Covering Principle: A New Approach to Address Multiplicity in Hypothesis Testing

Huajiang Li[1] and Hong Zhou[2*]
[1]Avanir Pharmaceuticals, Inc. 30 Enterprise, Aliso Viejo, CA, 92656
[2]Arkansas State University, P. O. Box 70, State University, AR 72467

Abstract
The closure and the partitioning principles have been used to build various multiple testing procedures in the past three decades. The essence of these two principles is based on parameter space partitioning. In this article, we propose a novel approach coined the covering principle from the perspective of rejection region coverage in the sample space. The covering principle divides the whole family of null hypotheses into a few overlapped subsets when there is a priority of making decisions for hypothesis testing. We have proven that the multiple testing procedure constructed by the covering principle strongly controls the familywise error rate as long as the multiple tests for each subset strongly control the type I error. The covering principle are applied to two real clinical trials to illustrate how to construct multiple testing procedures. It is also shown that the proposed method can reject more null hypotheses and gain some power in some scenarios compared to the graphical approaches.

Key Words: Familywise error rate, multiple hypotheses testing, closed test principle, partitioning principle, covering principle, gate-keeping

1. Introduction

The key issue in the multiple hypotheses testing is to strongly control the familywise error rate (Hochberg & Tamhane, 1987). Two important principles: the closure principle (Marcus et al. 1976) and the partitioning principle (Finner & Strassburger, 2002; Sonnemann, 2008), are widely used to construct various multiple test procedures that can strongly control the familywise error rate. A strong control of the familywise error rate for multiple hypotheses testing procedures is mandated by the regulatory agencies in all confirmatory clinical trials (Food & Drug Administration, 2002; Committee for Proprietary Medicinal Products, 2002).

First, a common way to handle the multiplicity issue is to cut the spending of the overall significance level α as it does in Bonferroni procedure and its modifications (Holm, 1979; Hochberg, 1988; Hommel, 1988; Li et al., 2017). Second, when the multiple study objectives exhibit a hierarchical structure which is usually divided into primary, secondary objectives in the clinical trials, the gate-keeping procedures are used to deal with the multiplicity issue (Dmitrienko et al. 2003; Dmitrienko & Tamhane 2007; Dmitrienko, Tamhane, Liu & Wiens 2008; Dmitrienko, Tamhane & Wiens 2008; Dmitrienko & Tamhane 2013). Recently, Bretz et al. (2009) introduced a graphical approach to construct multiple hypotheses testing procedures. Many multiple hypotheses testing procedures, such as Holm's (Holm 1979), the fixed sequence (Maurer et al. 1995; Westfall & Krishen 2001) and the fallback (Wiens 2003; Wiens & Dmitrienko 2005) procedures as well as gate-keeping procedures can be illustrated by the weighted and directed graphs. The graphical approach facilitates easy communication with medical doctors and clinicians and becomes very popular in clinical trials with the help of "gMCP" package in R programming language. (Bretz,

*Corresponding author: email: hzhou@astate.edu, Phone: +1-870-972-3090, Fax: +1-870-972-3950

Posch, Glimm, Klinglmueller, Maurer & Rohmeyer 2011; Bretz, Maurer & Hommel 2011; Bretz et al. 2014).

In this article, we introduce a novel principle termed the covering principle for the construction of the multiple testing procedures from the perspective of the sample space. It is different from both the closed test and the partitioning principles which work in the parameter space. The covering principle analyzes the rejection regions in the sample space based on the priorities of the decisions for testing the null hypotheses and divides the whole family of null hypotheses into a few overlapped sub-families, for which any multiple testing procedure can be used. The special contributions of this article include: (a) proposing a novel method to solve multiple hypotheses testing problems in the sample space; (b) proving a theoretical result that a multiple hypotheses testing procedure based on the proposed covering principle strongly control the familywise error rate for the whole family of hypotheses if each multiple hypotheses testing procedure on each subset can strongly control the familywise error rate.

Before we formally introduce the covering principle in Section 2, let us consider a simple motivating problem to understand the rationale of the proposed method.

Example 1: Suppose that there are three families F_1, F_2, F_3 in a simple serial gatekeeping problem and each family has only one null hypothesis, denoted as H_1, H_2 and H_3, respectively, with the corresponding rejection regions R_1, R_2 and R_3 in the sample space. Assume that the order of testing hypotheses is H_1 the first, H_2 the second, and H_3 the last, i.e. $H_1 \rightarrow H_2 \rightarrow H_3$. The sequential order of testing hypotheses is illustrated as a decision flowchart in Figure 1(a). The hypothesis H_1 is called the "gate-keeper" of H_2, meaning that H_2 will not be tested unless H_1 has been tested and rejected. Let us think reversely, if we know H_2 is rejected, then it implies that H_1 has already been tested and rejected because of the order of hypothesis testing and decision making.

Logically speaking, the rejection of H_2 indicates that H_1 has been tested and rejected. That means the rejection region R_2 must be within R_1 in the sample space. For the similar reason, the rejection region R_3 of the hypothesis H_3 also must be within R_2. Therefore, from the analysis of the order in testing null hypotheses and the decision making process, the relationship among rejection regions is depicted as $R_3 \subseteq R_2 \subseteq R_1$ and visualized by a Venn diagram in Figure 1(b). Imagine that we have an observed sample at hand. If the value of the test statistic from the sample falls in the rejection region R_3, then H_3 is rejected. Since $R_3 \subseteq R_2 \subseteq R_1$, consequently, both H_1 and H_2 must be also rejected. It is worth noting that the relationship of rejection regions is an abstract logical relation. H_1 could be related to an endpoint with a continuous measurement while H_2 might correspond to a binary outcome and H_3 could be associated with a time-to-event endpoint. We will re-visit this example after the covering principle is introduced in Section 2.

The article is organized as follows. Section 2 presents mathematical notations and introduces the theorem of the covering principle formally. In Section 3, we apply the covering principle to two real clinical trials to illustrate the use of the proposed principle. Section 4 is the discussion and closing remarks of using the covering principle. Finally, we prove that the multiple hypotheses testing procedures based on the covering principle strongly control the familywise error rate for the whole family in the Appendix.

2. The Covering Principle

Denote $N = \{1, 2, \ldots, n\}$ as the index set of a family of n null hypotheses $H_1, H_2, \ldots, H_n$ with the corresponding test functions $\phi = \{\phi_1, \phi_2, \ldots, \phi_n\}$ and rejection regions $R_1, R_2, \ldots,$

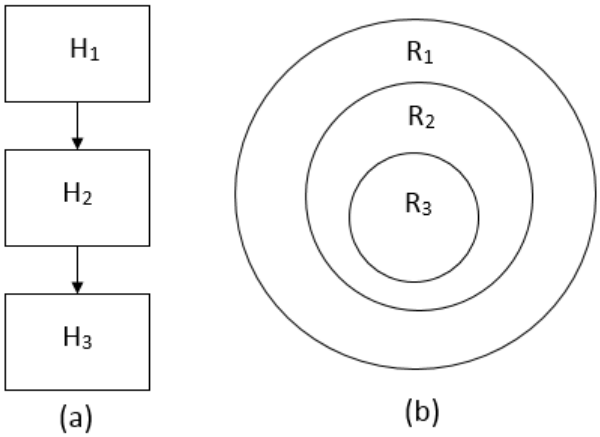

Figure 1: (a) Decision Flowchart (b) Venn Diagram of Rejection Regions of Example 1

R_n. Each ϕ_i $(i = 1, 2, \ldots, n)$ is an elementary test function, where

$$\phi_i = \begin{cases} 1, & \text{if } H_i \text{ is rejected} \\ 0, & \text{if } H_i \text{ is accepted.} \end{cases}$$

For $\emptyset \neq S \subseteq N$, let $\Phi_\alpha(S)$ denote the set of all α-level multiple tests for the family of null hypotheses with an index set S, where $0 < \alpha < 1$. If $\phi = \{\phi_i : i \in S\} \in \Phi_\alpha(S)$, then it indicates the multiple test ϕ strongly controls the familywise error rate on S at the significance level α. For a group of elementary test functions $\phi_i, i \in S$, define

$$\min_{i \in S} \phi_i = \begin{cases} 1, & \text{if } \forall i \in S, \phi_i = 1 \\ 0, & \text{otherwise.} \end{cases}$$

and

$$\max_{i \in S} \phi_i = \begin{cases} 1, & \text{if } \exists i \in S, \phi_i = 1 \\ 0, & \text{otherwise.} \end{cases}$$

For any two elementary test functions ϕ_1 and ϕ_2, denote $\phi_1 \leq \phi_2$ if $\{\phi_1 = 1\}$ implies $\{\phi_2 = 1\}$. Equivalently say, the rejection of H_1 implies the rejection of H_2. In terms of the rejection regions, it means $R_1 \subseteq R_2$.

Theorem 1 (Covering principle)**.** *Suppose* $\emptyset \neq I \subset N$, $\emptyset \neq J \subset N$, $I \cap J = \emptyset$, *and* $\bigcup_{i \in I} R_i \subseteq \bigcup_{j \in J} R_j$. *Denote*

$$\phi^j = \{\phi_i^j : i \in N \setminus j\}, \forall j \in J. \tag{1}$$

$$\phi^I = \{\phi_i^I : i \in N \setminus I\}. \tag{2}$$

where ϕ^j *denote an arbitrary multiple test which does not include the test function of the* j*th hypothesis* H_j, $j \in J$. *Each* ϕ^j *consists of* $n - 1$ *elementary test functions* $\phi_i^j, i \in N \setminus j$. *Similarly,* ϕ^I *denote a multiple test which does not include those test functions of hypotheses whose indices are in* I. ϕ^I *consists of* $n - |I|$ *elementary test functions* $\phi_i^I, i \in N \setminus I$. *Define a test function* ψ_i *as follows:*

$$\psi_i = \begin{cases} \min(\min_{j \in J} \phi_i^j, \phi_i^I), & \text{if } i \in N \setminus I \\ \min(\min_{j \in J} \phi_i^j, \max_{j \in J} \psi_j), & \text{if } i \in I. \end{cases} \tag{3}$$

If $\phi^j \in \Phi_\alpha(N \setminus j), \forall j \in J$ *and* $\phi^I \in \Phi_\alpha(N \setminus I)$, *then* $\{\psi_i : i \in N\} \in \Phi_\alpha(N)$.

The following is an explanation of Theorem 1. Suppose that there exist two nonempty index sets $I \subset N$ and $J \subset N$, $I \cap J = \emptyset$, i.e. these two sets of hypotheses $\{H_j, j \in J\}$, and $\{H_i, i \in I\}$, are not overlapped. Furthermore, there are orders when these hypotheses are tested. In order for H_i, to be tested, at least one of the hypothesis H_j, must be tested and rejected first. For example, the hypotheses H_j could be related to the primary endpoints and H_i could be related to the secondary endpoints in clinical trials. For parallel gate-keeping, if a null hypothesis on a secondary endpoint is rejected, then at least one hypothesis on one of primary endpoints already has been rejected. If this is the case, we say the set of hypotheses $\{H_j, j \in J\}$, dominates the set of hypotheses $\{H_i, i \in I\}$. In other words, there is a dominance relationship between two sets of hypotheses $\{H_i, i \in I\}$, and $\{H_j, j \in J\}$. From the perspective of hypotheses testing in the sample space, the dominance relationship between two sets of hypotheses: $\{H_i, i \in I\}$ and $\{H_j, j \in J\}$ can be defined by the logical relationship among their rejection regions: $\bigcup_{i \in I} R_i \subseteq \bigcup_{j \in J} R_j$. Furthermore, $H_j, j \in J$ is called a dominant hypothesis and $H_i, i \in I$ is a dominated hypothesis.

It may seem that the definition of ψ_i in Equation (3) is circular. In fact, domains for the index i are mutually exclusive. For the first part of the definition of the test function ψ_i, a hypothesis $H_i, i \in N \setminus I$, could be either one of hypotheses $H_j, j \in J$, which dominate $H_i, i \in I$, or one of non-constraint hypotheses. A non-constraint hypothesis is the one that has no dominance relationship with other hypotheses. The first part of ψ_i defines a test function to reject any hypothesis H_i whose index is not in I. That is, $H_i, i \in N \setminus I$, will be rejected if it is rejected in all subsets which contain it.

The second part of ψ_i in Equation (3) defines a test function for those hypotheses whose indices are within I. The set of hypotheses $\{H_i, i \in I\}$, are dominated by the set of hypotheses $\{H_j, j \in J\}$. In order for a hypothesis $H_i, i \in I$, to be rejected, not only at least one $H_j, j \in J$, must be rejected first, but also $H_i, i \in I$, must be rejected in all subsets which contain it.

Then, the covering principle in Theorem 1 states that the original whole family of n null hypotheses can be decomposed into $|J| + 1$ subsets with index sets $N \setminus j, \forall j \in J$ and $N \setminus I$. In other words, the original multiple testing problem on the family of n null hypotheses with the index set N can be divided into $|J| + 1$ multiple testing problems. The corresponding multiple tests are ϕ^I with the index set $N \setminus I$ and ϕ^j with index sets $N \setminus j, \forall j \in J$. Each subset has fewer null hypotheses than the original family and can be tested using any multiple testing procedure. The multiple testing procedure built on this divide-and-conquer strategy strongly controls the familywise error rate for the whole family at the significance level α if the multiple tests ϕ^I and ϕ^j can control their familywise error rate at the significance level α for their corresponding subsets. Finally, the decision rule for each individual hypothesis can be reached by summarizing the results as follows:

Step 1. A dominant or a non-constraint hypothesis, $H_i, i \in N \setminus I$, will be rejected if H_i is rejected in all decomposed subsets in which H_i is contained.

*Step 2.*A dominated hypothesis, $H_i, i \in I$, will be rejected if at least one of its dominant hypotheses $H_j, j \in J$, is rejected first. In addition, H_i must be also rejected in all subsets in which H_i is contained.

Let us continue Example 1 to illustrate how to use Theorem 1 to decompose a family of hypotheses into subsets. In this example, there are three null hypotheses, denoted by their indices $N = \{1, 2, 3\}$. Based on the sequential order of testing null hypotheses, $H_1 \rightarrow H_2 \rightarrow H_3$, the rejection regions exhibit the following coverage relations: $R_3 \subseteq R_2$, $R_2 \subseteq R_1$, $R_3 \subseteq R_1$ as shown by a Venn diagram in Figure 1(b). H_1 dominates H_2 and H_2 dominates H_3. Consequently, H_1 dominates H_3 through H_2 indirectly. There are three dominance relations among these hypotheses.

Now let us apply the covering principle to this example to build a multiple hypotheses

testing procedure. First, according to the relationship: $R_3 \subseteq R_2$, $J = \{2\}$, $I = \{3\}$, the whole family of null hypotheses $\{H_1, H_2, H_3\}$ can be divided into two subsets: $N \setminus I = \{1, 2\}$ and $N \setminus j = \{1, 3\}$. In addition, because $R_2 \subseteq R_1$, the subset $\{H_1, H_2\}$ can be further decomposed into two subsets $\{H_1\}$ and $\{H_2\}$ by using the covering principle again. Similarly, the subset $\{H_1, H_3\}$ is decomposed into $\{H_1\}$ and $\{H_3\}$ by using $R_3 \subseteq R_1$. Combining the results, we have three distinct subsets: $\{H_1\}$, $\{H_2\}$ and $\{H_3\}$. In this example readers may see that the covering principle is used iteratively as long as there is a dominance relation in a subset. The dimension of the original set of null hypotheses is reduced to 1 from 3.

The decision rule for each individual hypothesis is as follows. The null hypothesis H_1 will be rejected if it is tested and rejected at the α-level. In order for H_2 to be rejected, the null hypothesis H_1 must be rejected first, followed by H_2 being rejected in the test of H_2 at the α-level. In order for H_3 to be rejected, both null hypothesis H_1 and H_2 must be rejected first and H_3 is also rejected in the test of H_3 at the α-level. Interestingly, the multiple testing procedure constructed by using our method is equivalent to the fixed sequence procedure (Maurer et al. 1995; Westfall & Krishen 2001), which is well-known for strongly controlling the familywise error rate. In general, it can be extended in the case of n null hypotheses.

The novelty of the covering principle is that it works on the coverage relation of rejection regions in the sample space. The covering principle performs a sample space analysis using the union of rejection regions in contrast to the closed test principle using the intersection of hypotheses in the parameter space. By using the dominance relationship, the whole family of null hypotheses is decomposed into a few overlapped subsets and the decomposition reduces the dimension of the multiple testing problem. After the decomposition, any α-level multiple hypotheses testing procedure can be used to test hypotheses in these subsets. The decision on an individual hypothesis is made by consolidating the testing results from each subset. Simply speaking, in order for an individual hypothesis $H_i, i \in N$, to be rejected, not only one of its precedent and dominant hypotheses in the hierarchy of the hypotheses must be rejected first, but it must also be rejected in all subsets which contain H_i. The covering principle extends the closed test principle to a family of hypotheses with the priority of importance when making decisions.

3. Applications to Two Real Clinical Trials

3.1 Example from Gate-Keeping Problem

Example 2: For illustration purposes, let us consider a simple parallel gate-keeping problem. Cummings et al. (1999) studied breast cancers in post-menopausal women in a clinical trial. The study has two primary endpoints: the incidence of vertebral fractures and the incidence of breast cancer, and one secondary endpoint: the incidence of non-vertebral fractures. Each primary endpoint will result in an independent regulatory claim if the treatment effect on any one of two primary endpoints is effective. Denote H_1 and H_2 as the null hypotheses on the two primary endpoints, respectively, and H_3 on the secondary endpoint. The test on the secondary endpoint can only be carried out if at least one of the null hypotheses $\{H_1, H_2\}$ relating to two primary endpoints is rejected. In terms of gatekeeping approach, the two primary endpoints serve as the gatekeepers. The order of testing hypotheses can be displayed as a decision flowchart in Figure 2(a).

Let R_1, R_2 and R_3 denote three rejection regions in the sample space corresponding to tests on null hypotheses H_1, H_2 and H_3, respectively. Because the hypothesis on the secondary endpoint can not be tested and rejected unless one of two hypotheses on two

primary endpoints has been rejected, so logically speaking, the rejection of H_3 implies that at least one of H_1 and H_2 has already been rejected. In terms of the rejection region in the sample space, R_3 must be covered by the union of rejection regions R_1 and R_2. It can be denoted as $R_3 \subseteq R_1 \cup R_2$. Imagine that we have a sample at hand. If the value of the test statistic from this sample falls in the rejection region R_3, then H_3 is rejected. Since $R_3 \subseteq R_1 \cup R_2$, consequently, either H_1 or H_2, or both must be also rejected depending on where the value of the test statistic falls. Following the analysis of the decision flowchart, the coverage relation of rejection regions among R_1, R_2 and R_3 can be visualized by a Venn diagram as shown in Figure 2(b).

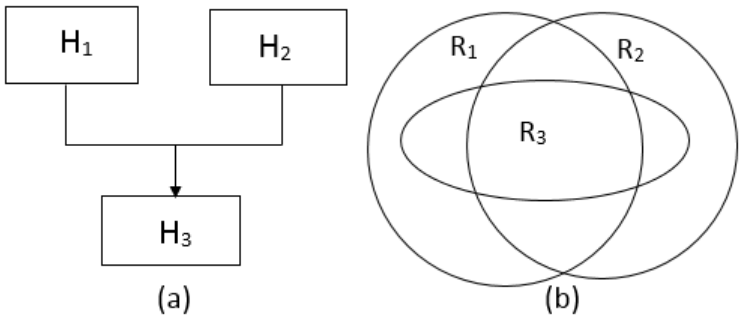

Figure 2: (a) Decision Flowchart (b) Venn Diagram of Rejection Regions of Example 2

H_1 and H_2 dominate H_3, i.e. $R_3 \subseteq R_1 \cup R_2$, so $J = \{1, 2\}$ and $I = \{3\}$. According to the covering principle, the whole family of 3 null hypotheses $\{H_1, H_2, H_3\}$ are then decomposed into three subsets: $N \setminus I = \{1, 2\}$, $N \setminus 1 = \{2, 3\}$ and $N \setminus 2 = \{1, 3\}$. The null hypothesis H_1 will be rejected if it is rejected in all subsets which include H_1: $\{H_1, H_2\}$ and $\{H_1, H_3\}$; similarly for H_2. However, in order for H_3 to be rejected, not only must either H_1 or H_2 be rejected first because H_3 is dominated by H_1 and H_2, but H_3 is also rejected in all subsets including H_3: $\{H_2, H_3\}$ and $\{H_1, H_3\}$. Suppose that observed p-values for three hypotheses: $p_1 = 0.024$, $p_2 = 0.06$, $p_3 = 0.003$ and the significance level $\alpha = 0.05$, Holm's procedure is used for testing all subsets for the sake of simplicity. Then, H_1 will be rejected in both subsets: $\{H_1, H_2\}$ and $\{H_1, H_3\}$; similarly for H_3 in subsets: $\{H_1, H_3\}$ and $\{H_2, H_3\}$; But H_2 can not be rejected in both subsets: $\{H_1, H_2\}$ and $\{H_2, H_3\}$. Finally, we consolidate the results from each subset and make conclusions on an individual hypothesis as follows: (a) The hypothesis H_1 on the primary endpoint is rejected because it has been rejected in both subsets: $\{H_1, H_2\}$ and $\{H_1, H_3\}$; (b) The hypothesis H_2 on the primary endpoint can not been rejected; (c) The hypothesis H_3 on the secondary endpoint is also rejected because not only has it been rejected in both subsets: $\{H_1, H_3\}$ and $\{H_2, H_3\}$, but also the hypothesis H_1 on the primary endpoint has been rejected.

3.2 Example from Graphical Approach

Example 3: Bretz et al. (2009) demonstrated a case study for multiple sclerosis by using the graphical approach. This trial compares two dose levels of a new treatment with a control for three hierarchical endpoints. Denote six hypotheses on different dose levels and different endpoints as H_{ij}, where $i = 1$ for the high dose level and $i = 2$ for the low dose level, and $j = 1, 2, 3$ for the primary, secondary and tertiary endpoint, respectively. The primary endpoint is the annualized relapse rate and the secondary endpoint is the number of lesions in the brain. The tertiary endpoint is the disability progression. A few graphical strategies were discussed in their paper. Only two main common graphical strategies are shown here in Figure 3. The left graph in Figure 3 divided six hypotheses into two groups

according to two dose levels: $\{H_{11}, H_{12}, H_{13}\}$ and $\{H_{21}, H_{22}, H_{23}\}$. The endpoints in each group were tested in a fixed sequence at $\alpha/2$ level. If three hypotheses in one group were rejected, then the local significance level can be transfered to the tests in the other group, meaning that the fixed sequence tests for the other group can be carried out at the α-level. For the left graph, you can start with testing the primary endpoint for either dose level.

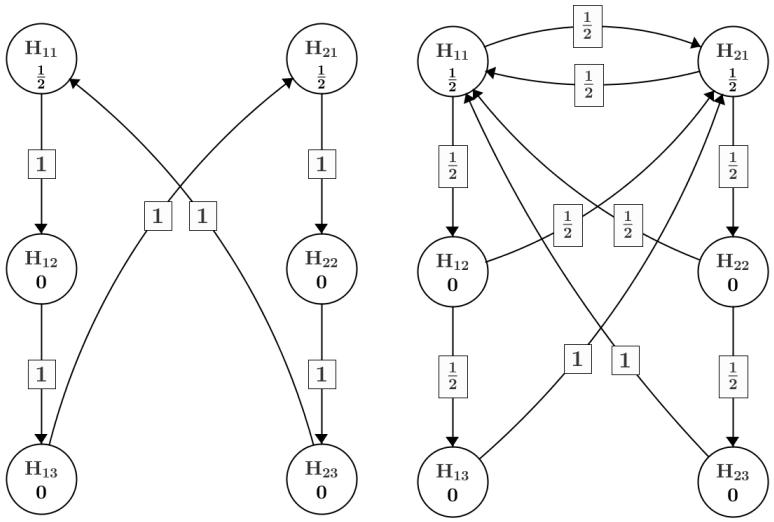

Figure 3: Visualization of two different Graphical Strategies of Example 3

The right graph in Figure 3 puts more weight on the precedent hypothesis in the hierarchy of hypotheses. The local significance level after a rejection of one hypothesis was split into two parts. One part was distributed to the precedent hypothesis in the other group which has not been rejected. The second part was re-allocated to the subsequent hypothesis in the same group.

For illustration purposes, consider Scenario 1: assume that the significance level α for the test is 0.05 and observed p-values are $p_{11} = p_{12} = p_{13} = 0.024$, $p_{21} = p_{22} = p_{23} = 0.04$. According to the left graph, all six hypotheses will be rejected. However, if the right graph is used, only H_{11} can be rejected.

It is apparent that different testing graphs could lead to different testing results. If one has sufficient prior knowledge on the testing problem, then the graphical approach can use such knowledge to set its weights and initial α allocations. However, such types of prior knowledge are not always available. Therefore, the determination of initial allocations of significance level and transition weights is a challenging problem. The proposed covering principle may help users avoid or at least alleviate such problem.

Now let us apply the covering principle to Example 3.

Step 1. Construct the coverage relations among rejection regions. Based on the dominance relations among hypotheses on primary, secondary, and tertiary endpoints and the decision-making flowchart in Figure 4(a), the coverage relations of rejection regions are constructed: $R_{13} \subseteq R_{12} \subseteq R_{11}$, $R_{23} \subseteq R_{22} \subseteq R_{21}$ as shown in Figure 4(b).

Step 2. Decompose the family of 6 hypotheses into 9 subsets as follows:

The family of six null hypotheses $\{H_{11}, H_{12}, H_{13}, H_{21}, H_{22}, H_{23}\}$ is decomposed into nine subsets containing only two hypotheses each: $\{H_{11}, H_{21}\}$, $\{H_{11}, H_{22}\}$, $\{H_{11}, H_{23}\}$, $\{H_{12}, H_{21}\}$, $\{H_{12}, H_{22}\}$, $\{H_{12}, H_{23}\}$, $\{H_{13}, H_{21}\}$, $\{H_{13}, H_{22}\}$, and $\{H_{13}, H_{23}\}$. The original 6-dimension multiple testing problem is now reduced into nine 2-dimension multiple testing problems.

Step 3. Test each hypothesis H_i in each subset in which H_i is contained. For purposes of simplicity and illustration, Holm's procedure are used for all nine subsets. All six hypotheses are rejected in their respective subsets according to the p-values given in Scenario 1.

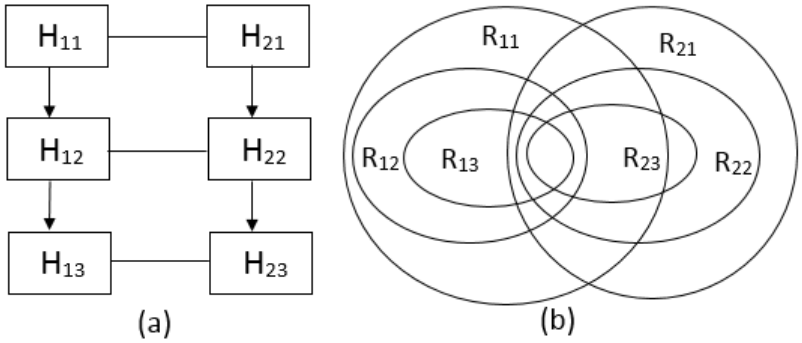

Figure 4: (a) Decision Flowchart (b) Venn Diagram of Rejection Regions of Example 3

Step 4. Consolidate results from each subset to make conclusions on each hypothesis. The decision rule for each individual hypothesis is as follows:

(*4a*) H_{11} is rejected since it is rejected in the decomposed subsets containing H_{11}: $\{H_{11}, H_{21}\}$, $\{H_{11}, H_{22}\}$, and $\{H_{11}, H_{23}\}$. H_{21} is also rejected due to similar reason.

(*4b*) H_{12} is rejected because the dominant hypothesis H_{11} in its upper level has been rejected in (*4a*) and H_{12} is also rejected in the decomposed subsets containing H_{12}: $\{H_{12}, H_{22}\}$, $\{H_{12}, H_{23}\}$, and $\{H_{12}, H_{21}\}$. H_{22} is rejected due to similar reason.

(*4c*) H_{13} is rejected because both dominant hypotheses H_{11} and H_{12} its upper levels have been rejected already in (*4a*) and (*4b*), and H_{13} is also rejected in the decomposed subsets containing H_{13}: $\{H_{13}, H_{23}\}$, $\{H_{13}, H_{21}\}$, and $\{H_{13}, H_{22}\}$. H_{23} is rejected due to similar reason.

In conclusion, all six hypotheses in Scenario 1 are rejected by the multiple testing procedure built on the covering principle.

Now let us consider Scenario 2. Suppose observed p-values: $p_{11} = 0.0374$, $p_{12} = 0.024$, $p_{13} = 0.024$, $p_{21} = 0.024$, $p_{22} = 0.04$, $p_{23} = 0.024$. According to the graphical procedures in Figure 3, the left graph only rejects H_{21}; the right graph rejects H_{11}, H_{12} and H_{21}.

The covering principle rejects all six hypotheses if Holm's procedure is used in three subsets: $\{H_{11}, H_{21}\}$, $\{H_{12}, H_{22}\}$, $\{H_{13}, H_{23}\}$ and the fixed sequence procedure is used in the remaining six subsets: $\{H_{11}, H_{22}\}$, $\{H_{11}, H_{23}\}$, $\{H_{12}, H_{21}\}$, $\{H_{12}, H_{23}\}$, $\{H_{13}, H_{21}\}$, $\{H_{13}, H_{22}\}$.

Readers might have observed that different multiple testing procedures could be employed to the different subsets. In Scenario 1, Holm's procedure is utilized for all nine subsets universally for simplicity. However, in Scenario 2, Holm's procedure is used for the subset containing hypotheses from same tier. The fixed sequence procedure is used for subsets with hypotheses from different tiers. The different choices between multiple hypotheses testing procedures for different subsets leave the flexibility for practitioners to develop a tailored multiple testing procedure. This type of choices may also increase the power of tests in practice.

4. Discussion and Closing Remarks

The covering principle provides a new approach and may play an important role in solving multiple hypotheses testing problems in which there are constraints/orders among null hypotheses. The covering principle can be viewed as an extension and generalization to the popular closed test principle and the partitioning principle in some sense. The constraints or pre-specified orders of testing hypotheses and making decisions can be described by the dominance relations between hypotheses and specified by the coverage relations among rejection regions in the sample space. The dominance/coverage relations are used to decompose a family of n null hypotheses into a few overlapped subsets. The process of decomposition continues in a subset as long as there exists a dominance relation within the subset. After the decomposition, the closed test principle and the partitioning principle can be applied to these decomposed subsets.

The merit of the decomposition is twofold. First, it reduces the dimension of the multiple testing problems since each subset has fewer hypotheses. Testing multiple hypotheses is easier in lower dimensions. Second, each decomposed subset forms a "new family" of null hypotheses. Any α-level multiple testing procedure can be used for testing hypotheses in each subset. Users may choose different multiple test procedures for different decomposed subsets according to different situations. In Example 3, Holm's procedure is chosen for subsets with hypotheses from same tier, whereas a fixed sequence procedure is used for subsets with hypotheses from different tiers. The choice of a multiple testing procedure for a particular subset might depend on the circumstances in practice. In fact, a Hochberg's procedure (1988) could be used for the subset $\{H_{11}, H_{21}\}$ if there is a positive correlation between the test statistics of the two hypotheses. The flexibility of choosing different multiple testing procedures for different subsets may facilitate users in practice and improve testing power in some situations.

5. Appendix

5.1 Proof of Theorem 1

By the results (Finner & Strassburger (2002),Eq.(2.2) on p. 1197; Sonnemann (2008), Eq.(3.4) on p. 645), we have $\{\psi_i : i \in N\} \in \Phi_\alpha(N)$ if and only if $\forall \emptyset \neq S \subseteq N$, $\forall \theta \in \cap_{i\in S} H_i$, $P_\theta(\max_{i\in S} \psi_i = 1) \leq \alpha$. For $\forall \emptyset \neq S \subseteq N$, consider two cases of the relationship between S and J.

Case I: $J \not\subseteq S$. There exists a $j_0 \in J$ such that $j_0 \notin S$, then $S \subseteq N \setminus j_0$. By the definition of ψ_i in equation (3), $\psi_i \leq \min_{j\in J} \phi_i^j$, $\forall i \in N$, we have $\psi_i \leq \phi_i^{j_0}$, $j_0 \neq i \in N$. Therefore, $\max_{i\in S} \psi_i \leq \max_{i\in S} \phi_i^{j_0}$. By the assumption of Theorem 1, $\{\phi_i^{j_0} : i \in N \setminus j_0\} \in \Phi_\alpha(N \setminus j_0)$, and $S \subseteq N \setminus j_0$, hence $\forall \theta \in \cap_{i\in S} H_i$, $P_\theta(\max_{i\in S} \phi_i^{j_0} = 1) \leq \alpha$. Therefore, $P_\theta(\max_{i\in S} \psi_i = 1) \leq P_\theta(\max_{i\in S} \phi_i^{j_0} = 1) \leq \alpha$.

Case II: $J \subseteq S$.

If $S \cap I = \emptyset$, then $S = S \setminus I$, hence $\max_{i\in S} \psi_i = \max_{i\in S\setminus I} \psi_i$.

If $S \cap I \neq \emptyset$, by the definition of ψ_i in equation (3), $\psi_i \leq \max_{j\in J} \psi_j, i \in I$, and since $\forall i \in S \cap I \subseteq I$, hence $\max_{i\in S\cap I} \psi_i \leq \max_{j\in J} \psi_j$. Because $J \subseteq S$ and $I \cap J = \emptyset$, then $J \subseteq S \setminus I$. Since $S = (S \cap I) \cup (S \setminus I)$, we have $\max_{i\in S} \psi_i = \max(\max_{i\in S\cap I} \psi_i, \max_{i\in S\setminus I} \psi_i)$ $\leq \max(\max_{j\in J} \psi_j, \max_{i\in S\setminus I} \psi_i) = \max_{i\in S\setminus I} \psi_i$. But $S \setminus I \subseteq S$, hence $\max_{i\in S} \psi_i \geq \max_{i\in S\setminus I} \psi_i$. Therefore

$\max_{i\in S} \psi_i = \max_{i\in S\setminus I} \psi_i$.

Similarly, by the definition of ψ_i in equation (3), $\psi_i \le \phi_i^I, i \in N \setminus I$, and since $\forall i \in S \setminus I \subseteq N \setminus I$, hence $\max_{i\in S\setminus I} \psi_i \le \max_{i\in S\setminus I} \phi_i^I$. By the assumption of Theorem 1, $\{\phi_i^I : i \in N\setminus I\} \in \Phi_\alpha(N\setminus I)$, and $S\setminus I \subseteq N\setminus I$, then $\forall \theta \in \bigcap_{i\in S\setminus I} H_i$, $P_\theta(\max_{i\in S\setminus I} \phi_i^I = 1) \le \alpha$. Therefore $\forall \theta \in \bigcap_{i\in S} H_i \subseteq \bigcap_{i\in S\setminus I} H_i$, $P_\theta(\max_{i\in S} \psi_i = 1) = P_\theta(\max_{i\in S\setminus I} \psi_i = 1) \le P_\theta(\max_{i\in S\setminus I} \phi_i^I = 1) \le \alpha$.

Combining Case I and II above, we have $\forall \emptyset \ne S \subseteq N$, $\forall \theta \in \bigcap_{i\in S} H_i$, $P_\theta(\max_{i\in S} \psi_i = 1) \le \alpha$, therefore $\{\psi_i : i \in N\} \in \Phi_\alpha(N)$.

References

Bretz, F., Maurer, W., Brannath, W. & Posch, M. (2009), 'A graphical approach to sequentially rejective multiple test procedure', *Statistics in Medicine* **28**, 586–604.

Bretz, F., Maurer, W. & Hommel, G. (2011), 'Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures', *Statistics in Medicine* **30**, 1489–1501.

Bretz, F., Maurer, W. & Maca, J. (2014), *Graphical Approaches to Multiple Testing. Clinical Trial Biostatistics and Biopharmaceutical Applications*, Taylor and Francis.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W. & Rohmeyer, K. (2011), 'Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests', *Biometrical Journal* **53**, 894–913.

Committee for Proprietary Medicinal Products, M. E. A. (2002), *Points to Consider on Multiplicity Issues in Clinical Trials*, European Medicines Agency.

Cummings, S. R., Eckert, S. & Krueger, K. (1999), 'The effect of raloxifene on risk of breast cancer in post-menopausal women', *The Journal of the American Medical Association* **281**, 2189–2197.

Dmitrienko, A., Offen, W. W. & Westfall, P. H. (2003), 'Gatekeeping strategies for clinical trials that do not require all primary effects to be significant', *Statistics in Medicine* **22**, 2387–2400.

Dmitrienko, A. & Tamhane, A. C. (2007), 'Gatekeeping procedures with clinical trial applications', *Pharmacetical Statistics* **6**, 171–180.

Dmitrienko, A. & Tamhane, A. C. (2013), 'General theory of mixture procedures for gatekeeping', *Biometrical Journal* **55**, 402–419.

Dmitrienko, A., Tamhane, A. C., Liu, W. & Wiens, B. L. (2008), 'A note on tree gatekeeping procedures in clinical trials', *Statistics in Medicine* **27**, 3446–3451.

Dmitrienko, A., Tamhane, A. C. & Wiens, B. L. (2008), 'General multistage gatekeeping procedures', *Biometrical Journal* **50**, 667–677.

Finner, H. & Strassburger, K. (2002), 'The partitioning principle: a powerful tool in multiple decision theory', *The Annals of Statistics* **30**, 1194–1213.

Food & Drug Administration, F. D. A. (2002), *Points to consider on multiplicity issues in clinical trials*, Washington, D. C.: Committee for Proprietary Medicinal Products, U.S.A. Food and Drug Administration.

Hochberg, Y. (1988), 'A sharper bonferroni procedure for multiple tests of significance', *Biometrika* **75**, 800–802.

Hochberg, Y. & Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: Wiley.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. (1988), 'A stagewise rejective multiple test procedure based on a modified bonferroni test', *Biometrika* **75**, 383–386.

Li, H. J., Yi, M. & Zhou, H. (2017), 'Generalized holm's procedure for multiple hypotheses testing problems', *Communications in Statistics - Theory and Methods* **46**, 7503–7510.

Marcus, R., Peritzl, E. & Gabriel, K. (1976), 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika* **63**, 655–660.

Maurer, W., Hothorn, L. & Lehmacher, W. (1995), *Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses*, Fischer-Verlag, Stutgart.

Sonnemann, E. (2008), 'General solutions to multiple testing problems', *Biometrical Journal* **50**, 641–656.

Westfall, P. H. & Krishen, A. (2001), 'Optimally weighted, fixed sequence, and gatekeeping multiple testing procedure', *Journal of Statistical Planning and Inference* **99**, 25–40.

Wiens, B. (2003), 'A fixed-sequence bonferroni procedure for testing multiple endpoints', *Pharmaceutical Statistics* **2**, 211–215.

Wiens, B. & Dmitrienko, A. (2005), 'The fallback procedure for evaluating a single family of hypotheses', *Journal of Biopharmaceutical Statistics* **15**, 929–942.