p$, let I_j be the index set of doses such that $t_{ij}, i \in I_j$ is selected for endpoint j , (if no t statistic from endpoint j is selected, $I_j = \emptyset$) $j = 1, \dots, m$, and $I = \cup_{j=1}^m I_j$ be the index set of doses with t statistics selected for a certain endpoint. For strong FWER control at level α , we calculate the starting candidate critical value $d_{\alpha, I}$ using the exact joint distribution of the t statistics for endpoint 1 (primary endpoint) with doses in I such that

$$\sup_{\cap_{i \in I} \{\theta_{i1} \leq \delta_1\}} \mathbf{P}_{\theta}(\max_{i \in I} T_{i1} > d_{\alpha, I}) = \alpha.$$

Note, $d_{\alpha, I}$ is calculated adjusting multiplicity for the doses in I (a subset of all doses) only. We then assign critical values as close to $d_{\alpha, I}$ as possible using independence or disjointness approximation. Specifically, the critical values c_j for t statistics selected for each endpoint $j, j = 1, \dots, m$ can be set recursively using the following steps:

- Step s , (representing each endpoint), $s = 1, \dots, m$:

- If no t statistic from endpoint s is selected, that is if $I_s = \emptyset$, go to step $s+1$.
- Else pretend to use $d_{\alpha,I}$ as the critical value for this endpoint, and check if the cumulative probability (endpoint 1 to s) of rejection is no larger than α . That is, if

$$\begin{aligned}
 * \text{ Independence: } & \prod_{\substack{j=1 \\ I_j \neq \emptyset}}^s \inf_{\boldsymbol{\theta} \in \Theta_p} P_{\boldsymbol{\theta}}(\max_{i \in I_j} T_{ij} \leq d_{\alpha,I}) \geq 1 - \alpha \\
 * \text{ Disjointness: if } & \sum_{\substack{j=1 \\ I_j \neq \emptyset}}^s \sup_{\boldsymbol{\theta} \in \Theta_p} P_{\boldsymbol{\theta}}(\max_{i \in I_j} T_{ij} > d_{\alpha,I}) \leq \alpha
 \end{aligned}$$

Note the two-sample t statistic, T_{ij} , depends on θ_{ij} only.

If true, set $c_s = d_{\alpha,I}$ and go to step $s + 1$.

Else set c_s to be somewhat larger than $d_{\alpha,I}$ to ensure the cumulative probability of rejection is within α . That is set c_s such that

* Independence:

$$\left(\prod_{\substack{j=1 \\ I_j \neq \emptyset}}^{s-1} \inf_{\boldsymbol{\theta} \in \Theta_p} P_{\boldsymbol{\theta}}(\max_{i \in I_j} T_{ij} \leq d_{\alpha,I}) \right) \left(\inf_{\boldsymbol{\theta} \in \Theta_p} P_{\boldsymbol{\theta}}(\max_{i \in I_s} T_{is} \leq c_s) \right) = 1 - \alpha$$

* Disjointness:

$$\sum_{\substack{j=1 \\ I_j \neq \emptyset}}^{s-1} \sup_{\boldsymbol{\theta} \in \Theta_p} P_{\boldsymbol{\theta}}(\max_{i \in I_j} T_{ij} > d_{\alpha,I}) + \sup_{\boldsymbol{\theta} \in \Theta_p} P_{\boldsymbol{\theta}}(\max_{i \in I_s} T_{is} > c_s) = \alpha$$

Then set the critical values for endpoints of lower order with at least one t statistic selected to infinity. That is,

set $c_l = \infty$, for $l = s + 1, \dots, m$ & $I_l \neq \emptyset$, and quit.

CHAPTER 4

WORK IN PROGRESS AND FUTURE RESEARCH

4.1 Stability of V

When simultaneously testing a large number of hypotheses, such as in microarray data analysis, FWER which guarantees that the probability of any false rejection is no more than some small percentage seems to be too stringent. Thus researchers developed procedures that control other types of error rates. Among those, $E(V)$ control is proposed with the hope that the number of false rejections for the experiment/analysis is “not too far away” from the pre-specified value. In statistical language, this is equivalent to hoping that this particular realization of the random variable V is close to its expectation. There are a lot of factors that could influence the result, however. The variance of V is obviously one important factor.

Gordon et al. (2007) advocate the use of what they call the “Bonferroni procedure”, the equivalent one-step procedure for $E(V)$ control introduced in Section 1.4.1. They showed through simulation that controlling $E(V)$ or FDR at about the same level, the one-step procedure is more stable than the BH step-up procedure (Benjamini and Hochberg, 1995). Owen (2005) also investigates the effect of correlation among test statistics on the variance of V .

In this section, we provide the insight behind the simulations showing why the step-up method is less stable than the one-step procedure.

Since the number of false rejections only relates to the true null hypotheses, we assume all k null hypotheses are true in the discussion. Thus, for notational convenience, we drop the notion of taking the supremum over Θ_I^* and assume the test statistics $T_1, \dots, T_k$ have a multivariate normal distribution with mean vector $\mathbf{0}$ under the null hypothesis with covariance matrix $(1 - \rho) * I_k + \rho * J_k$, where I_k is the k by k identity matrix and J is the k by k matrix with all entries being 1. This means that each T_i has standard deviation 1, and each pair of T_i and T_j has equal correlation ρ .

4.1.1 Var(V) increases with correlation for the one-step procedure

The one-step procedure compares each test statistic T_i with the critical value c , and rejects H_{0i} if $T_i > c$, $i = 1, \dots, k$. Thus the number of false rejections can be written as $V = \sum_{i=1}^k I(T_i > c)$. The expected value and variance of the number of false rejections can be calculated as follows:

$$\begin{aligned}
E(V) &= E\left(\sum_{i=1}^k I(T_i > c)\right) \\
&= \sum_{i=1}^k P(T_i > c) \\
&= kp \quad \text{where } p = 1 - \Phi(c)
\end{aligned}$$

$$\begin{aligned}
Var(V) &= Var\left(\sum_{i=1}^k I(T_i > c)\right) \\
&= \sum_{i=1}^k Var(I(T_i > c)) + 2 \sum_{\substack{i \neq j \\ i < j}} Cov(I(T_i > c), I(T_j > c)) \\
&= kp(1 - p) + k(k - 1)Cov(I(T_1 > c), I(T_2 > c))
\end{aligned}$$

$$\begin{aligned}
\text{where} \quad & Cov(I(T_1 > c), I(T_2 > c)) \\
&= E(I(T_1 > c)I(T_2 > c)) - E(I(T_1 > c))E(I(T_2 > c)) \\
&= P(T_1 > c \text{ and } T_2 > c) - p^2
\end{aligned}$$

From Slepian's inequality (Corollary A.3.1 on p. 229 of Hsu, 1996), we know that the probability $P(T_1 > c \text{ and } T_2 > c)$ increases with the correlation ρ . Therefore, the variance of the number of false rejections, $Var(V)$ increases in correlation ρ as well.

In analyzing real data, the test statistics are always dependent. If the correlation among test statistics is high, then even if one controls the expected value of V , due to increased variance, the number of false rejections may be far off from what we are expecting.

4.1.2 Comparing the one-step and BH step-up approaches

For illustration purposes we discuss the $k = 2$ case in this section. We also assume that the test statistics are independent, since the BH procedure is only valid under this assumption.

Let $[1], [2]$ be the random indices such that $T_{[1]} \leq T_{[2]}$. Then the BH procedure executes in the step-up fashion shown below controlling FDR at level β .

- Step 1: If $T_{[1]} > z_\beta$ (normal β quantile) then reject both hypotheses, that is $V = 2$ and stop; otherwise, go to next step.
- Step 2: If $T_{[2]} > z_{\beta/2}$ then reject $H_{0[1]}$, that is $V = 1$ and stop.

The distribution of V_{BH} can then be summarized as follows:

$$V_{BH} = \begin{cases} 2 & \text{if } T_{[1]} > z_\beta \\ 1 & \text{if } T_{[1]} \leq z_\beta \text{ and } T_{[2]} > z_{\frac{\beta}{2}} \\ 0 & \text{otherwise} \end{cases}$$

The expected value of V_{BH} is:

$$\begin{aligned} E(V_{BH}) &= 2P(T_{[1]} > z_\beta) + P(T_{[1]} \leq z_\beta \text{ and } T_{[2]} > z_{\frac{\beta}{2}}) \\ &= 2P(T_1 > z_\beta \text{ and } T_2 > z_\beta) + 2P(T_1 \leq z_\beta \text{ and } T_2 > z_{\frac{\beta}{2}}) \\ &= 2\beta^2 + 2\frac{\beta}{2}(1 - \beta) \\ &= \beta^2 + \beta \end{aligned}$$

The variance of V_{BH} is:

$$\begin{aligned}
& Var(V_{BH}) \\
&= E(V^2) - E(V)^2 \\
&= 2^2 P(T_{[1]} > z_\beta) + P\left(T_{[1]} \leq z_\beta \text{ and } T_{[2]} > z_{\frac{\beta}{2}}\right) - (\beta^2 + \beta)^2 \\
&= 3\beta^2 + \beta - (\beta^4 + 2\beta^3 + \beta^2) \\
&= \beta(-\beta^3 - 2\beta^2 + 2\beta + 1)
\end{aligned}$$

For the one-step procedure, suppose $E(V_{OS})$ is controlled at level α and the critical value for the one-step procedure is $z_{\alpha/2}$. Under independence, $V = \sum_{i=1}^2 I(T_i > z_{\alpha/2})$ follows a Binomial distribution with $p = \alpha/2$. The expectation of V_{OS} is then $2p = \alpha$. For both procedures to control $E(V)$ at the same level, we require the following equality: $\beta^2 + \beta = \alpha$.

The variance of V_{OS} for the one-step procedure is thus

$$\begin{aligned}
Var(V_{OS}) &= 2(\alpha/2)(1 - \alpha/2) = \alpha(1 - \alpha/2) = \beta(1 + \beta)(1 - \beta^2 - \beta) \\
&< \beta(-\beta^3 - 2\beta^2 + 2\beta + 1) = Var(V_{BH}) \text{ when } \beta \in [0, 1]
\end{aligned}$$

It is clear from the calculation shown above that the one-step procedure yields smaller variance of V than the BH step-up procedure while controlling $E(V)$ at the same level. However, to provide the insight as to why that is the case we need to investigate the distribution of V for these two procedures.

Table 4.1 shows that the one-step method has smaller probability of V taking on values 0 and 2 compared to the BH method while the means of the two distributions are exactly the same. The reason, shown in Figure 4.1, is that the region $\{V = 2\}$ for the one-step procedure is a subset of the one for the BH procedure, and it is the

Table 4.1: Distribution of the number of false rejections for the one-step and BH step-up procedure

	One-step	Relationship	BH
$P(V=0)$	$(1 - \alpha/2)^2 = 1 - \beta + \beta^2(1 + \beta)^2/4 - \beta^2$	$<$	$1 - \beta$
$P(V=1)$	$\alpha(1 - \alpha/2) = \beta(1 + \beta)(1 - \frac{\beta^2 + \beta}{2})$	$>$	$\beta(1 - \beta)$
$P(V=2)$	$\alpha^2/4 = \beta^2(1 + \beta)^2/4$	$<$	β^2

Figure 4.1: Rejection region of the one-step procedure (left) and the BH step-up procedure (right).

opposite with $\{V = 1\}$. Even with the same “center”, when probabilities on the two extreme ends increase (see Figure 4.2), the variance obviously increases.

Although Figure 4.2 is in reference to $E(V)$ control procedures, a similar situation happens in the case of FWER control. That is, for example, the step-down Holm’s or step-up Hochberg’s procedure are more powerful than the one-step procedure because the rejection region of $\{V = 2\}$ is larger for step-wise procedures. However, since the

Figure 4.2: Probability distribution of the number of false rejections for the one-step procedure (left) and the BH step-up procedure (right).

FWER control procedure only concerns the region $\{V > 0\} = \{V = 1\} \cup \{V = 2\}$, one can increase $\{V = 2\}$ and decrease $\{V = 1\}$ while still maintaining the same $\{V > 0\}$. Unfortunately, this does not happen with $E(V)$ control since every piece of the probability may change the “average”. This is the exact reason why the one-step procedure is not necessarily less powerful than step-wise procedures for $E(V)$ control as mentioned in Section 1.4.

4.2 Non-inferiority and superiority tests

In dose response studies, when an active control (a drug that is already proven effective in the market) is included in the clinical trial, efficacy of the new test drug is then shown if it is non-inferior to the active control. However, sometimes given data, the test drug may seem even “better” than the active control. But without proper statistical design beforehand, the test drug can not be claimed to be superior

to the active control even if it is shown in the data. Driven by the potential advertisement benefit, there is an increasing desire from pharmaceutical companies for testing superiority if non-inferiority has been shown.

Wiens and Dmitrienko (2005) proposed a multiple testing method called the “fallback” procedure. It operates in a fixed sequential manner, but allows one to continue each test even if one test can not be rejected early on. Thus it is quite commonly used in, for example, non-inferiority and superiority problems. The downside, however, is that no test in each step can be tested at the full α level as predetermined steps allow. The “fallback” procedure, in fact, is also a special case of partitioning. We can show that it is uniformly less powerful than the sample-determined steps, such as Holm’s or step-down tests based on independence.

In this section, we use the partitioning approach to develop non-inferiority and superiority tests that are more powerful than the “fallback” procedure. In addition, we provide insight into the non-inferiority and superiority problem and show that the proposed test works even with multiple endpoints with ordering between primary and secondary endpoints. We show that the ordering from non-inferiority to superiority is a path and the problem can be solved using the decision paths approach discussed in Chapter 3. Thus, for all scenarios investigated in this section, we only focus on the process of getting the set of partition hypotheses and choosing rejection rules. The rest of the steps are exactly the same as described in Section 3.5.

4.2.1 Paths between non-inferiority and superiority

Suppose there are k doses included in the clinical trial, and let $\theta_i = \mu_i - \mu_a$ be the mean difference between dose group i and the active control group. Then the

hypotheses of interest are:

$$\begin{aligned} H_{0i}^{NI} : \theta_i &\leq \delta_{NI} \\ H_{0i}^{SUP} : \theta_i &\leq \delta_{SUP} \end{aligned} \quad i = 1, \dots, k \quad (4.2.1)$$

where δ_{NI} is the clinically meaningful non-inferiority margin, and δ_{SUP} is the clinically meaningful superiority margin. Usually, the non-inferiority margin, δ_{NI} , is a negative number and the superiority margin, δ_{SUP} , is a positive number if a larger response is better. One of the most problematic issues for non-inferiority test is the determination of the non-inferiority margin δ_{NI} . (See ICH E10 for further discussion.)

We first discuss the case in which, for each dose, if the test drug is claimed to be non-inferior to the active control, we further test to see if the test drug is superior to the active control. This setting is a decision paths problem as discussed in Chapter 3 with paths from the non-inferiority test to the superiority test for each dose, but no paths between the doses, as shown in Figure 4.3.

Figure 4.3: Decision paths for k doses, with paths from non-inferiority to superiority for each dose.

It is clear from Figure 4.3 that the partition test described in Chapter 3 is directly applicable here, with NI (non-inferiority) as the primary endpoint, and SUP

(superiority) as the secondary endpoint. The difference, however, is that the parameters involved in the NI and SUP test are the same in this case. Thus we believe a more powerful test can be derived if this information is incorporated into the decision making process.

4.2.2 Additional paths between doses for the non-inferiority test

When there is prior information available that indicates a non-decreasing dose-response relationship, it is usually more powerful to order the tests for doses. In this section, we discuss the scenario when we order the NI test from high dose to low dose. The partition principle is applied and then illustrated for the case of $k = 3$.

Figure 4.4: Decision paths for the non-inferiority (NI) and superiority (SUP) test with three doses. Paths are from high to low for the NI test, and from NI to SUP for each dose.

Figure 4.4 presents the paths between the doses for the NI test as well as the paths between NI and SUP for each dose in the case of three doses. Based on these decision paths, one can use partitioning to form hypotheses that are compatible with

the pre-specified paths. Once we have the set of partition hypotheses available, the rest of the steps fall into the description of Section 3.5. Thus, in this section, we focus on the steps of partitioning the parameter space and getting the set of partition hypotheses.

We follow the two-stage partitioning plan mentioned in Chapter 3 with the more complicated paths in Figure 4.4. The first stage is to partition following the paths. Based on Figure 4.4, we first test the high dose NI, $H_{03}^{NI} : \theta_3 \leq \delta_{NI}$. If H_{03}^{NI} is rejected, then the high dose is judged to be non-inferior to the active control for the test drug.

Then follow the path to either the medium dose NI test or the high dose SUP test knowing that the high dose NI test is already rejected. Thus we test $H_{02}^{NI*} : \theta_2 \leq \delta_{NI}$ and $\theta_3 > \delta_{NI}$ and $H_{03}^{SUP*} : \theta_3 \leq \delta_{SUP}$ and $\theta_3 > \delta_{NI}$.

We then continue with similar steps testing the rest of the hypotheses following the paths with the knowledge that the hypotheses in the previous steps have already been rejected.

We present the process of a stage I path partition in a decision tree representation in Figure 4.5. One starts with the high dose NI test, presented in the root (top) of the tree, and then form hypotheses by following the paths in Figure 4.4 with combined information that hypotheses in the path upstream have already been rejected. As a result of this stage I path partition, hypotheses along the same “branch” of the decision tree are disjoint, meaning they can not be true at the same time.

The second stage is a disjointness partition, as described in Chapter 3. We take intersections between the branches of the decision tree in Figure 4.4 to make all hypotheses disjoint. We start from the two lowest branches on the bottom of the tree with the same node (M, NI), and take intersections, resulting in $2*3=6$ hypotheses.

Figure 4.5: Decision tree from stage I: path partition for the decision paths problem in Figure 4.4.

Then these 6 hypotheses, together with the hypothesis representing the node, can be regarded as one big branch, and they are intersected with the left branch. This gives us $2 \times 7 = 14$ hypotheses. Adding the top hypothesis (H, NI) and subtracting the one hypothesis that does not need to be tested (representing all doses are superior to the active control), there are $14 + 1 - 1 = 14$ partition hypotheses presented in Table 4.2.

Table 4.2: Partition hypotheses from a stage II disjointness partition for decision paths in Figure 4.4

Partition hypotheses				Rejection rule			
				$\theta_3 \in (-\infty, \delta_{NI}]$			
$\theta_1 \in (-\infty, \delta_{NI}]$	and	$\theta_2 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$	$t_1^{NI} > c_3$	or	$t_2^{SUP} > c_3$ or $t_3^{NI} > c_1$
$\theta_1 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_2 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$	$t_1^{SUP} > c_3$	or	$t_2^{SUP} > c_3$ or $t_3^{SUP} > c_3$
$\theta_1 \in (\delta_{SUP}, +\infty)$	and	$\theta_2 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$			$t_2^{SUP} > c_2$ or $t_3^{SUP} > c_2$
$\theta_1 \in (-\infty, \delta_{NI}]$	and	$\theta_2 \in (\delta_{SUP}, +\infty)$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$	$t_1^{NI} > c_2$		or $t_3^{SUP} > c_2$
$\theta_1 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_2 \in (\delta_{SUP}, +\infty)$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$	$t_1^{SUP} > c_2$		or $t_3^{SUP} > c_2$
$\theta_1 \in (\delta_{SUP}, +\infty)$	and	$\theta_2 \in (\delta_{SUP}, +\infty)$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$			$t_3^{SUP} > c_1$
				$\theta_2 \in (-\infty, \delta_{NI}]$	and	$\theta_3 \in (\delta_{NI}, \delta_{SUP}]$	$t_2^{NI} > c_2$ or $t_3^{SUP} > c_2$
$\theta_1 \in (-\infty, \delta_{NI}]$	and	$\theta_2 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_3 \in (\delta_{SUP}, +\infty)$	$t_1^{NI} > c_2$	or	$t_2^{SUP} > c_2$
$\theta_1 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_2 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_3 \in (\delta_{SUP}, +\infty)$	$t_1^{SUP} > c_2$	or	$t_2^{SUP} > c_2$
$\theta_1 \in (\delta_{SUP}, +\infty)$	and	$\theta_2 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_3 \in (\delta_{SUP}, +\infty)$			$t_2^{SUP} > c_1$
$\theta_1 \in (-\infty, \delta_{NI}]$	and	$\theta_2 \in (\delta_{SUP}, +\infty)$	and	$\theta_3 \in (\delta_{SUP}, +\infty)$	$t_1^{NI} > c_1$		
$\theta_1 \in (\delta_{NI}, \delta_{SUP}]$	and	$\theta_2 \in (\delta_{SUP}, +\infty)$	and	$\theta_3 \in (\delta_{SUP}, +\infty)$	$t_1^{SUP} > c_1$		
				$\theta_2 \in (-\infty, \delta_{NI}]$	and	$\theta_3 \in (\delta_{SUP}, +\infty)$	$t_2^{NI} > c_1$

Note: The statistics t_i^{NI}, t_i^{SUP} are two-sample t statistics for hypotheses $H_{0i}^{NI}, H_{0i}^{SUP}, i = 1, 2, 3$. The critical values have the ordering: $c_1 < c_2 < c_3$.

Let n_i and n_a be the sample sizes for dose group i and the active control group, $\bar{Y}_i$, $\bar{Y}_a$ be the sample means for dose group i and the active control group, and $\hat{\sigma}$ be the sample standard deviation. Then the two-sample t-statistics defined as

$$t_i^{NI} = \frac{\bar{Y}_i - \bar{Y}_a - \delta_{NI}}{\hat{\sigma} \sqrt{1/n_i + 1/n_a}} \quad t_i^{SUP} = \frac{\bar{Y}_i - \bar{Y}_a - \delta_{SUP}}{\hat{\sigma} \sqrt{1/n_i + 1/n_a}} \quad i = 1, \dots, 3$$

are used in the rejection rule for each partition hypothesis in Table 4.2. They are chosen according to the recommendation described in Chapter 3. The difference, however, is that unlike the primary and secondary endpoints case, the NI and SUP tests involve the same parameters and almost the same test statistics, except for the different constant in the numerator. Thus the correlation between $(T_1^{SUP}, T_2^{SUP}, T_3^{SUP})$ is the same as the one between $(T_1^{NI}, T_2^{SUP}, T_3^{SUP})$, for example. Since all partition hypotheses are tested at the same α level, the critical values for these two sets of test statistics are then the same as shown in Table 4.2. By calculating the critical values based on the exact correlation between test statistics, this procedure clearly is more powerful than the primary secondary endpoint situation where a slightly conservative measure (independence or disjointness) has to be used.

4.2.3 Non-inferiority and superiority tests with two doses and two endpoints

In this section, we discuss non-inferiority and superiority tests in dose-response studies with multiple endpoints. To keep the problem setup simple, we assume there are two doses, low and high, and two endpoints, primary and secondary.

The hypotheses of interest are the following eight, testing whether each dose-endpoint combination is inferior or non-superior to the active control.

Primary	$H_{0i}^{P,NI} : \theta_i^P \leq \delta_{NI}^P$	$H_{0i}^{P,SUP} : \theta_i^P \leq \delta_{SUP}^P$	$i = 1, 2.$
Secondary	$H_{0i}^{S,NI} : \theta_i^S \leq \delta_{NI}^S$	$H_{0i}^{S,SUP} : \theta_i^S \leq \delta_{SUP}^S$	

In testing these eight hypotheses, we assume the following three sets of orderings.

- O1. For each dose and endpoint combination, superiority is tested only if non-inferiority is shown.
- O2. For each dose, non-inferiority for the secondary endpoint is tested only if non-inferiority for the primary endpoint is shown.
- O3. For the primary endpoint, non-inferiority for the low dose is tested only if non-inferiority for the high dose is shown.

The rationale for O2 is that, given the fact that the test drug is shown non-inferior to the active control for the primary endpoint, further testing either superiority for the primary endpoint or non-inferiority for the secondary endpoint is reasonable. However, requiring such an ordering between the superiority of primary and secondary endpoint would not make much sense. Even if superiority can not be claimed for the primary endpoint, a test for the superiority of the secondary endpoint is still of interest given that non-inferiority of the secondary endpoint has already been shown. The orders in O3 are proposed in scenarios where prior information indicates that the primary endpoint is more efficacious with a higher dose.

These three layers of orders between the eight hypotheses are graphically represented in Figure 4.6.

Based on the decision paths in Figure 4.6, the stage I path partition starts with the NI test for the high dose primary endpoint. Then, following the paths, one can get the decision tree in Figure 4.7.

The advantage of using the decision tree representation is that one can easily calculate the number of the final set of partition hypotheses to be tested after the stage

Figure 4.6: Graphical representation of the decision paths (O1, O2, and O3) for non-inferiority and superiority tests with two doses (low, and high) and two endpoints (primary, and secondary).

II disjointness partition. In the case of Figure 4.7, this number is $1+2*3*(1+2*3)-1 = 42$. Compared to $2^8 - 1 = 255$ tests for closed testing, the number of tests is reduced dramatically. This set of 42 partition hypotheses obtained by taking all possible intersections, is not presented since the process is quite straight forward.

The two scenarios discussed in this section for non-inferiority and superiority tests are decision path problems. The general framework introduced in Chapter 3 should be followed. In this particular setting, more powerful tests are developed due to the known correlation between test statistics for non-inferiority and superiority tests. The rest of the steps in Section 3.5 are omitted since they are easy to follow.

4.3 Co-primary endpoints problem

Chapters 2 and 3 discussed multiple endpoints problems in the setting of a single primary endpoint and multiple secondary endpoints. However, sometimes, due to the disease mechanism, more than one primary endpoint needs to be assessed for the

Figure 4.7: Decision tree from stage I: path partition, for the decision paths problem in Figure 4.6

overall efficacy of the new treatment. *Co-primary endpoints* are such that efficacy of a new treatment in *all* multiple primary endpoints constitutes evidence of efficacy. For example, four co-primary endpoints are used in Migraine treatments: pain free

at 2 hours, nausea at 2 hours, photosensitivity at 2 hours, and phonosensitivity at 2 hours (Offen et al. 2007).

In multiple testing, this problem is regarded as a *reverse* multiplicity problem (Offen et al. 2007) because no multiplicity adjustment is needed. Each primary endpoint can be tested at α level while maintaining the FWER at the same level α .

The co-primary endpoints problem is in fact an *intersection-union* test, where the null hypothesis is a union of two or more individual null hypotheses. Common practice is to reject if all individual hypotheses are rejected. In other words, the rejection region is the intersection of all individual rejection regions. Thus the Type I error rate is controlled naturally and is not of great concern. This test, however, suffers from a lack of power, especially when all the individual null hypotheses are true or when the test statistics are independent.

Several alternative methods were proposed from both statistical and medical standpoints. Offen et al. (2007) advocates the identification of a single primary endpoint to avoid the co-primary endpoints situation from a medical perspective, by either defining a composite primary endpoint or demoting all but one endpoints to become key secondary endpoints. This approach involves important medical decisions, and may not be feasible in all cases. As one of the statistical solutions, Offen and Helterbrand (1986) propose the use of a restricted null space because they think some of the regions of the null space are unrealistic in practice. Offen et al. (2007) takes a Bayesian approach that places priors on the mean vector as well as on the covariance matrix of the endpoints. A mixed Bayesian/Frequentist approach was also proposed by Chuang-Stein et al. (2007) by assigning a discrete prior distribution on three points $(+\infty, 0)$, $(0, +\infty)$, and $(0, 0)$ in the two co-primary endpoints case. All

these methods try to modify the definition of the Type I error rate as taking the supremum over the entire parameter space as on p. 361 of Casella and Berger (2001). We do not feel that this is the right direction for constructing more powerful test procedures.

Thus, a question of fundamental importance is whether the loss of power in executing the usual intersection-union test can be avoided. Inspired by the idea of strong partitioning (Finner and Strassburger 2002), which utilizes information on partition hypotheses that have been rejected when testing the rest of the partition hypotheses, we propose a new partitioning test that is proven to control the Type I error rate with Casella and Berger (2001)’s definition. It is also shown to be more powerful than the usual intersection-union test when the parameter of the rejected partition hypotheses lie in certain region.

4.3.1 Problem setting

We investigate the co-primary endpoints problem in the setting of one primary and one secondary endpoint with one treatment group because a “win” on both is equivalent to a “win” on the first, and then a “win” on the second.

Let (μ_i^P, μ_i^S) be the means of the primary and secondary endpoints for group i where $i = 1$ is the treatment group and $i = 0$ is the placebo group. Then $(\theta^P, \theta^S) = (\mu_1^P - \mu_0^P, \mu_1^S - \mu_0^S)$ are the corresponding mean differences. Denote the clinically meaningful differences as (δ^P, δ^S) . Then we are interested in testing the following two hypotheses:

$$\begin{aligned} H_0^P : \theta^P &\leq \delta^P \\ H_0^S : \theta^S &\leq \delta^S \end{aligned} \tag{4.3.1}$$

In addition, since the secondary endpoint is of interest only if the primary endpoint has shown efficacy, we test H_0^S only if H_0^P is rejected first.

Assume the measurements (Y_{ir}^P, Y_{ir}^S) for patient r in group i come from a model:

$$\begin{pmatrix} Y_{ir}^P \\ Y_{ir}^S \end{pmatrix} = \begin{pmatrix} \mu_i^P \\ \mu_i^S \end{pmatrix} + \begin{pmatrix} \epsilon_{ir}^P \\ \epsilon_{ir}^S \end{pmatrix} \quad i = 0, 1; \quad r = 1, \dots, n_i. \quad (4.3.2)$$

where $\begin{pmatrix} \epsilon_{ir}^P \\ \epsilon_{ir}^S \end{pmatrix}$ is *i.i.d.* $\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\sigma^P)^2 & \rho\sigma^P\sigma^S \\ \rho\sigma^P\sigma^S & (\sigma^S)^2 \end{pmatrix}\right)$.

The usual two-sample t-statistics (T^P, T^S) for hypotheses in (4.3.1) are then defined as follows:

$$T^L = \frac{\bar{Y}_{1\cdot}^L - \bar{Y}_{0\cdot}^L - \delta^L}{\hat{\sigma}^L \sqrt{1/n_1 + 1/n_0}} \quad L = P, S, \quad (4.3.3)$$

where $\bar{Y}_{i\cdot}^L = \frac{1}{n_i} \sum_{r=1}^{n_i} Y_{ir}^L$, $\hat{\sigma}^L = \sqrt{\frac{\sum_{i=0}^2 \sum_{r=1}^{n_i} (Y_{ir}^L - \bar{Y}_{i\cdot}^L)^2}{\sum_{i=0}^2 (n_i - 1)}}$, $i = 0, 1$, $L = P, S$.

4.3.2 Predetermined steps

The standard procedure for this primary/secondary endpoint setting is the predetermined steps proposed in Hsu and Berger (1999). First test the the primary endpoint at level α . If that gets rejected, then tests the secondary endpoint at level α . Otherwise stop. Specifically, if $t_{\alpha, \nu}$ denotes the upper α quantile of the central t-distribution with $\nu = n_0 + n_1 - 2$ degrees of freedom, then

- Step 1: Is $T^P > t_{\alpha, \nu}$?

If yes, claim the primary endpoint is efficacious and go to step 2;

otherwise stop.

- Step 2: is $T^S > t_{\alpha, \nu}$?

If yes, claim the secondary endpoint is efficacious and stop;

otherwise stop.

In the co-primary endpoint context, this is exactly the same test as the intersection-union test that rejects if both hypotheses are rejected.

4.3.3 Motivation

From a partitioning point of view, the predetermined steps tests the following two partition hypotheses:

- Step 1: $H_0^P : \theta^P \leq \delta^P$ (the primary endpoint is not efficacious).
- Step 2: $H_0^{S*} : \theta^S \leq \delta^S$ and $\theta^P > \delta^P$ (the secondary endpoint is not efficacious but the primary is already efficacious).

The predetermined steps use the rejection region $\{T^P > t_{\alpha,\nu}\}$ for H_0^P , and $\{T^S > t_{\alpha,\nu}\}$ for H_0^{S*} because a level α test for H_0^S is also a level α test for H_0^{S*} . However, this test ignores the fact that when we test H_0^{S*} , the test statistic T^P is already in the rejection region of H_0^P . Incorporating this information can sometimes help one develop more powerful procedures. This is the main idea of strong partitioning introduced in Finner and Strassburger (2002). Thus we artificially make the rejection region for secondary endpoint inside the one for primary endpoint to see if more power can be gained. We now reject H_0^{S*} if $\{T^S > c\} \cap \{T^P > t_{\alpha,\nu}\}$. The Type I error for testing H_0^{S*} is calculated as follows:

$$\sup_{\{\theta^S \leq \delta^S, \theta^P > \delta^P\}} \Pr_{(\theta^P, \theta^S)} (\{T^S > c\} \cap \{T^P > t_{\alpha,\nu}\}) \quad (4.3.4)$$

For any fixed value c , the probability increases with θ^P and θ^S . Therefore, the supremum is obtained at $\theta^S = \delta^S$ and $\theta^P = +\infty$. Thus,

$$(4.3.4) = \Pr_{(\delta^S, \infty)}(\{T^S > c\} \cap \{T^P > t_{\alpha, \nu}\}) \quad (4.3.5)$$

$$= \Pr_{\delta^S}(T^S > c) \quad (4.3.6)$$

which brings us back the same critical value $t_{\alpha, \nu}$ as before.

Even though the new rejection region does not work, it provides us a new perspective. The failure to improve the predetermined steps is due to the fact that we used the same critical value c for every different primary endpoint θ^P . In reality, since the true mean difference between treatment and placebo for primary endpoint θ^P is most likely to be not very large, we can potentially use a smaller critical value for “small” θ^P s compared to “large” θ^P s.

Figure 4.8 shows the proposed rejection region on the right side compared to the usual intersection-union test on the left. The idea is to get smaller critical value $c_1 (< t_{\alpha, \nu})$ for the secondary endpoint when θ^P is small, by sacrificing the large θ^P s with a larger critical value $c_2 (\geq t_{\alpha, \nu})$.

4.3.4 Pre-specifying d for the primary endpoint

Before executing the procedure, an appropriate $d (> t_{\alpha, \nu})$ needs to be specified beforehand, as an upper bound on T^P . The choice may be based on prior information about θ^P from a clinical perspective. Then depending on the investigator’s goal, a set of (c_1, c_2) that is suitable in his/her specific situation also needs to be determined. Intuitively, increasing c_1 results in decreasing c_2 to keep the probability of the entire region under α . And the limiting case is $c_1 = c_2 = t_{\alpha, \nu}$.

Figure 4.8: Rejection region of pre-determined steps (left) and of the new proposed method (right) for testing $H_0^{S*} : \theta^S \leq \delta^S$ and $\theta^P > \delta^P$.

4.3.5 Calculating critical values

Given the specification of d ($> t_{\alpha,\nu}$) from the previous section, the critical values need to satisfy the following equation for controlling Type I error rate at level α .

$$\begin{aligned} & \sup_{\{\theta^S \leq \delta^S, \theta^P > \delta^P\}} \Pr_{(\theta^P, \theta^S)} \{T^S > c_1 \text{ and } T^P \in (t_{\alpha,\nu}, d]\} \\ & + \Pr_{(\theta^P, \theta^S)} \{T^S > c_2 \text{ and } T^P > d\} := \alpha \end{aligned} \quad (4.3.7)$$

For illustration purposes, we assume the clinically meaningful difference $\delta^P = \delta^S = 0$ and the test statistics T^P and T^S are independent with $N(\theta^P, 1)$ and $N(\theta^S, 1)$ distributions. In order to take the supremum over $\Theta^{S*} = \{\theta^S \leq 0 \text{ and } \theta^P > 0\}$, we

investigate the probability as a function of θ^P and θ^S .

$$\begin{aligned}
& \Pr_{(\theta^P, \theta^S)} \{T^S > c_1 \text{ and } T^P \in (z_\alpha, d]\} + \Pr_{(\theta^P, \theta^S)} \{T^S > c_2 \text{ and } T^P > d\} \\
&= \Pr_{(\theta^P, \theta^S)} \{T^S - \theta^S > c_1 - \theta^S \text{ and } T^P - \theta^P \in (z_\alpha - \theta^P, d - \theta^P]\} \\
&+ \Pr_{(\theta^P, \theta^S)} \{T^S - \theta^S > c_2 - \theta^S \text{ and } T^P - \theta^P > d - \theta^P\} \\
&= (1 - \Phi(c_1 - \theta^S)) (\Phi(d - \theta^P) - \Phi(z_\alpha - \theta^P)) \\
&+ (1 - \Phi(c_2 - \theta^S)) (1 - \Phi(d - \theta^P)) \tag{4.3.8}
\end{aligned}$$

where Φ is the cdf of standard normal distribution and $z_\alpha = \Phi^{-1}(1 - \alpha)$ is the $1 - \alpha$ quantile of standard normal distribution. We then take the partial derivative with respect to θ^S and θ^P .

$$\begin{aligned}
& \frac{\partial(4.3.8)}{\partial \theta^S} \\
&= \phi(c_1 - \theta^S) (\Phi(d - \theta^P) - \Phi(z_\alpha - \theta^P)) + \phi(c_2 - \theta^S) (1 - \Phi(d - \theta^P)) \\
&> 0 \quad \text{since we choose } d > z_\alpha \\
& \frac{\partial(4.3.8)}{\partial \theta^P} \\
&= (1 - \Phi(c_1 - \theta^S)) (-\phi(d - \theta^P) + \phi(z_\alpha - \theta^P)) \\
&+ (1 - \Phi(c_2 - \theta^S)) \phi(d - \theta^P) \\
&= (1 - \Phi(c_1 - \theta^S)) \phi(z_\alpha - \theta^P) - (\Phi(c_2 - \theta^S) - \Phi(c_1 - \theta^S)) \phi(d - \theta^P)
\end{aligned}$$

When $\theta^P = 0$, since $1 - \Phi(c_1 - \theta^S) > \Phi(c_2 - \theta^S) - \Phi(c_1 - \theta^S)$ and $\phi(z_\alpha) > \phi(d)$, the partial derivative with respect to θ^P is positive. Then when θ^P increases to certain large value, the derivative turns negative. In the limit when θ^P goes to infinity, it approaches 0. Thus, we know the supremum of the probability of rejection takes place

when

$$\begin{aligned} (1 - \Phi(c_1 - \theta^S)) \phi(z_\alpha - \theta^P) &= (\Phi(c_2 - \theta^S) - \Phi(c_1 - \theta^S)) \phi(d - \theta^P) \\ \Rightarrow \theta_m^P &= \frac{-1}{d - z_\alpha} \log \frac{\Phi(c_2 - \theta^S) - \Phi(c_1 - \theta^S)}{1 - \Phi(c_1 - \theta^S)} + \frac{d + z_\alpha}{2} \end{aligned} \quad (4.3.9)$$

Combining results from both θ^S and θ^P , we have

$$\begin{aligned} &\sup_{\Theta^{S*}} \Pr_{(\theta^P, \theta^S)} \{T^S > c_1 \text{ and } T^P \in (z_\alpha, d]\} + \Pr_{(\theta^P, \theta^S)} \{T^S > c_2 \text{ and } T^P > d\} \\ &= (1 - \Phi(c_1)) (\Phi(d - \theta_m^P) - \Phi(z_\alpha - \theta_m^P)) + (1 - \Phi(c_2)) (1 - \Phi(d - \theta_m^P)) \\ &:= \alpha \end{aligned} \quad (4.3.10)$$

Since θ_m^P involves both c_1 and c_2 , there is no closed form expression of c_2 as a function of c_1 in (4.3.10). Figure 4.9 shows plots of c_2 as a function of c_1 for $d = 2.1$ (blue) and $d = 2.5$ (red) through numerical calculation using *Mathematica 7.0*. It is expected that c_2 decreases in c_1 for fixed d , and c_2 increases in d for fixed c_1 . For moderately large d , for example $d = 2.1$, c_1 is 0.5, much smaller than $z_\alpha = 1.96$, while c_2 is 2.19, only slightly over 1.96. When c_1 increases to 1.96, c_2 decreases to 1.96 as well for both d values, leading to the special case of predetermined steps.

We will do a power comparison between the predetermined steps (or equivalently the intersection-union test) and the proposed new rejection region in Figure 4.8 with $c_1 = 0.5$, $c_2 = 2.19$, $d = 2.1$. The proposed new test is not uniformly more powerful. But for small θ^P values which are more important and more likely to be the case in practice, the new test should be more powerful.

Figure 4.9: Plot of c_2 as function of c_1 for $d = 2.1$ in blue and $d = 2.5$ in red.

4.4 Finding Minimum Effective Dose using confidence bounds in dose-response studies

Dose-response studies can be encountered in almost every stage of drug development. It is especially important in Phase II studies for finding an appropriate dose range and Phase III studies for identifying the lowest dose at which the drug is effective. In this section, we propose a simultaneous confidence bounds through which the minimum effective dose (MED) can be identified in a typical Phase III drug trial setting. This approach is in a general form, and thus is suitable for either discrete or continuous dose cases.

There are generally two strategies in analyzing dose-response data. One is the use of multiple comparisons procedures, where inferences are made on the observed discrete dose levels $\{d_1, \dots, d_k\}$. The other is the use of modeling techniques, where

a functional relationship between dose and response can be found through a pre-specified parametric model with inferences made on the entire dose range $[d_1, d_k]$. The multiple comparisons approach has been the one used in practice, however, because

1. The inverse regression aspect makes the modeling approach not straight forward;
2. Non-linear curve fitting has not been very well developed.

To make inference on the entire dose range $[d_1, d_k]$ while still preserving the nice properties about multiple comparisons procedures, we extend the multiple comparisons approach by allowing the number of hypotheses to go to infinity. Since FWER is still well defined with an infinite number of hypotheses, we can use the partitioning principle to form a test procedure that provides a $1 - \alpha$ level upper bound on MED using pair-wise confidence bounds without a multiplicity adjustment. What motivates us to develop this new procedure is

1. Partitioning gives insight into the inverse regression aspect;
2. With the work of our co-authors Jose Pinheiro and Frank Bretz, non-linear modeling has become more feasible.

Note this new procedure works with either discrete or continuous doses.

4.4.1 Preliminaries

Let d_i denote the increasing dose levels for group $i, i = 1, \dots, k$ (group 1 is the placebo group) and D denote the dose set upon which inferences are made. Depending on which approach one uses, D could be either discrete $\{d_1, \dots, d_k\}$ or continuous $[d_1, d_k]$, but the true dose-response is always a function on the entire dose range

$[d_1, d_k]$. Let

$$\mathcal{F} = \{f : [d_1, d_k] \rightarrow \mathbb{R}\} \quad (4.4.1)$$

be the function space containing all the functions from the dose range $[d_1, d_k]$ to the real line in which the true dose response profile f lies.

We extend the multiple hypothesis testing framework to the continuous dose range case by allowing the number of hypotheses to be infinity. The family of hypotheses is then defined as:

$$H_{0d} : \text{dose } d \text{ is not efficacious, } d \in D$$

Let Θ_d be the collection of functions that satisfy $\{f \in \mathcal{F} : f(d) - f(d_1) \leq \delta\}$ then we could write equivalently,

$$H_{0d} : f \in \Theta_d, d \in D \quad (4.4.2)$$

Thus the familywise error rate (FWER) is defined as the probability of making any false rejection:

$$\text{FWER} = \sup_{f \in \mathcal{F}} P_f\{V > 0\} \quad (4.4.3)$$

where the supremum is now taken over $\mathcal{F}$. Other types of error rates, such as the false discovery rate (FDR) or generalized familywise error rate (gFWER) do not make much sense in this case because they require the counts of false or correct rejections.

Hsu and Berger (1999) and Bretz, Pinheiro and Branson (2005) defined the minimum effective dose (MED) as the minimum dose that the drug is effective in a multiple comparison and modeling setting, respectively.

$$\text{MED}_1 = \begin{cases} \min\{i : \mu_i - \mu_1 > \delta\} \\ \operatorname{argmin}_{d \in (d_1, d_k]} \{f(d) - f(d_1) > \delta\} \end{cases} \quad (4.4.4)$$

where δ is a positive quantity defining a clinically meaningful difference. However, neither definitions restrict the responses for doses higher than MED. In other words,

the doses higher than MED are not guaranteed to be efficacious. To avoid such situations, we define MED (4.4.5) as the minimum dose such that not only itself but also all dose higher are efficacious:

$$\text{MED}_2 = \operatorname{argmin}_{d \in D} \{f(d') - f(d_1) > \delta, \forall d' \geq d, d' \in D\} \quad (4.4.5)$$

where $D = \{d_2, \dots, d_k\}$ for discrete doses and $D = (d_1, d_k]$ for continuous doses. This definition agrees with the one discussed in Section 3.2 of Bretz, Hsu, Pinheiro and Liu (2008) in the multiple comparison setting. By definition, MED_2 is no smaller than MED_1 .

Some multiple comparison procedures such as predetermined steps in Hsu and Berger (1999) for finding MED do not place any assumption on the true dose response profile. That is, the true dose response function could be any $f \in \mathcal{F}$. Some other procedures, however, are valid only if the true dose response profile satisfies certain assumptions. Tamhane et al. (2001) assumes weak monotonicity (in multiple comparison setting, in terms of maximum safe dose) which states that there exists one dose $d^* \in D$ such that every dose below it is not efficacious and every dose greater than or equal to it is efficacious.

$$f(d) > f(d_1) + \delta, \quad \text{for all } d \geq d^* \quad (4.4.6)$$

$$f(d) \leq f(d_1) + \delta, \quad \text{for all } d < d^*$$

If the dose response function f satisfies weak monotonicity assumption (4.4.6), then the two definitions of MED, MED_1 and MED_2 , are the same.

4.4.2 Finding MED using confidence bounds without multiplicity adjustment

Suppose we have the one-sided point-wise confidence bounds $(L_{d-d_1}, +\infty)$, $d \in D$ for the difference in response between dose d and d_1 $f(d) - f(d_1)$, with confidence level $1 - \alpha$. That is

$$\sup_{f \in \Theta_d} \Pr_f\{L_{d-d_1} \geq \delta\} \leq \alpha \quad (4.4.7)$$

We can use these confidence bounds to make inference on MED_2 without multiplicity adjustment. The procedure starts from the highest dose d_k , and finds the lowest dose in D such that the lower confidence limits for the doses higher than or equal to it exceed δ . Formally, we infer

$$\widehat{MED} = \operatorname{argmin}_{d \in D} \{L_{d'-d_1} > \delta, \forall d' \geq d, d' \in D\} \quad (4.4.8)$$

This procedure works for dose-response functions in any subsets of $\mathcal{F}$ (if some assumption on the dose-response function is made) or the entire $\mathcal{F}$ (if no such assumption is made). It also allows dose set D to be either discrete or continuous. If the dose set $D = \{d_1, \dots, d_k\}$, this procedure (4.4.8) is equivalent to the pre-determined steps (Hsu and Berger, 1999). If weak monotonicity or strong monotonicity (non-decreasing responses) is assumed or the true dose response curve is not inverted-U shaped, the procedure will not stop too soon by starting with the highest dose level.

We now prove in Theorem 4.4.1 that the estimated MED in (4.4.8) provides an upper bound on the true MED.

Theorem 4.4.1. *For either discrete or continuous doses, the estimated MED defined in (4.4.8) is an upper bound on MED_2 . That is*

$$\inf_{\{f: MED = MED_2\}} \Pr_f\{\widehat{MED} \geq MED_2\} \geq 1 - \alpha. \quad (4.4.9)$$

The proof of the Theorem is given in Appendix Section 4.4.3.

In this section, we unified the two separate techniques in dose-response analysis and proposed a procedure that can be viewed as an extension of predetermined steps in continuous dose range setting. The procedure provides an upper bound estimate of MED and works with no assumption on the dose-response profile.

4.4.3 Appendix

Proof of Theorem 4.4.1

We first partition the parameter space $\cup_{d \in D} \Theta_d$ where Θ_d is defined in (4.4.2). For each dose $d \in D$, we partition the parameter space into

$$\begin{aligned} \Theta_d^+ &= \{f \in \mathcal{F} : f(d) - f(d_1) \leq \delta < f(d') - f(d_1), d' > d, d' \in D\} \\ &= \{f \in \mathcal{F} : \text{dose } d \text{ is not efficacious, but every dose higher is efficacious}\} \\ &= \{f \in \mathcal{F} : \text{MED}_2 = d^+\} \end{aligned} \tag{4.4.10}$$

where d^+ is the lowest dose among all the doses higher than d .

We now show that $\Theta_d^+, d \in D$ is indeed a partition of the parameter. That is, the following two conditions are met with $\mathcal{F}^+ = \{f \in \mathcal{F} : f(d) - f(d_1) > \delta, \forall d > d_1, d \in D\}$.

1. $\forall s \neq t \in D, \Theta_s^+ \cap \Theta_t^+ = \emptyset$.

WLOG, suppose $s < t$, and $f^* \in \Theta_s^+$, then by definition,

$$f^*(s) - f^*(d_1) \leq \delta < f^*(d') - f^*(d_1), \text{ for } d' > s, d' \in D.$$

Since $t > s, t \in D$, we have $f^*(t) - f^*(d_1) > \delta$. Thus

$$f^* \notin \Theta_t^+ = \{f \in \mathcal{F} : f(t) - f(d_1) \leq \delta < f(d') - f(d_1), \forall d' > t, d' \in D\}.$$

That is $\Theta_s^+ \cap \Theta_t^+ = \emptyset$.

$$2. \underline{\cup_{d \in D} \Theta_d^+ = \mathcal{F} \setminus \mathcal{F}^+ = \{f \in \mathcal{F} : \exists d > d_1, d \in D, \text{ s.t. } f(d) - f(d_1) \leq \delta\}}.$$

Since $\cup_{d \in D} \Theta_d^+ \subseteq \mathcal{F} \setminus \mathcal{F}^+$, we only need to show $\cup_{d \in D} \Theta_d^+ \supseteq \mathcal{F} \setminus \mathcal{F}^+$. For any $f \in \mathcal{F} \setminus \mathcal{F}^+$, let

$$d^* = \operatorname{argmin}_{d \in D} \{f(d') - f(d_1) > \delta, \forall d' > d, d' \in D\},$$

then $f(d^*) - f(d_1) \leq \delta$, and $f(d') - f(d_1) > \delta, d' > d^*, d' \in D$. That is $f \in \Theta_{d^*}^+$.

We then reject the partition hypothesis $H_{0d}^+ : f \in \Theta_d^+$ if $L_{d-d_1} > \delta$, the same rejection region as H_{0d} . This is a level α test for H_{0d}^+ because $\Theta_d^+ \subseteq \Theta_d$. By partitioning principle, the FWER of the family of hypotheses $H_{0d}, d \in D$ is controlled at level α because the type I error rate of each partition test H_{0d}^+ is α .

Suppose MED_2 is the true minimum effective dose, then

$$\sup_{\{f: MED=MED_2\}} \Pr_f\{\widehat{MED} < MED_2\} \leq \sup_{\{f: MED=MED_2\}} \Pr_f\{L_{MED_2-d_1} > \delta\} \leq \alpha$$

APPENDIX A

R FUNCTION “DOSEENDPOINTS” IMPLEMENTING DECISION PATH APPROACH

A.1 Documentation

Description

DoseEndpoints compares k dose groups to a negative or an active control group, for efficacy claim on m endpoints, while controlling the Familywise Error Rate (FWER). For each dose group, decisions on endpoints follow specific paths. That is, endpoints with lower orders are not tested unless higher ordered ones have shown efficacy.

Usage

```
DoseEndpoints(object,alpha=0.025,delta=0,alternative="greater")
```

Arguments

object	a list object with at least two components: <code>coef</code> and <code>vcov</code> (typically from some modeling output such as <i>lm</i>)
coef	m by k matrix with each column as mean difference between dose and control group for m endpoints.
df.residual	m by 1 vector representing error degrees of freedom for the m endpoints. If not specified, the default value is 0, and multivariate normal distribution is used in the calculations instead of multivariate t distribution.
vcov	m by k by k array with first index representing each endpoint, 2nd and 3rd indices representing estimated covariance between estimated group mean differences.
alpha	FWER to be controlled.
delta	m by 1 vector representing clinically meaningful differences for the m endpoints (see Details for different ways of specifying delta).
alternative	“greater” is the default value indicating alternative hypothesis is of the form “>”. Similarly, “less” indicates “<” while “two.sided” indicates “≠”.

Details

Let β_{ij} be the true mean difference between group i and the control group for endpoint j , $i = 1, \dots, m$, $j = 1, \dots, k$. Let $\hat{\beta}_{ij}$, the (i, j) th element of matrix `object$coef`, be the estimated mean difference between group i and the control group for endpoint j , $i = 1, \dots, m$, $j = 1, \dots, k$. Also, denote by δ_i , the i th element of delta, ν_i , the i th element of `object$df.residual`, and s_{ij}^2 , the variance of $\hat{\beta}_{ij}$, the (i, j, j) th element of `object$vcov`.

Family of hypotheses

The family of hypotheses being tested for each alternative option is:

$$\text{alternative}=\text{"greater"} \quad H_{0ij} : \beta_{ij} \leq \delta_i \quad \text{vs.} \quad H_{aij} : \beta_{ij} > \delta_i$$

$$\text{alternative}=\text{"less"} \quad H_{0ij} : \beta_{ij} \geq \delta_i \quad \text{vs.} \quad H_{aij} : \beta_{ij} < \delta_i$$

$$\text{alternative}=\text{"two.sided"} \quad H_{0ij} : \beta_{ij} = \delta_i \quad \text{vs.} \quad H_{aij} : \beta_{ij} \neq \delta_i$$

$$i = 1, \dots, m; \quad j = 1, \dots, k$$

Thus,

- If superiority is desired comparing the dose groups to the negative control, delta is specified as a positive number in the case of alternative="greater", and a negative number for alternative="less".
- If non-inferiority is desired comparing the dose groups to the active control, delta is specified as a negative number in the case of alternative="greater", and a positive number for alternative="less".

Decision Paths

The endpoints are ordered from 1 to m for each dose. There is no ordering between the doses. Decision paths (orders in testing the hypotheses) are:

- H_{0ij} is tested only if $H_{0i'j}$ for all $i' < i$ are rejected for the same dose j .
- Tests for different doses are not ordered.

Partition testing is used to generate $(m + 1)^k - 1$ partition hypotheses. Each partition hypothesis can be tested at level alpha to maintain the FWER at the same

alpha level, since only one contains the true parameter. The decisions after collating results from testing these partition hypotheses are guaranteed to follow the Decision Paths.

Test Statistics

The test statistics for hypothesis H_{0ij} are

$$T_{ij} = (\hat{\beta}_{ij} - \delta_i) / \sqrt{s_{ij}^2}$$

The test statistics have t-distribution marginally with ν_i as the degrees of freedom for endpoint i . Note, if $\nu_i=0$, then normal distribution is used instead.

Within the same endpoint i , $(T_{i1}, \dots, T_{ik})$ has a multivariate t distribution with known correlation. However, for test statistics from different endpoints, the distribution is no longer multivariate t (see Liu, Hsu, and Ruberg 2007).

Computation of the critical values is exactly as described in section 4.4 of Liu and Hsu (2009). Within each endpoint, it makes use of the exact joint distribution of test statistics. Therefore, our method is more powerful than methods that ignore the joint distribution, as illustrated in our Examples. In the one endpoint case, our partition testing method (acting as a step-down Dunnett's method), is more powerful than Holm's method and 1-step Dunnett's method, and can be more powerful than Hochberg's method. The simulation study in Liu and Hsu (2009) also indicates our method is more powerful than the gatekeeping method of Dmitrienko et al. (2006).

Value

A list containing the rejection matrix and all the input information.

final.reject	m by k rejection matrix with (i, j) th element being 1 (reject) or 0 (fail to reject) for testing H_{0ij} .
coefficient	object\$coef
covariance	object\$vcov
df	object\$df.residual
FWER.control	alpha
delta	delta
alternative.hypothesis	alternative

Examples

Two examples are used to illustrate the use of *DoseEndpoints*.

Anti-psychotic drug “Seroquel” (Quetiapine)

This example is a 26-center double-blind trial comparing the effect of five doses of the anti-psychotic drug “Seroquel” (Quetiapine, or Qtp for short) and the placebo with parallel design on a total of 361 patients, as described in Arvanitis et al. (1997). An active control, “Haloperidol” is also included in the trial with a standard dose to compare with the test drug. Three of the endpoints, with orders from high to low, BPRS total score (primary), CGI Severity of Illness score (most important secondary), and SANS summary score (least important secondary) are used in our illustration. Figure A.1 contains plots of mean changes from baseline for each of the endpoints. The plots also show the negative superiority margins as well as the positive non-inferiority margins since a lower score indicates a better result.

The following analyses are done with FWER controlled at 0.025.

- 1) Test superiority of Seroquel by comparing the 5 dose groups to the Placebo with $\delta = -0.30$ for BPRS total score, -0.18 for CGI Severity of Illness score, and 0.00 for SANS summary score. Controlling FWER at 0.025, the resulting rejection matrix is:

Figure A.1: Mean change from baseline for placebo (P), five dose groups (Qtp75, Qtp150, Qtp300, Qtp600, Qtp750) and active control group (A) in BPRS total score, CGI Severity of Illness score, and SANS summary score. The two dotted horizontal lines in each graph are mean change of placebo group plus a negative delta (for superiority) and mean change of active control plus a positive delta (for non-inferiority).

```
$final.reject
      [,1] [,2] [,3] [,4] [,5]
[1,]    0    1    1    1    1
[2,]    0    1    1    0    1
[3,]    0    0    1    0    0
```

Therefore, four doses (except Qtp 75) are superior to the placebo, showing efficacy for BPRS total score. Dose 150, 300 and 750 show additional efficacy in the CGI Severity of Illness score. In addition, Dose 300 is also efficacious for the SANS summary score.

- 2) Compare to Holm's, Hochberg's, and 1-step Dunnett's method, based on dose group 1, 3, 4 for the SANS summary score only. With $\delta = -0.11$, dose group 3 and 4 are rejected using *DoseEndpoints*.

```
$final.reject
[1] 0 1 1
```

However, only dose group 3 is rejected by either Holm's or Hochberg's with the following raw p-values. *DoseEndpoints* (equivalent to step-down Dunnett in the one endpoint case, section 3.1.1.2 of Hsu 1996, appropriately computed for unbalanced design) is more powerful in this case because it utilizes the joint distribution among the test statistics.

```
> rawp
[1] 0.042833501 0.001241597 0.012596262
```

The critical value for Dunnett’s method and the t-statistics (taken negative to be consistent with a positive critical value) are:

```
[1] 2.361696
> -Tstat
[1] 1.723875 3.048817 2.248675
```

Again, only the dose group 3 is rejected by the 1-step Dunnett’s method while *DoseEndpoints* (equivalent to the step-down Dunnett in the case of a single endpoint) infers both dose 3 and 4 as efficacious.

- 3) Test non-inferiority of the test drug comparing to the active control Haloperidol with $\delta=(6.6,0.7,1.8)$ for the three endpoints. The rejection matrix is:

```
$final.reject
      [,1] [,2] [,3] [,4] [,5]
[1,]    0    1    1    0    0
[2,]    0    0    1    0    0
[3,]    0    0    0    0    0
```

Dose group 2 and 3 can be claimed non-inferior to the active control by 6.6 for BPRS total score. In addition, dose group 3 is non-inferior to the active control by 0.7 for CGI Severity of Illness score.

- 4) Test assay sensitivity by comparing the positive control “Haloperidol” with the placebo, and see if it is superior with $\delta=(-0.3,-0.18,0)$ for the three endpoints. The rejection matrix is:


```
$final.reject
```

```
[,1]
```

```
[1,]    1
```

```
[2,]    1
```

```
[3,]    1
```

The results indicate that Haloperidol is in fact superior to the placebo by delta for all endpoints. Note, using *DoseEndpoints* in the one dose multiple endpoints case is the same as the pre-determined steps method in Hsu and Berger (1999).

Had we included Haloperidol in the analysis of 1), it would have made no difference in the results for the five Qtp dose groups in this case, because of Dose Multiplicity Adjustment: If a dose is shown to be efficacious in all endpoints, then that dose should no longer be included in multiplicity adjustment, as described in section 4.4 of Liu and Hsu (2009). (This would be expected in testing multiple doses, in the case of a single endpoint.)

Hypertension drug trial

This example is a dose-response study conducted to evaluate the effect of a hypertension drug, as described by Dmitrienko, Offen and Westfall (2003). The study compares three doses (low, medium and high) to the placebo with respect to two endpoints, systolic blood pressure (primary) and diastolic blood pressure (secondary). The mean changes from baseline on both endpoints for different dose groups of this study are plotted in Figure A.2. As lower blood pressure is better for hypertensive patients, a negative delta is specified for each endpoint to show superiority.

The following analyses are done controlling FWER at 0.025 unless otherwise noted.

Figure A.2: Mean change from baseline in systolic and diastolic blood pressure. The two dotted horizontal lines are mean change for placebo group plus delta (-0.4 for systolic blood pressure and -0.2 for diastolic blood pressure).

- 1) Show how to pipe output from a modeling function such as *lm* into the function *DoseEndpoints*.

An artificial data set is created matching the summary statistics from Dmitrienko et al. (2003). The linear modeling function *lm* is used to fit the data for each endpoint. Combine values from each endpoint to form the list object required by *DoseEndpoints*.

- 2) Test superiority for both systolic blood pressure and diastolic blood pressure with $\delta = (-0.4, -0.2)$. The rejection matrix is:

```
$final.reject
      [,1] [,2] [,3]
[1,]     1     1     1
[2,]     0     1     0
```

All dose groups are superior to the placebo group in reducing systolic blood pressure by at least 0.4 mmHg. In addition, Medium dose group shows a significant reduction by at least 0.2 mmHg in diastolic blood pressure.

- 3) Test superiority for diastolic blood pressure (secondary endpoint) only with the same $\delta = -0.2$.

```
$final.reject
[1] 0 1 0
```

The result is the same as the second row in 2), testing superiority for both endpoints. This illustrates Endpoint Prioritization by *DoseEndpoints*: If efficacy

in a higher ordered endpoint is shown for all doses, then testing proceeds as if the next highest ordered endpoint were one order higher, as described in section 4.4 of Liu and Hsu (2009).

- 4) Test superiority with the same delta as 2), using two-sided tests at 0.05 level.

```
$final.reject
      [,1] [,2] [,3]
[1,]    1    1    1
[2,]    0    1    0
```

The decisions are the same as those in 2 for this example. However, in MULTIPLE testing, a 0.05 two-sided test has an error rate slightly higher than 0.025 when viewed as a one-sided test.

A.2 Sample Code

```
#####  
####      Psychotic drug example (Arvanitis et al. 1997)      ####  
####      This example shows how to test for superiority and   ####  
####      non-inferiority and how to specify delta in each case. ####  
####      It shows that our partition testing method (which acts ####  
####      as step-down Dunnett's method in the one endpoint case) ####  
####      is more powerful than Holm's and 1-step Dunnett, and can ####  
####      be more powerful than Hochberg's.                    ####  
#####  
### input summary data including mean, standard error, sample  
### size for each endpoint  
  
BPRStot <- c(1.71,-2.24,-8.67,-8.59,-7.68,-6.33,-7.58)  
BPRStotse <- c(2.06,2.04,2.14,2.06,2.08,2.02,2.10)  
BPRStotn <- c(51,52,48,51,51,53,50)  
  
CGIsevn <- c(0.25,-0.15,-0.49,-0.69,-0.46,-0.46,-0.69)  
CGIsevnse <- c(0.15,0.15,0.16,0.15,0.16,0.15,0.16)  
CGIsevn <- c(51,52,48,51,51,54,50)  
  
SANSsum <- c(0.76,-0.62,-0.78,-1.56,-0.98,-0.50,-1.83)  
SANSsumse <- c(0.51,0.53,0.54,0.51,0.52,0.52,0.51)  
SANSsumn <- c(50,46,45,49,49,48,50)  
  
### number of endpoints  
  
m <- 3
```

```

### number of dose groups
k <- 6

### mean matrix
mean <- rbind(BPRStot,CGIsev,SANSsum)

### mean difference between dose group and placebo group
meandif <- mean[,2:(k+1)]-mean[,1]

### group size for each endpoint
gsize <- rbind(BPRStotn,CGIsevn,SANSsumn)

### degree of freedom for each endpoint
df <- apply(gsize-1,1,sum)

### pooled standed deviation for each endpoint
poolsd <-
  c(sqrt(sum((BPRStotn-1)*BPRStotn*BPRStotse^2)/sum(BPRStotn-1)),
    sqrt(sum((CGIsevn-1)*CGIsevn*CGIsevse^2)/sum(CGIsevn-1)),
    sqrt(sum((SANSsumn-1)*SANSsumn*SANSsumse^2)/sum(SANSsumn-1)))

### covariance matrix for each endpoint
vcov <- array(0,dim=c(m,k,k))
for (en in 1:m){
  vcov[en,,] <- (diag(c(1/gsize[en,2],1/gsize[en,3],1/gsize[en,4],
    1/gsize[en,5],1/gsize[en,6],1/gsize[en,7]),k)
    +matrix(1/gsize[en,1],ncol=k,nrow=k))*poolsd[en]^2

```

```
}
```

```
### test whether 5 dose groups of the test drug (excluding
### Haloperidol group) is superior to the placebo by at least delta
sup.lst <- list(coef=meandif[,-6],df.residual=df,vcov=vcov[,-6,-6])
DoseEndpoints(sup.lst,0.025,delta=c(-0.3,-0.18,0),alternative="less")
```

```
### DoseEndpoints acts as step-down Dunnett for one endpoint case
### (since step-down testing is a shortcut to partition testing).
### As described in Chapter 3 of Hsu (1996), partition testing
### is step-down Dunnett in the one endpoint case, and confidence
### intervals are available for this case.
```

```
### Comparing to Holm's, Hochberg's, and single step Dunnett's method,
### the following example based on the last endpoint with dose 1,3,4
### shows that our method rejects more dose groups than these
### three methods.
```

```
sup1e3d.lst <- list(coef=meandif[3,c(1,3,4)],df.residual=df[3],
                    vcov=vcov[3,c(1,3,4),c(1,3,4)])
```

```
### Note: in one endpoint case, coef and vcov can be specified in two
### ways. One way is to specify coef as a vector and vcov as a matrix.
### The other way is to specify coef as a matrix and vcov as an array.
DoseEndpoints(sup1e3d.lst,0.025,delta=-0.11,alternative="less")
```

```

### dose 3 and 4 are rejected

Tstat <- (sup1e3d.lst$coef-(-0.11))/sqrt(diag(sup1e3d.lst$vcov))

rawp <- pt(Tstat,df=sup1e3d.lst$df.residual)

orderedp <- rawp[order(rawp)]

orderedp

c(0.025/3,0.025/2,0.025)

### only one dose is rejected by either Holm's or Hochberg's method.

### compare to 1-step Dunnett method

getcor <- function(covmat){
  d <- dim(covmat)[1]
  diag(1/sqrt(diag(covmat)),d)%*%covmat%*%
    diag(1/sqrt(diag(covmat)),d)
}

cormat <- getcor(sup1e3d.lst$vcov)

qmvtn(1-.025,df=sup1e3d.lst$df.residual,tail="lower.tail",sigma=cormat,
abseps=0.00001)$quantile

-Tstat

### only one dose group is rejected by one step Dunnett's method.

### test whether 5 dose groups of the test drug is non-inferior to

### the active control Haloperidol

### mean difference between dose group and active control group

meandif.noninf <- mean[,2:k]-mean[,k+1]

```



```

### covariance matrix for each endpoint
vcov.noninf <- array(0,dim=c(m,5,5))
for (en in 1:m){
  vcov.noninf[en,,] <-
    (diag(c(1/gsize[en,2],1/gsize[en,3],1/gsize[en,4],
    1/gsize[en,5],1/gsize[en,6])),5)
    +matrix(1/gsize[en,7],ncol=5,nrow=5))*poolsd[en]^2
}

noninf.lst <-
list(coef=meandif.noninf,df.residual=df,vcov=vcov.noninf)
DoseEndpoints(noninf.lst,0.025,delta=c(6.6,0.7,1.8),alternative="less")

### test whether the positive control Haloperidol is able to show
### superiority comparing to Placebo.
### Note: for multiple dose, one endpoint case, coef and df.residual
### must be an m by 1 matrix, and the covariance need to be
### m by 1 by 1 array.

positive.lst <- list(coef=matrix(meandif[,6]),df.residual=matrix(df),
vcov=array(vcov[,6,6],dim=c(3,1,1)))
DoseEndpoints(positive.lst,0.025,delta=c(-0.3,-0.18,0),
              alternative="less")

```

```
#####

#####      hypertension example (Dmitrienko et al. 2003) #####

##### This example illustrates how to feed the output of a #####

##### model such as lm into the function DoseEndpoints. #####

#####

### create artificial data set from summary statistics

Summary2Data <-

function (summary = stop("Argument 'summary' is missing"))

{

  coef <- summary$coef

  N <- summary$N

  sd <- summary$sd

  k <- length(coef)

  x <- rep(0, sum(N))

  y <- rep(0, sum(N))

  e <- rnorm(N[1])

  x[1:sum(N[1:1])] <- rep("Trt1", N[1])

  y[1:sum(N[1:1])] <- coef[1] + sd[1]*(e-mean(e))/sqrt(var(e))

  for (i in 2:k) {

    e <- rnorm(N[i])

    x[(sum(N[1:i-1])+1):sum(N[1:i])] <-

                                rep(paste("Trt",i,sep=""), N[i])

    y[(sum(N[1:i-1])+1):sum(N[1:i])] <- coef[i] +

    sd[i]*(e-mean(e))/sqrt(var(e))

  }

}
```

```

    }

    data.frame(y,group=x)
}

### create input list for the function Summary2Data
systolic <- list(coef=c(4.02,-2.16,-5.03,-2.60),N=c(29,23,24,24)
, sd=c(7.85,8.71,11.01,8.96))
diastolic <- list(coef=c(2.31,-0.67,-3.18,-1.44),
N=c(29,23,24,24),sd=c(5.88,5.51,6.89,6.15))
systolic.dat <- Summary2Data(systolic)
diastolic.dat <- Summary2Data(diastolic)

### run artificial data through lm
systolic.lm <- lm(y~group,data=systolic.dat)
diastolic.lm <- lm(y~group,data=diastolic.dat)

### input number of doses and endpoints
k <- 3
m <- 2

### combine results from two endpoints
### the coef from lm is already estimates of mean differences
hyper.coef <- rbind(systolic.lm$coef,diastolic.lm$coef)[,-1]
hyper.df <- c(systolic.lm$df.residual,diastolic.lm$df.residual)
hyper.vcov <- array(0,dim=c(m,k,k))
hyper.vcov[1,,] <- vcov(systolic.lm)[-1,-1]

```

```

hyper.vcov[2,,] <- vcov(diastolic.lm)[-1,-1]

### test superiority for both endpoints
hyper.lst <- list(coef=hyper.coef,df.residual=hyper.df,
                 vcov=hyper.vcov)
DoseEndpoints(hyper.lst,0.025,delta=c(-.4,-0.2),alternative="less")

### test superiority for secondary endpoint only
hyper1vec.lst <- list(coef=hyper.coef[2,],df.residual=hyper.df[2],
                    vcov=hyper.vcov[2,,])
DoseEndpoints(hyper1vec.lst,0.025,delta=-0.2,alternative="less")
### The result is the same as the result from testing both endpoints
### shown above, because all doses for primary endpoint are inferred
### to be efficacious. In that case, our method acts as if the
### secondary endpoint is the only endpoint.

### try the other way specifying coef as vector and vcov as matrix
hyper1emat.lst <- list(coef=t(-hyper.coef[2,]),
                     df.residual=hyper.df[2],
                     vcov=array(hyper.vcov[2,,],dim=c(1,k,k)))
DoseEndpoints(hyper1emat.lst,delta=.2)

### try two-sided testing
DoseEndpoints(hyper.lst,0.05,delta=c(-.4,-0.2),

```

```
alternative="two.sided")
```

A.3 R codes for “DoseEndpoints”

```
#####  
#### The main purpose of this code file is to calcualte critical ####  
#### values with k doses, m endpoints. The paths are from ####  
#### endpoint 1 (primary) to E2, E3, ect for each dose. One gets ####  
#### to test the next endpoint only if the endpoint with higher ####  
#### orders are rejected for each dose. ####  
#### Copyright(c) 2009 Yi Liu and Jason C. Hsu ####  
#### Last modified Aug, 9, 2009, version 0.9.6 ####  
#### This software is distributed WITHOUT ANY WARRANTY. ####  
#####  
  
#Let m be the number of endpoints, and k be the number of dose groups,  
#then the total number of groups (including the control group) is k+1.  
#The inputs is a list object which has at least three components  
#1. coef -- m * k matrix with each column as mean difference  
# between dose and control group for m endpoints.  
#2. df.residual -- m*1 vector representing error degrees  
# of freedom for each endpoint  
#3. vcov -- m*k*k array with first index representing each endpoint,  
# 2nd and 3rd index representing covariance between the group mean  
# differences.
```

```

#alpha -- FWER to be controlled

#delta -- m*1 vector representing clinically meaningful difference

#           for each endpoint

#alternative -- "greater" is the default value indicating alternative
#hypothesis is of the form ">". Similarly, "less" indicates "<" while
#"two.sided" indicates "\=".

```

```

DoseEndpoints <-
function(object,alpha=0.025,delta=0,alternative="greater"){

#check whether list object has at least two components
if (length(object)<2) stop("object needs to have at least two
components: coef and vcov")

#check whether the option alternative is one of the three choices
if (alternative!="greater" & alternative!="less" & alternative!="
two.sided") stop("alternative has to be one of greater, less,
and two.sided")

library(mvtnorm)

#####if more than one endpoint #####

if (is.matrix(object$coef)) {

#number of endpoints

```

```

numde <- dim(object$coef)

m <- numde[1]

#number of dose groups

k <- numde[2]

#get the test statistics

Tstat <- matrix(0,ncol=k,nrow=m)

#if delta is the default value, get to the same dimension
if (length(delta)==1) {
  #if the default, make the dimension match the number of endpoints
  if (delta==0) delta <- rep(0,m)
}

#check whether the dim of delta is correct if not using the default
if(length(delta)!=m) stop("the dimentstion of delta does not
match the number of endpoints")

#check whether the covariance matrix has the correct dimension
if (length(dim(object$vcov))!=3)
  {stop("the dimension of object$vcov does not match
object$coef")} else
  if (sum(dim(object$vcov)==c(m,k,k))!=3)
    {stop("the dimension of object$vcov does not match
object$coef")}

#otherwise, the nonzero single delta should indicate that m=1

```

```

if(k>1) {
  for (en in 1:m){
    Tstat[en,] <- (object$coef[en,]-delta[en])/
                                sqrt(diag(object$vcov[en,,]))
  }
} else {
  for (en in 1:m){
    Tstat[en,] <- (object$coef[en,]-delta[en])/
                                sqrt(object$vcov[en,,])
  }
}

### if the option "alternative" indicates "less",
### take negative of the original Tstat
if (alternative=="less") Tstat <- -Tstat
if (alternative=="two.sided") Tstat <- abs(Tstat)

#get the degree of freedom for each endpoint
#if df.residual is not specified, use 0 for normal distribution
if (length(object)==2) object$df.residual=rep(0,m)
#check whether the degrees of freedom match the number of endpoints
if (length(object$df.residual)!=m) stop("the dimension of
object$df.residual does not match the number of endpoints")
nu <- object$df.residual

```



```

##get the correlation matrix by the function getcor
getcor <- function(covmat){
  d <- dim(covmat)[1]
  diag(1/sqrt(diag(covmat)),d)%*%covmat%*%diag(1/sqrt(diag(covmat)),d)
}
if (k>1) {
  cormat <- array(0,dim=c(m,k,k))
  for (en in 1:m){
    cormat[en,,] <- getcor(object$vcov[en,,])
  }
} else {
  cormat <- array(1,dim=c(m,k,k))
}

```

```

#get cv by Dunnett's method for diff combinations of selected doses
#so that in testing each partition hypothesis the computer does not
#spend time calculating the same cv many times
#typically, one row of Dcomb looks like (1,0,1,0) meaning dose 1,3
#are selected
#All possible 0,1 for k doses, are  $2^k$ , except for (0,0,0,0) case.
getcvDun <- function(cormat,m,k,nu,alpha,alternative){
  Dcomb <- array(0,dim=c(m,2^k-1,k))
  #initialize critical value vector
  cvDun <- matrix(0,nrow=m,ncol=2^k-1)

```

```

if (alternative=="two.sided") {
  for (en in 1:m){
    for (i in 1:(2^k-1)){
      temp <- i
      for (j in 1:k){
        Dcomb[en,i,j] <- temp%%2^(k-j)
        temp <- i%%2^(k-j)
      }
      #get the index of columns that have value 1
      colindex <- which(Dcomb[en,i,]==1)
      #get the correlation matrix for selected doses
      cormati <- cormat[en,colindex,colindex]
      cvDun[en,i] <- qmvt(1-alpha,df=nu[en],tail="both.tails",
                          sigma=cormati,abseps=0.00001)$quantile
    }
  }
} else {
  for (en in 1:m){
    for (i in 1:(2^k-1)){
      temp <- i
      for (j in 1:k){
        Dcomb[en,i,j] <- temp%%2^(k-j)
        temp <- i%%2^(k-j)
      }
    }
  }
}

```

```

        #get the index of columns that have value 1
        colindex <- which(Dcomb[en,i,]==1)

        #get the correlation matrix for selected doses
        cormati <- cormat[en,colindex,colindex]

        cvDun[en,i] <- qmvt(1-alpha,df=nu[en],tail="lower.tail",
                           sigma=cormati,abseps=0.00001)$quantile
    }
}

}

list(Dcomb=Dcomb, cvDun=cvDun)
}

Dcomb.lst <- getcvDun(cormat,m,k,nu,alpha,alternative)
Dcomb <- Dcomb.lst$Dcomb
cvDun <- Dcomb.lst$cvDun
#print("done cvDun")

####get all possible partition hypotheses represented by rows of
####"combm" with each dose (column) ranging from 1 to m+1
getpartitionhyp <- function(m,k){
    #(m+1) possible vectors for each column
    v <- cbind(diag(m),rep(0,m))

    #number of partition hypotheses
    numpar <- (m+1)^k

    #the matrix recording all (m+1)^k possible choices,

```

```

#each row is a choice.

combmata <- matrix(0,numpar,k)

#getting all possible choices without doing all k for loops

#this can be done similarly in the above 0,1 case

combmata[,k] <- rep(seq(1,m+1,by=1),(m+1)^(k-1))

for (i in 1:(k-1)){

  combmata[,i] <- rep(sort(combmata[,k][1:(m+1)^(k-i+1)]),(m+1)^(i-1))

}

list(v=v,combmata=combmata)

}

combmata.lst <- getpartitionhyp(m,k)

combmata <- combmata.lst$combmata

v <- combmata.lst$v


#record rejection of each partition hypothesis

getreject <-

function(Tstat,v,combmata,cormata,cvDun,m,k,nu,alpha,alternative){

  numpar <- (m+1)^k

  reject <- rep(0,numpar)

  if (alternative=="two.sided") {

    #loop over each partition hypothesis

    for (i in 1:numpar){

      #each row of combata indicates the columns of v to select

```

```

parmat <- v[,combm[i,]]

#####calculate critical values for each part. hypo.

#initialize cv matrix

cvmat <- matrix(0,m,k)

#get Tstat associated with this particular hypothesis

Tstati <- Tstat*parmat

#number of test stat in each endpoint/row

### the apply function does not work when m=1

if (m==1) {

  r <- sum(parmat)

  csum <- parmat } else {

  r <- apply(parmat,1,sum)

  #which doses are involved over all endpoints

  csum <- apply(parmat,2,sum)

}

#try to get the corresponding Dunnet critical value

rowindex <- 0

for (j in 1:k){rowindex <- rowindex + csum[j]*2^(k-j)}

#initial cv candidate is calculated for doses involved over

#all endpoints, pretending they all come from endpoint 1

cv <- cvDun[1,rowindex]

#number of endpoints involved

tot <- sum(r>0)

#####if all test stat come from one endpoint

```

```

#####use Dunnett cv adjusting the number of doses involved
if (tot==1){cvmat <- parmat*cvDun[which(r>0),rowindex]} else
if (tot>1){

#####otherwise if test stats come from different endpoints
#####split the alpha levels by Bonferroni b/t endpoints
#setting cv recursively from 1st endpoint to last one
#For E1, set cv to Dunnett(ki) where ki is
#the number of doses involved in this partition hypothesis
#start from 1st endpoint
mi <- 0

#for each endpoint/row, define different cv
cvrow <- rep(0,m)
cumalphai <- 0
while (mi<m) {
  mi <- mi + 1

  #only for endpoints involved
  if (r[mi]>0) {

    #correlation matrix involved in each endpoint
    cormati <- cormat[mi,which(parmat[mi,]==1),
                                which(parmat[mi,]==1)]

    ##pretending cv is used for this endpoint

    ##get cumulative alpha levels spent including this
    checkalpha <-
    cumalphai + 1-pmvt(lower=rep(-cv,r[mi]),upper=

```

```

        rep(cv,r[mi]),df=nu[mi],sigma=cormati,
        abseps=0.00001)[1]

    #if alpha level has not been used up, set cv
    if (checkalpha<alpha) {
        cvrow[mi] <- cv
        cumalphai <- checkalpha} else {
        #o.w., set cv for this endpoint to Dun(alpha_left)
        #and cv for endpoints
        #of lower order to 100
        cvrow[mi] <-
        qmvt(1-(alpha-cumalphai),df=nu[mi],tail="both.tails",
        sigma=cormati,abseps=0.00001)$quantile
        if (mi<m) {cvrow[(mi+1):m] <- 100}
        #break out of the loop
        mi <- m}
    }
}

#matrix multiply vector is multiplication by each row
##this is ok b/c each row shares the same cv
cvmat <- parmat*cvrow
}

##if any Tstat exceeds the corresponding cv, then reject
if (sum(Tstat>cvmat)>0) reject[i] <- 1
}

```

```

} else {

  #loop over each partition hypothesis
  for (i in 1:numpar){

    #each row of combat indicates the columns of v to select
    parmat <- v[,combat[i,]]

    #####calculate critical values for each part. hypo.

    #initialize cv matrix
    cvmat <- matrix(0,m,k)

    #get Tstat associated with this particular hypothesis
    Tstati <- Tstat*parmat

    #number of test stat in each endpoint/row

    ###the apply function does not work when m=1
    if (m==1) {

      r <- sum(parmat)

      csum <- parmat } else

    if (k==1) {

      r <- parmat

      csum <- sum(parmat)

    } else {

      r <- apply(parmat,1,sum)

      #which doses are involved over all endpoints

      csum <- apply(parmat,2,sum)

    }

    #try to get the corresponding Dunnet critical value

```



```

rowindex <- 0
for (j in 1:k){rowindex <- rowindex + csum[j]*2^(k-j)}
#initial cv candidate is calculated for doses involved
#over all endpoints, pretending they all come from endpoint 1
cv <- cvDun[1,rowindex]
#number of endpoints involved
tot <- sum(r>0)
#####if all test stat come from one endpoint
#####use Dunnett cv adjusting the number of doses involved
if (tot==1){cvmat <- parmat*cvDun[which(r>0),rowindex]} else
if (tot>1){
  #####otherwise if test stats come from different endpoints
  #####split the alpha levels by Bonferroni b/t endpoints
  #setting cv recursively from 1st endpoint to last one
  #For E1, set cv to Dunnett(ki) where ki is
  #the number of doses involved in this partition hypothesis
  #start from 1st endpoint
  mi <- 0
  #for each endpoint/row, define different cv
  cvrow <- rep(0,m)
  cumalphai <- 0
  while (mi<m) {
    mi <- mi + 1
    #only for endpoints involved

```

```

if (r[mi]>0) {

  #correlation matrix involved in each endpoint
  cormati <- cormat[mi,which(parmat[mi,]==1),
                    which(parmat[mi,]==1)]

  ##pretending cv is used for this endpoint
  ##get cumulative alpha levels spent including this
  checkalpha <-
    cumalphai + 1-pmvt(lower=rep(-Inf,r[mi]),upper=
      rep(cv,r[mi]),df=nu[mi],sigma=cormati,
      abseps=0.00001)[1]

  #if alpha level has not been used up, set cv
  if (checkalpha<alpha) {
    cvrow[mi] <- cv
    cumalphai <- checkalpha} else {
    #o.w., set cv for this endpoint to Dun(alpha_left)
    #and cv for endpoints
    #of lower order to 100
    cvrow[mi] <-
      qmvt(1-(alpha-cumalphai),df=nu[mi],tail="lower.tail",
      sigma=cormati,abseps=0.00001)$quantile
    if (mi<m) {cvrow[(mi+1):m] <- 100}

    #break out of the loop
    mi <- m}
}

```

```

    }

    #matrix multiply vector is multiplication by each row
    ##this is ok b/c each row shares the same cv
    cvmat <- parmat*cvrow

  }

  ##if any Tstat exceeds the corresponding cv, then reject
  if (sum(Tstat>cvmat)>0) reject[i] <- 1
}

}

reject

}

reject <-

getreject(Tstat,v,combmata,cormat,cvDun,m,k,nu,alpha,alternative)

####collate results from all partition hypotheses
collate <- function(combmata,reject,m,k){
  ##initialize final rejection matrix as rejected
  numpar <- (m+1)^k
  freject <- matrix(1,m,k)
  for (i in 1:numpar){
    #if a partition hypothesis is not rejected, make final decision rules
    #on all corresponding individual hypotheses 0
    if (reject[i]==0){
      for (j in 1:k){

```

```

#corresponding endpoint for dose j this particular par. hyp
#is assoc.with
endptj <- combmat[i,j]
if (endptj<=m) {
    #lower endps for the same dose can't be rejected either
    if (freject[endptj,j]==1) freject[endptj:m,j] <- 0
}
}
}
}
freject
}
freject <- collate(combmat,reject,m,k)
cov.lst <- list(endpoint1=object$vcov[1,,])
if (m>1) {
    length(cov.lst) <- m
    for (i in 2:m) {
        cov.lst[[i]] <- object$vcov[i,,]
        names(cov.lst)[i] <- paste("endpoint",i,sep="")
    }
}

} else {

```

```

### for one endpoint case, do a different simpler version ###
#coef - vector, df.residual -- number,
#vcov -- k*k matrix, delta -- number
m <- 1
k <- length(object$coef)

#check whether the dim of delta is correct if not using the default
if(length(delta)!=1) stop("the dimentstion of delta does not match
the number of endpoints")

#check whether the dimension of vcov is correct
if (length(dim(object$vcov))!=2)
  {stop("the dimension fo object$vcov does not match number
of doses")} else
if (sum(dim(object$vcov)==c(k,k))!=2)
  {stop("the dimension fo object$vcov does not match number of doses")
}

Tstat <- (object$coef-delta)/sqrt(diag(object$vcov))

### if the option "alternative" indicates "less", take negative
### of the original Tstat
if (alternative=="less") Tstat <- -Tstat
if (alternative=="two.sided") Tstat <- abs(Tstat)

#get the degree of freedom
if (length(object)==2) object$df.residual=0

```

```

#check whether the degrees of freedom is correct

if (length(object$df.residual)!=1) stop("object$df.residual does
not match number of endpoints")

nu <- object$df.residual

##get the correlation matrix by the function getcor
getcor <- function(covmat){
  d <- dim(covmat)[1]
  diag(1/sqrt(diag(covmat)),d)%*%covmat%*%diag(1/sqrt(diag(covmat)),d)
}

cormat <- getcor(object$vcov)

getreject <- function(Tstat,cormat,k,nu,alpha,alternative){
  #each row of Dcomb represent a partition hypothesis,
  #"1" means the dose is invovled
  #the partition hypothesis with all "0"s were not included
  #since it does not need to be tested
  Dcomb <- matrix(0,nrow=2^k-1,ncol=k)
  #initialize critical value vector corresponding the rows of Dcomb
  cvDun <- rep(0,2^k-1)
  #record rejection of each partition hypothesis
  reject <- rep(0,2^k-1)
  if (alternative=="two.sided") {
    cv1d <- qt(1-alpha/2,df=nu)

```

```

#loop over each partition hypothesis
for (i in 1:(2^k-1)){
  temp <- i
  for (j in 1:k){
    Dcomb[i,j] <- temp%%2^(k-j)
    temp <- i%%2^(k-j)
  }
  parmat <- Dcomb[i,]
  ##if one dose selected, critical value is the same for sure
  if (sum(parmat)==1) {cvDun[i] <- cv1d} else {
    #get the index of columns that have value 1
    colindex <- which(parmat==1)
    #get the correlation matrix for selected doses
    cormati <- cormat[colindex,colindex]
    cvDun[i] <-
      qmvt(1-alpha,df=nu,tail="both.tails",sigma=cormati,
          abseps=0.00001)$quantile
  }
  #get the test stat
  Tstati <- Tstat*parmat
  cvmat <- cvDun[i]*parmat
  #if any Tstat exceeds the corresponding cv, then reject
  if (sum(Tstati>cvmat)>0) reject[i] <- 1
}

```

```

} else {

  cv1d <- qt(1-alpha,df=nu)

  #loop over each partition hypothesis
  for (i in 1:(2^k-1)){

    temp <- i

    for (j in 1:k){

      Dcomb[i,j] <- temp%%2^(k-j)

      temp <- i%%2^(k-j)

    }

    parmat <- Dcomb[i,]

    ##if one dose selected, critical value is the same for sure
    if (sum(parmat)==1) {cvDun[i] <- cv1d} else {

      #get the index of columns that have value 1
      colindex <- which(parmat==1)

      #get the correlation matrix for selected doses
      cormati <- cormat[colindex,colindex]

      cvDun[i] <-

      qmvt(1-alpha,df=nu,tail="lower.tail",sigma=cormati,

          abseps=0.00001)$quantile

    }

    #get the test stat

    Tstati <- Tstat*parmat

    cvmat <- cvDun[i]*parmat

    #if any Tstat exceeds the corresponding cv, then reject

```



```

        if (sum(Tstati>cvmat)>0) reject[i] <- 1
    }
}

list(partitionhyp=Dcomb,reject=reject)
}

reject.lst <- getreject(Tstat,cormat,k,nu,alpha,alternative)
reject <- reject.lst$reject
partitionhyp <- reject.lst$partitionhyp

####collate results from all partition hypotheses
collate <- function(Dcomb,reject,k){
    ##initialize final rejection matrix as rejected
    freject <- rep(1,k)
    for (i in 1:(2^k-1)){
        #if a partition hypothesis is not rejected,
        #make final decision rules on all
        #corresponding individual hypotheses 0
        if (reject[i]==0){
            colindex <- which(Dcomb[i,]==1)
            freject[colindex] <- rep(0,sum(Dcomb[i,]))
        }
    }
    freject
}

```

```

freject <- collate(partitionhyp,reject,k)
cov.lst <- object$vcov
}

####return the final results with some of the original inputs
list(final.reject=freject, coefficient=object$coef, covariance=cov.lst,
df=object$df.residual, FWER.control=alpha, delta=delta,
alternative.hypothesis=alternative)
}

```

BIBLIOGRAPHY

- [1] Lisa Arvanitis, Barbara Miller, and the Seroquel Trial 13 Study Group. Multiple fixed doses of "seroquel" (quetiapine) in patients with acute exacerbation of schizophrenia: A comparison with haloperidol and placebo. *Biological Psychiatry*, 42:233–246, 1997.
- [2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(2):289–300, 1995.
- [3] Yoav Benjamini and Wei Liu. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82:163–170, 1999.
- [4] Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300, 1982.
- [5] Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Statistical Science*, 11:283–315, 1996.
- [6] Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28:586–604, 2009.
- [7] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2001.
- [8] Christy Chuang-Stein, Paul Stryszak, Alex Dmitrienko, and Walter Offen. Challenge of multiple co-primary endpoints: new approaches. *Statistics in Medicine*, 26(6):1181–1192, 2007.
- [9] Alex Dmitrienko, Walt Offen, Ouhong Wang, and Dan Xiao. Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics*, 5:19–28, 2006.
- [10] Alex Dmitrienko, Walter W. Offen, and Peter H. Westfall. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine*, 22:2387–2400, 2003.

- [11] Charles W. Dunnett and Milton Sobel. A bivariate generalization of Student's t -distribution, with tables for certain special cases. *Biometrika*, 41:153–169, 1954.
- [12] Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50:1096–1121, 1955.
- [13] Helmut Finner and Marcus Roter. On the False Discovery Rate and expected Type I errors. *Biometrical Journal*, 43:985–1005, 2001.
- [14] Helmut Finner and Klaus Strassburger. The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, 30:1194–1213, 2002.
- [15] Frank Bretz, José C. Pinheiro, and Michael Branson. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61:738–748, 2005.
- [16] Frank Bretz, Jason Hsu, José C. Pinheiro, and Yi Liu. Dose finding - a challenge in statistics. *Biometrical Journal*, 50:480–504, 2008.
- [17] Alan Genz and Frank Bretz. Numerical computation of multivariate t -probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63:361–378, 1999.
- [18] Alexander Gordon, Galina Glazko, Xing Qiu, and Andrei Yakovlev. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Annals of Applied Statistics*, 1:179–190, 2007.
- [19] Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.
- [20] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.
- [21] Jason C. Hsu. The factor analytic approach to simultaneous inference in the general linear model. *Journal of Graphical and Computational Statistics*, 1:151–168, 1992.
- [22] Jason C. Hsu. *Multiple Comparisons: Theory and Methods*. Chapman & Hall, London, 1996.
- [23] Jason C. Hsu and Roger L. Berger. Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association*, 94:468–482, 1999.

- [24] Yifan Huang and Jason C. Hsu. Hochberg’s step-up method: Cutting corners off Holm’s step-down method. *Biometrika*, 94(4):965–975, 2007.
- [25] ICH E10. *Choice of Control Groups in Clinical Trials*. CPMP (Committee for Proprietary Medical Products), EMEA (The European Agency for the Evaluation of Medical Products), London, Draft ICH (International Conference on Harmonisation) Guideline edition, May 1999. <http://www.ifpma.org/pdfifpma/e10.pdf>.
- [26] Mark E. Johnson. *Multivariate Statistical Simulation*. Wiley, 1987.
- [27] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 5 edition, 2002.
- [28] Samuel Karlin and Yosef Rinott. Total positivity properties of absolute value multinormal variables with applications to confidence interval estimation and related probabilistic inequalities. *Annals of Statistics*, 9:1035–1049, 1981.
- [29] Shili Lin, James Rogers, and Jason C. Hsu. A confidence set approach for finding tightly linked genomic regions. *American Journal of Human Genetics*, 68:1219–1228, 2001.
- [30] Yi Liu, Jason C. Hsu, and Stephen Ruberg. Partition testing in dose-response studies with multiple endpoints. *Pharmaceutical Statistics*, 6:181–192, 2007.
- [31] Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660, 1976.
- [32] Walter Offen, Christy Chuang-Stein, Alex Dmitrienko, Gary Littman, Jeff Maca, Laura Meyerson, Robb Muirhead, Paul Stryszak, Alex Boddy, Kun Chen, Kati Copley-Merriman, Willard Dere, Sam Givens, David Hall, David Henry, Joseph D. Jackson, Alok Krishen, Thomas Liu, Steve Ryder, A. J. Sankoh, Julia Wang, and Chyon-Hwa Yeh. Multiple co-primary endpoints: Medical and statistical solutions a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of America. *Drug Information Journal*, 41:31–46, 2007.
- [33] Walter W. Offen and Jeffrey D. Helterbrand. Multiple comparison adjustments when two or more co-primary endpoints must all be statistically significant. In *Proceedings of the Annual Meeting of the American Statistical Association*, Chicago, IL, August 1986.
- [34] Art B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society B*, 67:411–426, 2005.
- [35] Stephen Ruberg and Victoria Cairns. Providing evidence of efficacy for a new drug. *Statistics in Medicine*, 17(15-16):1813–1823, August 1998.

- [36] Stephen J. Ruberg. Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association*, 84:816–822, 1989.
- [37] Stephen J. Ruberg. Dose response studies. I. some design considerations. *Journal of Biopharmaceutical Statistics*, 5(1):1–14, 1995.
- [38] Stephen J. Ruberg. Dose response studies. II. Analysis and interpretation. *Journal of Biopharmaceutical Statistics*, 5:15–42, 1995.
- [39] Mohommed Siddiqui. A bivariate t distribution. *The Annals of Mathematical Statistics*, 38(1):162–166, 1967.
- [40] Robert J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.
- [41] Gunnar Stefansson, W.C. Kim, and Jason C. Hsu. On confidence sets in multiple comparisons. In Shanti S. Gupta and James O. Berger, editors, *Statistical Decision Theory and Related Topics IV*, volume 2, pages 89–104. Springer-Verlag, New York, 1988.
- [42] Ajit C. Tamhane, Charles W. Dunnett, John W. Green, and Jefferey D. Wetherington. Multiple test procedures for identifying the maximum safe dose. *Journal of the American Statistical Association*, 96:835–843, 2001.
- [43] Ajit C. Tamhane and Brent R. Logan. Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *Journal of the American Statistical Association*, 97(457):293–301, 2002.
- [44] Robert J. Temple. Effect size – can the effect be too small. Presentation at FDA Center for Drug Evaluation and Research, Cardiovascular and Renal Advisory Committee meeting, 2006. <http://www.fda.gov/ohrms/dockets/ac/06/slides/2006-4215S1-02-FDA.ppt>.
- [45] Brian L. Wiens and Alex Dmitrienko. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*, 15(6):929–942, 2005.
- [46] Haiyan Xu, Isaac Nuamah, Jingyi Liu, Pilar Lim, and Allan Sampson. A Dunnett-Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints is available at <http://www3.interscience.wiley.com/cgi-bin/fulltext/121552605/pdfstart>. *Pharmaceutical Statistics*, 2008.